TITLE

Human leukocyte antigen susceptibility map for SARS-CoV-2

AUTHORS

Austin Nguyen (1, 2), Julianne K. David (1, 2), Sean K. Maden (1, 2), Mary A. Wood (1, 3), Benjamin R. Weeder (1, 2), Abhinav Nellore* (1, 2, 4), Reid F. Thompson* (1, 2, 5, 6, 7)

AFFILIATIONS

1. Computational Biology Program; Oregon Health & Science University; Portland, OR, 97239; USA
2. Department of Biomedical Engineering; Oregon Health & Science University; Portland, OR, 97239; USA
3. Portland VA Research Foundation; Portland, OR, 97239; USA
4. Department of Surgery; Oregon Health & Science University; Portland, OR, 97239; USA
5. Department of Radiation Medicine; Oregon Health & Science University; Portland, OR, 97239; USA
6. Department of Medical Informatics and Clinical Epidemiology; Oregon Health & Science University; Portland, OR, 97239; USA
7. Division of Hospital and Specialty Medicine; VA Portland Healthcare System; Portland, OR, 97239; USA

* co-corresponding authors [nellore@ohsu.edu, thompsre@ohsu.edu]

ABSTRACT

Genetic variability across the three major histocompatibility complex (MHC) class I genes (human leukocyte antigen [HLA] A, B, and C) may affect susceptibility to and severity of severe acute respiratory syndrome 2 (SARS-CoV-2), the virus responsible for coronavirus disease 2019 (COVID-19). We execute a comprehensive *in silico* analysis of viral peptide-MHC class I binding affinity across 145 HLA -A, -B, and -C genotypes for all SARS-CoV-2 peptides. We further explore the potential for cross-protective immunity conferred by prior exposure to four common human coronaviruses. The SARS-CoV-2 proteome is successfully sampled and presented by a diversity of HLA alleles. However, we found that HLA-B*46:01 had the fewest predicted binding peptides for SARS-CoV-2, suggesting individuals with this allele may be particularly vulnerable to COVID-19, as they were previously shown to be for SARS (1). Conversely, we found that HLA-B*15:03 showed the greatest capacity to present highly conserved SARS-CoV-2 peptides that are shared among common human coronaviruses, suggesting it could enable cross-protective T-cell based immunity. Finally, we report global distributions of HLA types with potential epidemiological ramifications in the setting of the current pandemic.

IMPORTANCE

Individual genetic variation may help to explain different immune responses to a virus across a population. In particular, understanding how variation in HLA may affect the course of COVID-19 could help identify individuals at higher risk from the disease. HLA typing can be fast and inexpensive. Pairing HLA typing with COVID-19 testing where feasible could improve assessment of viral severity in the population. Following the development of a vaccine against SARS-CoV-2, the virus that causes COVID-19, individuals with high-risk HLA types could be prioritized for vaccination.

INTRODUCTION

Recently, a new strain of betacoronavirus (severe acute respiratory syndrome coronavirus 2, or SARS-CoV-2) has emerged as a global pathogen, prompting the World Health Organization in January 2020 to declare an international public health emergency (2). In the large coronavirus family, comprising enveloped positive-strand RNA viruses, SARS-CoV-2 is the seventh encountered strain that causes respiratory disease in humans (3) ranging from mild -- the common cold -- to severe -- the zoonotic Middle East Respiratory Syndrome (MERS-CoV) and Severe Acute Respiratory Syndrome (SARS-CoV). As of April 2020, there are over one million presumed or confirmed cases of coronavirus disease 19 (COVID-19) worldwide, with total deaths exceeding 50,000 (4). While age and many comorbidities, including cardiovascular and pulmonary disease, appear to increase the severity and mortality of COVID-19 (5–10), approximately 80% of infected individuals have mild symptoms (11). As with SARS-CoV (12, 13) and MERS-CoV (14, 15), children seem to have low susceptibility to the disease (16–18); despite similar infection rates as adults (19) only 5.9% of pediatric cases are severe or critical, possibly due to lower binding ability of the ACE2 receptor in children or generally higher levels of antiviral antibodies (20). Other similarities (21–23) including genomic (24, 25) and immune system response (26–34) between SARS-CoV-2 and other coronaviruses (35), especially SARS-CoV and MERS-CoV, are topics of ongoing active research, results of which may inform an understanding of the severity of infection (36) and improve the ongoing work of immune landscape profiling (37–41) and vaccine discovery (29, 38, 42–49).

Human leukocyte antigen (HLA) alleles, which are critical components of the viral antigen presentation pathway, have been shown in previous studies to confer differential viral susceptibility and severity of disease. For instance, disease caused by the closely-related SARS-CoV (24, 25) shows increased severity among individuals with the HLA-B*46:01 genotype (1). Associations between HLA genotype and disease severity extend broadly to several other unrelated viruses. For example, in human immunodeficiency virus 1 (HIV-1), certain HLA types (e.g. HLA-A*02:05) may reduce risk of seroconversion (50), and in dengue virus, certain HLA alleles (e.g. HLA-A*02:07, HLA-B*51) are associated with increased secondary disease severity among ethnic Thais (51).

While a detailed clinical picture of the COVID-19 pandemic continues to emerge, there remain substantial unanswered questions regarding the role of individual genetic variability in the immune response against SARS-CoV-2 (51). We hypothesize that individual HLA genotypes may differentially induce the T-cell mediated antiviral response, and could potentially alter the course of disease and its transmission. In this study, we performed a comprehensive *in silico* analysis of viral peptide-MHC class I binding affinity~~,~~ across 145 different HLA types~~,~~ for the entire SARS-CoV-2 proteome.

RESULTS

To explore the potential for a given HLA allele to produce an antiviral response, we assessed the HLA binding affinity of all possible 8- to 12-mers from the SARS-CoV-2 proteome (n=48,395 unique peptides). We then removed from further consideration 16,138 peptides that were not predicted to enter the MHC class I antigen processing pathway via proteasomal cleavage. For the remaining 32,257 peptides, we repeated binding affinity predictions for a total of 145 different HLA types, and we show here the SARS-CoV-2-specific distribution of per-allele proteome presentation (predicted binding affinity threshold <500nM, Figure 1, Supplementary Table S1). Importantly, we note that the putative capacity for SARS-CoV-2 antigen presentation is unrelated to the

HLA allelic frequency in the population (Figure 1). We identify HLA-B*46:01 as the HLA allele with the fewest predicted binding peptides for SARS-CoV-2. We performed the same analyses for the closely-related SARS-CoV proteome (Supplementary Figure S1) and similarly note that HLA-B*46:01 was predicted to present the fewest SARS-CoV peptides, keeping with previous clinical data associating this allele with severe disease (1).

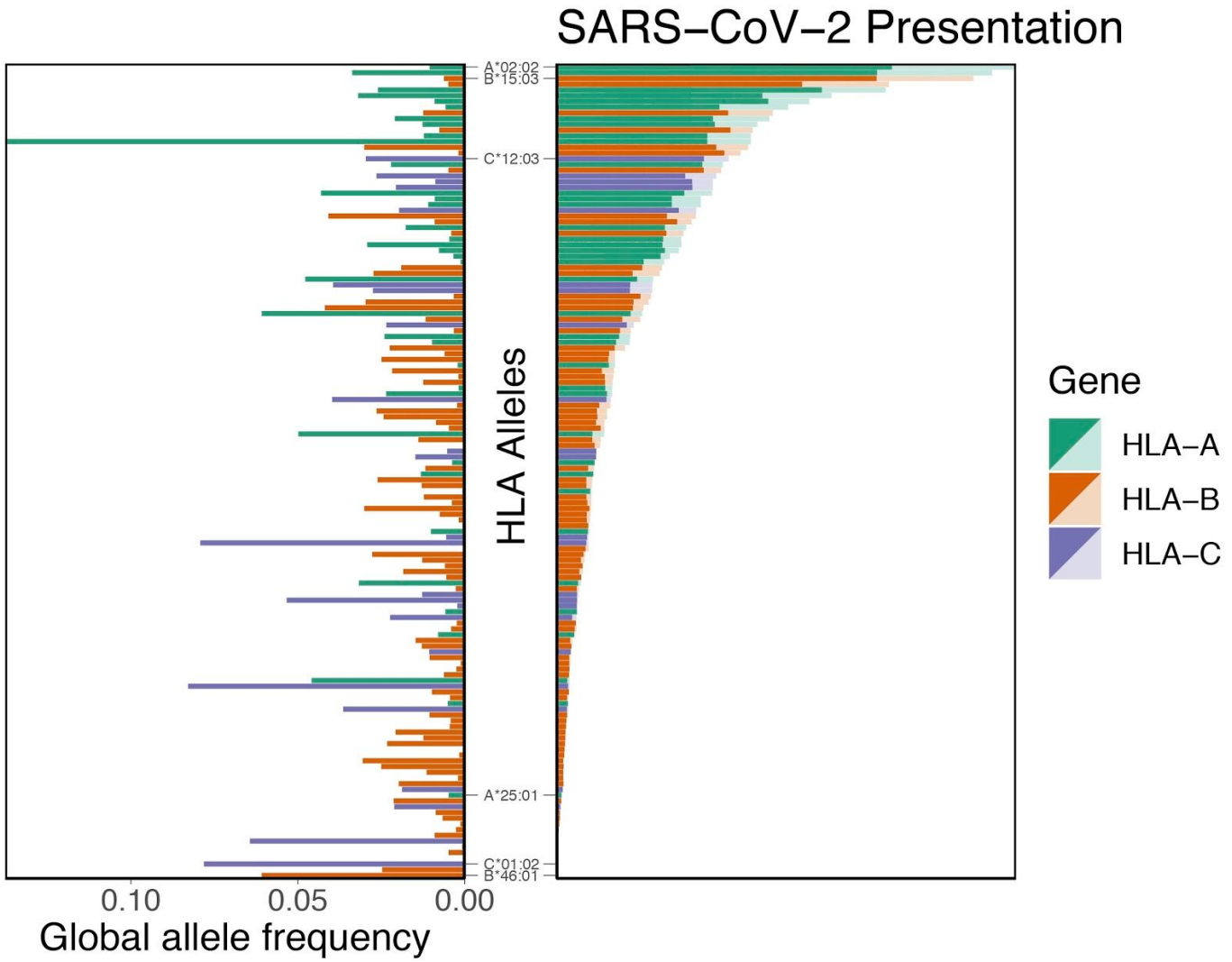


Figure 1: Distribution of HLA allelic presentation of 8- to 12-mers from the SARS-CoV-2 proteome. At right, the number of peptides (see Supplementary Table S1) that putatively bind to each of 145 HLA alleles is shown as a series of horizontal bars, with dark and light shading indicating the number of tightly (<50nM) and loosely (<500nM) binding peptides respectively, and with green, orange, and purple colors representing HLA-A, -B, and -C alleles, respectively. Alleles are sorted in descending order based on the number of peptides they bind (<500nM). The corresponding estimated allelic frequency in the global population is also shown (to left), with length of horizontal bar indicating absolute frequency in the population.

To assess the potential for cross-protective immunity conferred by prior exposure to common human coronaviruses (i.e. HKU1, OC43, NL63, and 229E), we next sought to characterize the conservation of the SARS-CoV-2 proteome across diverse coronavirus subgenera to identify highly conserved linear epitopes. After aligning reference proteome sequence data for 5 essential viral components (ORF1ab, S, E, M, and N proteins) across 34 distinct alpha- and betacoronaviruses, including all known human coronaviruses, we identified 48 highly conserved amino acid sequence spans (Appendix 1). Acknowledging the challenges inferring cross-protective immunity among closely related peptides, we confined attention exclusively to identical peptide matches. Among conserved sequences, 44 SARS-CoV-2 sequences would each be anticipated to generate at least one 8- to 12-mer linear peptide epitope also present within at least one other common human coronavirus (Supplementary Table S2, Figure 2). In total, 564 such 8- to 12-mer peptides were found to share 100% identity with corresponding OC43, HKU1, NL63, and 229E sequences (467, 460, 179, and 157 peptides, respectively) (Supplementary Table S3).

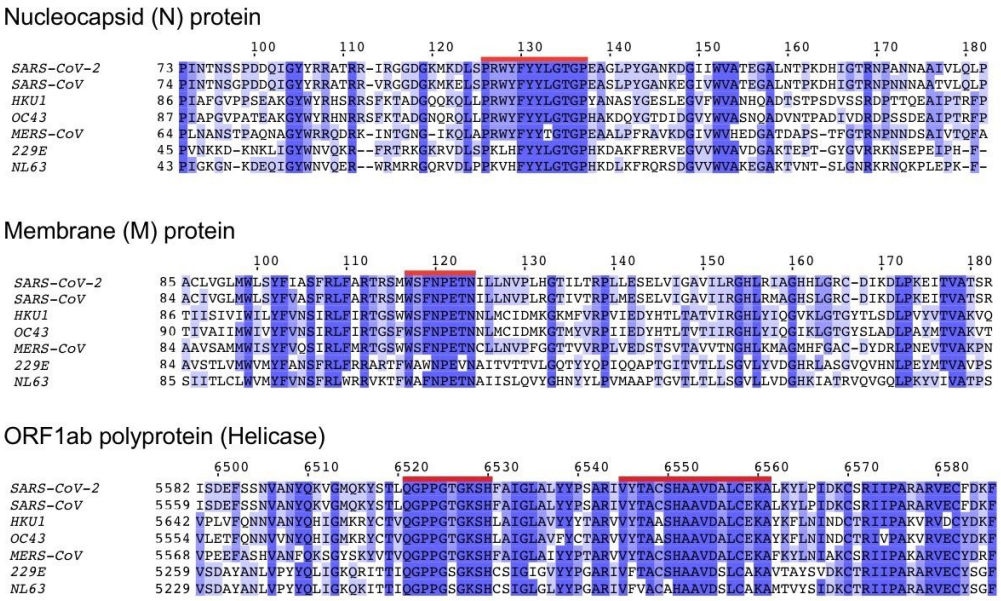


Figure 2: Amino acid sequence conservation of four linear peptide example sequences from three human coronavirus proteins. Protein sequence alignments are shown for nucleocapsid (N), membrane (M), and ORF1ab polyprotein (Helicase) across all five known human betacoronaviruses (SARS-CoV-2, SARS-CoV, HKU1, OC43, and MERS-CoV) and two known human alphacoronaviruses (229E and NL63). Each row in the three depicted sequence alignments corresponds to the protein sequence from the indicated coronavirus, with starting coordinate of the viral protein sequence shown at left and position coordinates of the overall alignment displayed above. Blue shading indicates the extent of sequence identity, with the darkest blue shading indicating 100% match for that amino acid across all sequences. The four red highlighted sequences correspond to highly conserved peptides ≥8 amino acids in length (PRWYFYYLGTGP, WSFNPETN, QPPGTGKSH, VYTACSHAAVDALCEKA, see Supplementary Table S2).

For the subset of these potentially cross-protective peptides that are anticipated to be generated via the MHC class I antigen processing pathway, we performed binding affinity predictions across 145 different HLA-A, -B, and -C alleles. As above, we demonstrate the SARS-CoV-2-specific distribution of per-allele presentation for these conserved peptides. We found that alleles HLA-A*02:02, HLA-B*15:03, and HLA-C*12:03 were the top presenters of conserved peptides. Conversely, we note that 56 different HLA alleles demonstrated no appreciable binding affinity (<500nM) to any of the conserved SARS-CoV-2 peptides, suggesting a concomitant lack of potential for cross-protective immunity from other human coronaviruses. We note, in particular, HLA-B*46:01 is among these alleles. Note also that the putative capacity for conserved peptide presentation is unrelated to the HLA allelic frequency in the population (Figure 3). Moreover, we see no appreciable global correlation between conservation of the SARS-CoV-2 proteome and its predicted MHC binding affinity, suggesting a lack of selective pressure for or against the capacity to present coronavirus epitopes (p=0.27 [Fisher's exact test], Supplementary Figure S2).

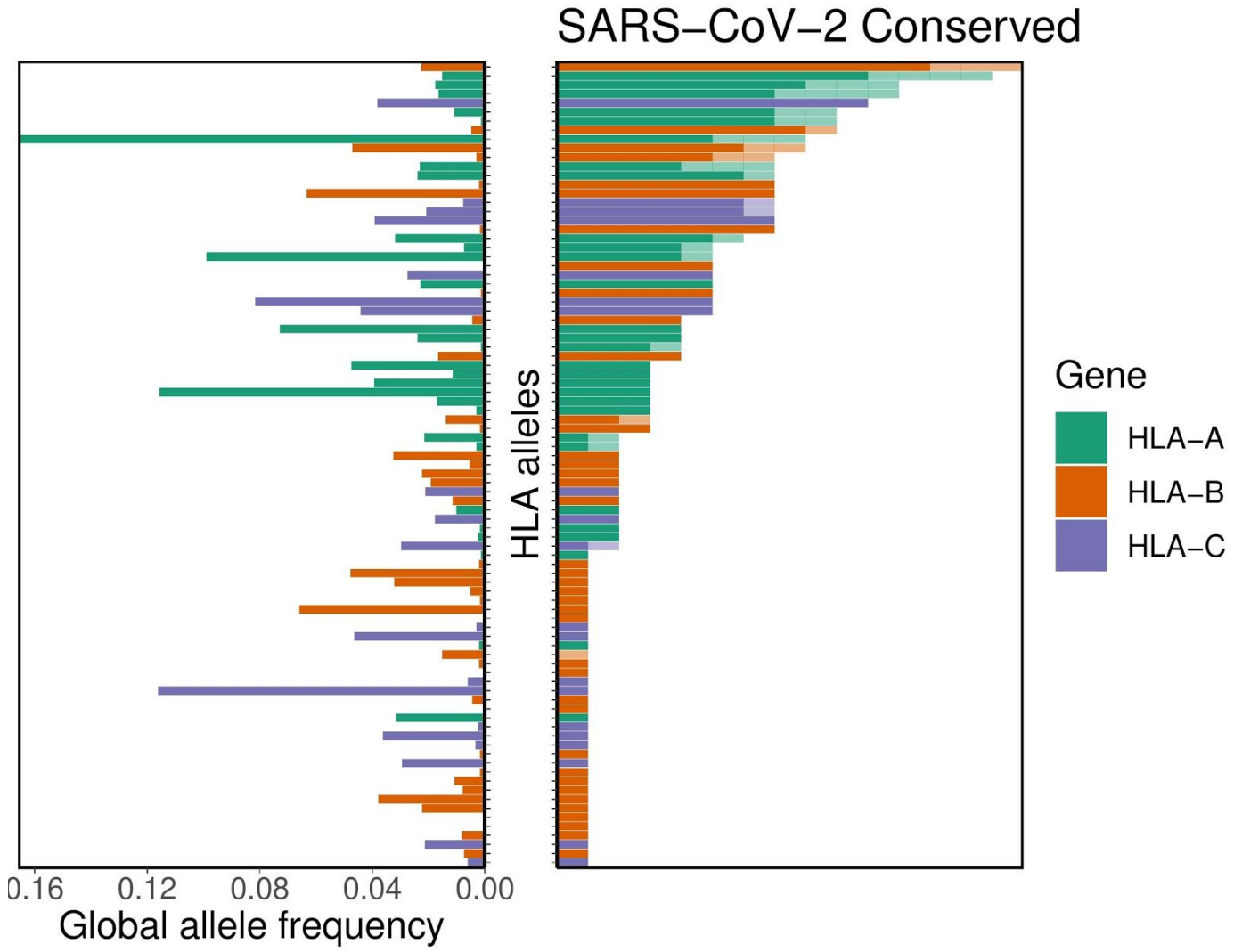


Figure 3: Distribution of HLA allelic presentation of highly conserved human coronavirus peptides with potential to elicit cross-protective immunity to COVID-19. At right, the number of conserved peptides (see

Supplementary Table S3) that putatively bind to a subset of 89 HLA alleles is shown as a series of horizontal bars, with dark and light shading indicating the number of tightly (<50nM) and loosely (<500nM) binding peptides, respectively, and with green, orange, and purple colors representing HLA-A, -B, and -C alleles, respectively. Alleles are sorted in descending order based on the number of peptides they are anticipated to present (binding affinity <500nM). The corresponding allelic frequency in the global population is also shown (to left), with length of horizontal bar indicating absolute frequency in the population.

We were further interested in whether certain regions of the SARS-CoV-2 proteome showed differential presentation by the MHC class I pathway. Accordingly, we surveyed the distribution of antigen presentation capacity across the entire proteome, highlighting its most conserved peptide sequences (Figure 4). Throughout the entire proteome, HLA-A and HLA-C alleles exhibited the relative largest and smallest capacity to present SARS-CoV-2 antigens, respectively. However each of the three major class I genes exhibited a very similar pattern of peptide presentation across the proteome (Supplementary Figure S3). We additionally note that peptide presentation appears to be independent of estimated time of peptide production during viral life cycle, with indistinguishable levels of peptide presentation of both early and late SARS-CoV-2 peptides (Supplementary Figure S4).

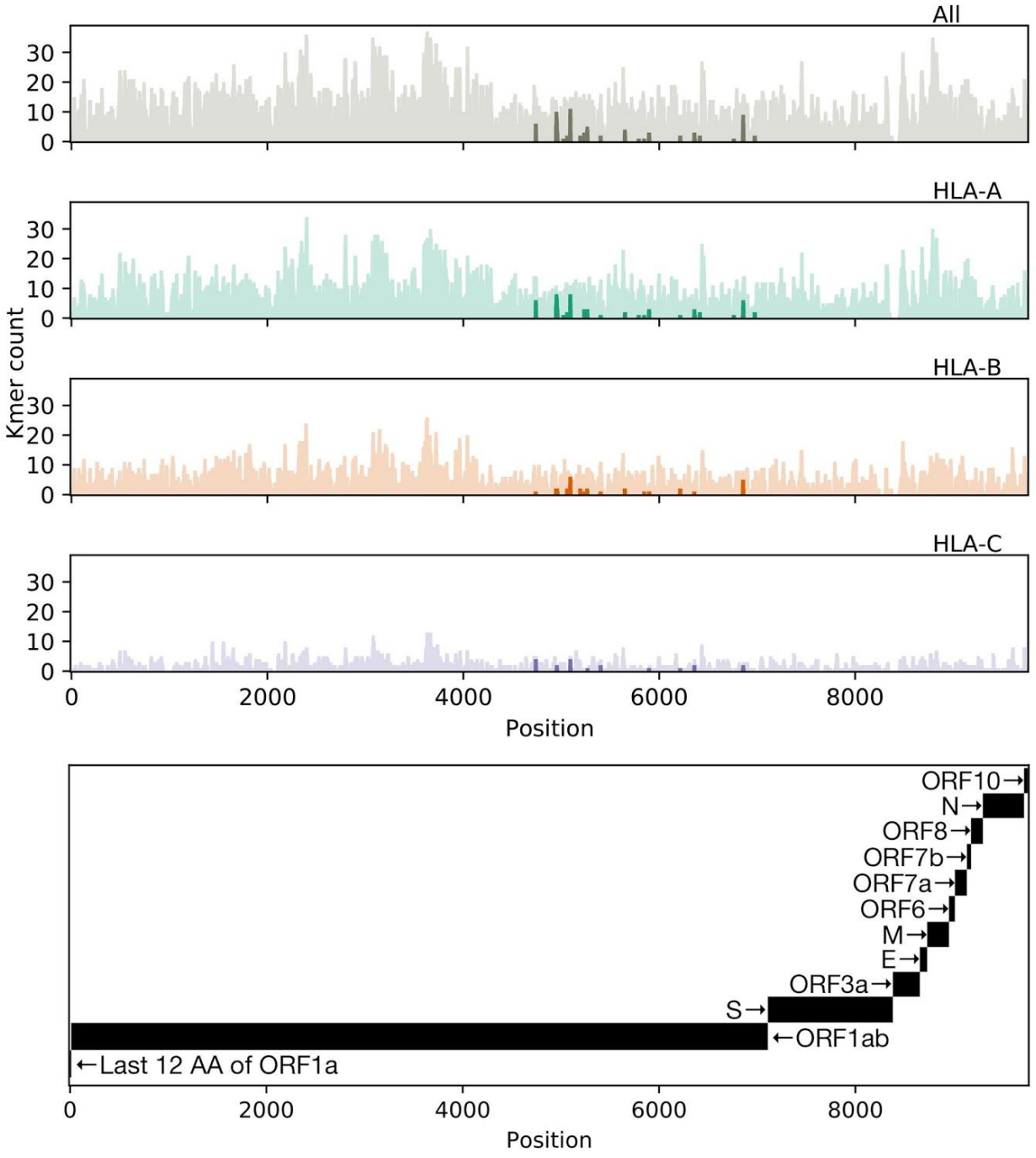


Figure 4: Distribution of allelic presentation of conserved 8- to 12-mers across the entire SARS-CoV-2 proteome for all HLA alleles and individually for HLA-A, HLA-B, and HLA-C (first, second, third, and fourth plots from top, respectively) with dark and light shading indicating the number of tightly (<50nM) and loosely (<500nM) binding peptides, respectively. Positions are derived from a concatenation of coding sequences (CDSes) indicated in the bottom panel. Tightly binding peptides are confined to ORF1ab. The sequence begins

with only the last 12 amino acids of ORF1a because all but the last four amino acids of ORF1a are contained in ORF1ab, and we considered binding peptides up to 12 amino acids in length.

Given the global nature of the current COVID-19 pandemic, we sought to describe population-level distributions of the HLA alleles best (and least) able to generate a repertoire of SARS-CoV-2 epitopes in support of a T-cell based immune response. While we present global maps of individual HLA allele frequencies for the full set of 145 different alleles studied herein (Appendix 2), we specifically highlight the global distributions of the three best-presenting (A*02:02, B*15:03, C*12:03) and three of the worst-presenting (A*25:01, B*46:01, C*01:02) HLA-A, -B, and -C alleles (Figure 5). Note that all allelic frequencies are aggregated by country, but implicitly reflect the distribution of HLA data available on the Allele Frequency Net Database (52).

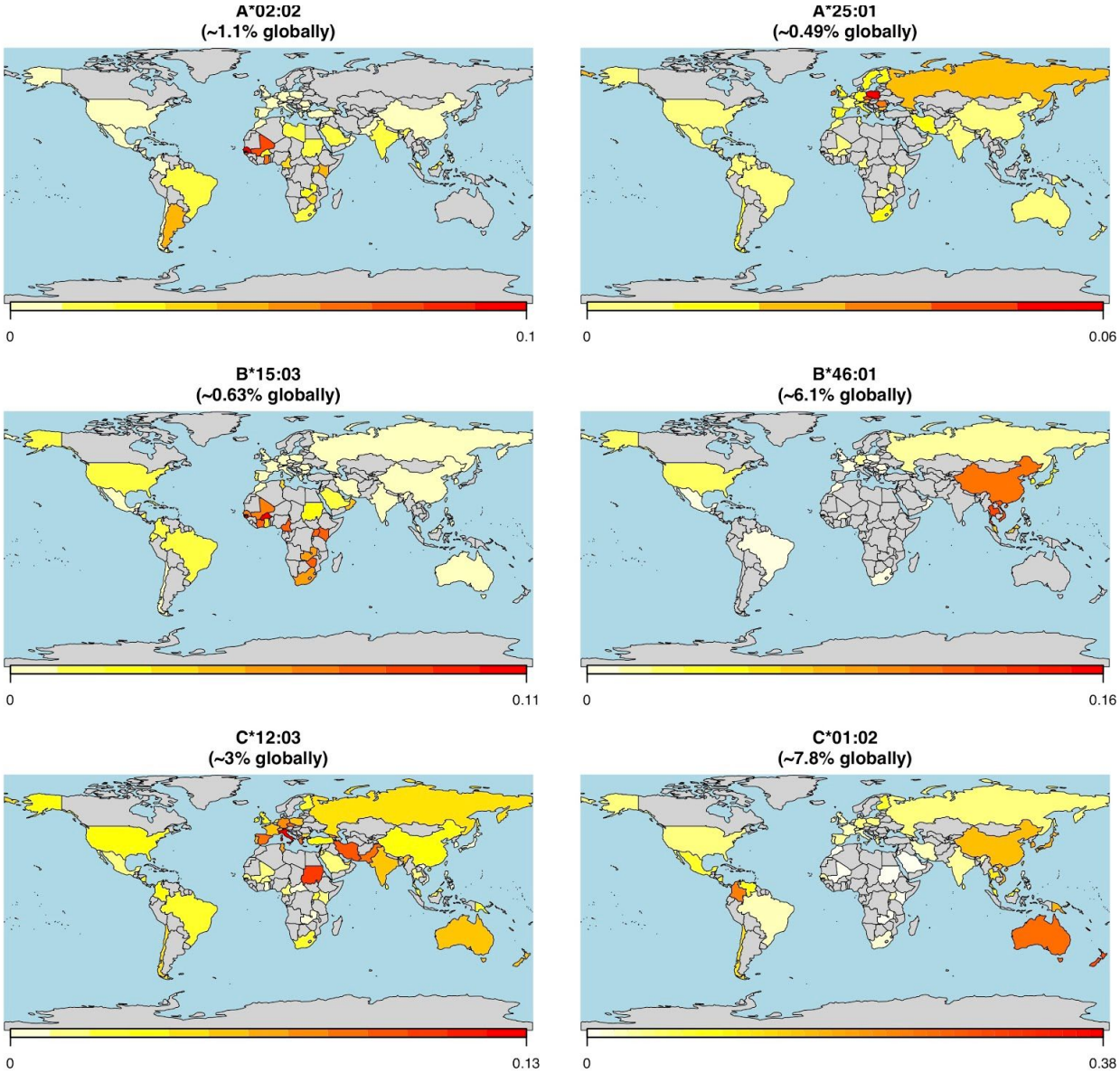


*Figure 5: Global HLA allele frequency distribution heatmaps for six HLA-A, -B, and -C alleles. The leftmost panels show the global allele frequency distributions by country for three representative alleles (HLA-A*02:02, HLA-B*15:03, and HLA-C*12:03) with the predicted capacities to present the greatest repertoire of epitopes from the SARS-CoV-2 proteome (21.1%, 19.1%, and 7.9% of presentable epitopes, respectively). Conversely, the rightmost panels show the global allele frequency distributions by country for three representative alleles (HLA-A*25:01, HLA-B*46:01, and HLA-C*01:02) with the least predicted epitope presentation from the SARS-CoV-2 proteome (0.2%, 0%, 0% of presentable epitopes, respectively). Heatmap color corresponds to the individual HLA allele frequency within each country ranging from least (white/yellow) to most (red) frequent as indicated in the legend below each map.*

Finally, we acknowledge that nearly all individuals have two HLA-A/B/C haplotypes constituting as few as three but as many as six distinct alleles, potentially buffering against the lack of presentation from a single

poorly-presenting allele. We sought to describe whether allele-specific variability in SARS-CoV-2 presentation extends to full HLA haplotypes and to whole individual HLA genotypes. For six representative alleles with the highest (HLA-A*02:02, HLA-B*15:03, and HLA-C*12:03) and lowest (HLA-A*25:01, HLA-B*46:01, and HLA-C*01:02) predicted capacity for SARS-CoV-2 epitope presentation, these differences remain significant at the haplotype level, albeit with wide variability in presentation among different haplotypes (Figure 6). Haplotype-level data for all 145 alleles is included in Supplementary Figure S5 and Appendix 2. We then identified 3,382 individuals with full HLA genotype data and noted wide variability in their capacity to present peptides from the SARS-CoV-2 proteome, albeit with a small minority of individuals at either extreme (Supplementary Figure S6).

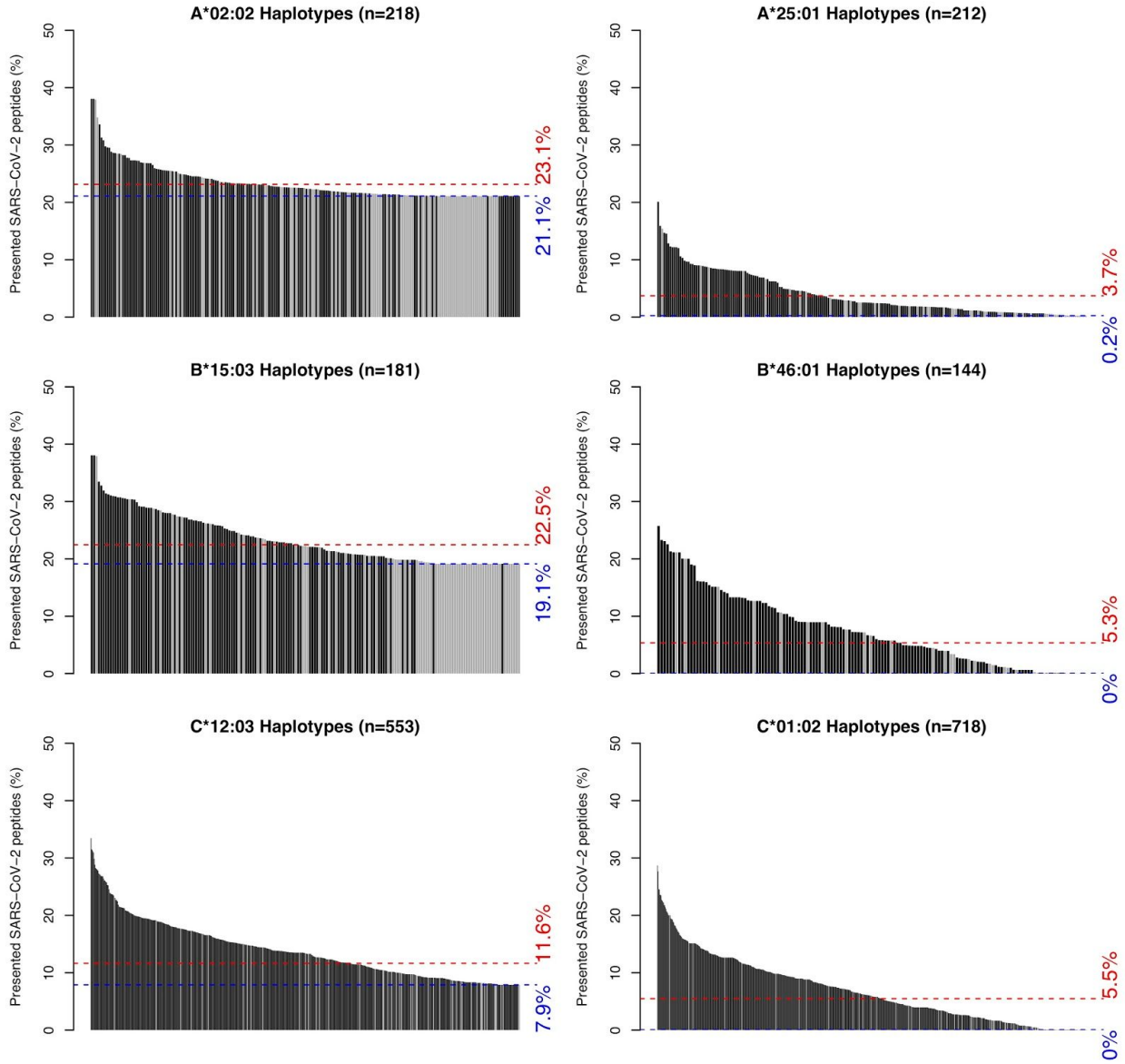


*Figure 6: Distributions of SARS-CoV-2 peptide presentation across HLA haplotypes. The leftmost panels show the distributions of SARS-CoV-2 peptide presentation capacity for haplotypes containing one of three representative HLA alleles (HLA-A*02:02, HLA-B*15:03, and HLA-C*12:03) with the greatest predicted*

*repertoire of epitopes from the SARS-CoV-2 proteome. Conversely, the rightmost panels show the distributions of SARS-CoV-2 peptide presentation capacity for haplotypes containing one of three representative alleles (HLA-A*25:01, HLA-B*46:01, and HLA-C*01:02) with the least predicted epitope presentation from the SARS-CoV-2 proteome. Black and gray bars represent full or partial haplotypes, respectively. Blue and red Dashed lines represent the percent of presented SARS-CoV-2 peptides for the indicated allele itself (blue) and its global population frequency weighted average presentation across its observed haplotypes (red).*

DISCUSSION

To the best of our knowledge, this is the first study to evaluate per-allele viral proteome presentation across a wide range of HLA alleles using MHC-peptide binding affinity predictors. This study also introduces the relationship between coronavirus sequence conservation and MHC class I antigen presentation. We show that individual HLA, haplotype, and full genotype variability likely influence the capacity to respond to SARS-CoV-2 infection, and we note certain alleles in particular (e.g. HLA-B*46:01) that could be associated with more severe infection, as previously shown with SARS-CoV (1). Indeed, we further compare SARS-CoV and SARS-CoV-2 peptide presentation and note a high degree of similarity between the two across HLA types. Finally, this is the first study to report global distributions of HLA types and haplotypes with potential epidemiological ramifications in the setting of the current pandemic. We found that in general, there is no correlation between the HLA allelic frequency in the population and allelic capacity to bind SARS-CoV or SARS-CoV-2 peptides, irrespective of estimated timing of peptide production during the viral replication cycle. While we are not aware of any studies explicitly reporting the relationship between human coronavirus epitope abundance and immune response, there is data in vaccinia virus that suggests that early peptide antigens are more likely to generate CD8+ T-cell responses while antibody and CD4+ T-cell responses are more likely to target later mRNA expression with higher peptide abundance in the virion (53).

We note, however, several limitations to ~~this~~ our work. First and foremost, while we note that a handful of our binding affinity predictions were borne out in experimentally validated SARS-CoV peptides (Supplementary Table S4), we acknowledge that this is an entirely *in silico* study. As we are unable to obtain individual-level HLA typing and clinical outcomes data for any real-world COVID-19 populations at this time, the data presented is theoretical in nature, and subject to many of the same limitations implicit to the MHC binding affinity prediction tool(s) upon which it is based. As such, we are unable to assess the relative importance of HLA type compared to known disease-modifying risk factors such as age and clinical comorbidities (5–10). We further note that peptide-MHC binding affinity is limited as a predictor of subsequent T-cell responses (54–56), and we do not study T-cell responses herein. As such, we are ill-equipped to explore phenomena such as original antigenic sin (57–59), where prior exposure to closely related infection(s) may trigger T-cell anergy (60–62) or immunopathogenesis (63) in the setting of a novel infection. We explored only a limited set of 145 well-studied HLA alleles, but note that this analysis could be performed across a wider diversity of genotypes (49). Additionally, we did not assess genotypic heterogeneity or *in vivo* evolution of SARS-CoV-2, which could modify the repertoire of viral epitopes presented, or otherwise modulate virulence in an HLA-independent manner (64, 65) (https://nextstrain.org/ncov). We also do not address the potential for individual-level genetic variation in other proteins (e.g. angiotensin converting enzyme 2 [ACE2] or transmembrane serine protease 2 [TMPRSS2], essential host proteins for SARS-CoV-2 priming and cell entry (66) to modulate the host-pathogen interface.

Unless and until the findings we present here are clinically validated, they should not be employed for any clinical purposes. However, we do at this juncture recommend integrating HLA testing into clinical trials and pairing HLA typing with COVID-19 testing where feasible to more rapidly develop and deploy predictor(s) of viral severity in the population, and potentially to tailor future vaccination strategies to genotypically at-risk populations. This approach may have additional implications for the management of a broad array of other viruses.

MATERIALS AND METHODS

Sequence retrieval and alignments

Full polyprotein 1ab (ORF1ab), spike (S) protein, membrane (M) protein, envelope (E) protein, and nucleocapsid (N) protein sequences were obtained for each of 34 distinct but representative alpha and betacoronaviruses from a broad genus and subgenus distribution, including all known human coronaviruses (i.e. SARS-CoV, SARS-CoV-2, MERS-CoV, HKU1, OC43, NL63, and 229E). FASTA-formatted protein sequence data (full accession number list available in Supplementary Table ~~S4~~S5) was retrieved from the National Center of Biotechnology Information (NCBI) (67). For each protein class (i.e. ORF1ab, S, M, E, N), all 34 coronavirus sequences were aligned using the Clustal Omega v1.2.4 multisequence aligner tool employing the following parameters: sequence type [Protein], output alignment format [clustal_num], dealign [false], mBed-like clustering guide-tree [true], mBed-like clustering iteration [true], number of combined iterations [0], maximum guide tree iterations [-1], and maximum HMM iterations [-1] (68). For the purposes of estimating time of viral peptide production, we classified ORF1a and ORF1b peptides as “early” while all other peptides produced by subgenomic mRNAs were classified as “late” (69, 70).

Conserved peptide assessment

Aligned sequences were imported into Jalview v. 2.1.1 (71) with automated generation of the following alignment annotations: 1) sequence consensus, calculated as the percentage of the modal residue per column, 2) sequence conservation (0-11), measured as a numerical index reflecting conservation of amino acid physico-chemical properties in the alignment, 3) alignment quality (0-1), measured as a normalized sum of BLOSUM62 ratios for all residues at each position, 4) occupancy, calculated as the number of aligned residues (not including gaps) for each position. In all cases, sequence conservation was assessed for each of three groups: only human coronaviruses (n=7), all betacoronaviruses (n=16), and combined alpha- and betacoronavirus sequences (n=34). Aligned SARS-CoV-2 sequence and all annotations were manually exported for subsequent analysis. Conserved human coronavirus peptides were defined as those with a length ≥8 consecutive amino acids, each with an agreement of SARS-CoV-2 and ≥4 other human coronavirus sequences with the consensus sequence (Supplementary Table S2). For each of these conserved peptides, we also assessed the component number of 8- to 12-mers sharing identical amino acid sequence between SARS-CoV-2 and each of the four other common human coronaviruses (i.e. OC43, HKU1, NL63, 229E) (Supplementary Table S3). For all peptides, human, beta, and combined conservation scores were obtained using a custom R v.3.6.2 script as the mean sequence conservation (minus gap penalties where relevant) (see https://github.com/pdxgx/covid19).

Peptide-MHC class I binding affinity predictions

FASTA-formatted input protein sequences from the entire SARS-CoV-2 and SARS-CoV proteomes were obtained from NCBI RefSeq database (67) under accession numbers NC_045512.2 and NC_004718.3. We

kmerized each of these sequences into 8- to 12-mers to assess MHC class I-peptide binding affinity across the entire proteome. MHC class I binding affinity predictions were performed using 145 different HLA alleles for which global allele frequency data was available as described previously (72) (see Supplementary Table ~~S1~~ S5) with netMHCpan v4.0 (73) using the '-BA' option to include binding affinity predictions and the '-l' option to specify peptides of lengths 8-12 (Supplementary Table S1). Binding affinity was not predicted for peptides containing the character '|' in their sequences. Additional MHC class I binding affinity predictions were performed on all 66 MHCflurry supported alleles (--list-supported-alleles, Supplementary Table S6) using both MHCnuggets 2.3.2 (74) and MHCflurry 1.4.3 (75) (Supplementary Tables S7 and S8, Supplementary Figures S7-S10). We further cross-referenced these lists of peptides with existing experimentally validated SARS-CoV epitopes present in the Immune Epitope Database (Supplementary Table S4) (76). We then performed consensus binding affinity predictions for the 66 supported alleles shared by all three tools by taking the union set of alleles and filtering for peptide-allele pairs matching the union set of alleles. For the SARS-CoV and SARS-CoV-2 specific distribution of per-allele proteome presentation, we exclude all peptides-allele pairs with >500nM predicted binding. In all cases, we used the netchop v3.0 (77) "C-term" model with a cleavage threshold of 0.1 to further remove any peptides that were not predicted to undergo canonical MHC class I antigen processing via proteasomal cleavage (of the peptide's C-terminus).

Global HLA allele and haplotype frequencies

HLA-A, -B, and -C allele and haplotype frequency data were obtained from the Allele Frequency Net Database (52) for 805 distinct populations pertaining to 101 different countries and 2628 distinct major/minor (4-digit) alleles, corresponding to 20,478 distinct haplotypes (https://github.com/pdxgx/covid19). We also identified full HLA genotype data for 3,382 individuals whose HLA types were confined to the 145 HLA alleles studied herein. Population allele and haplotype frequency data were aggregated by country as a mean of all constituent population allele or haplotype frequencies weighted by sample size of the population, but not accounting for representative ethnic demographic size of the population. Global allele frequency maps were generated using the rworldmap v1.3-6 package (78), with total global allele and haplotype frequency estimates calculated as the mean of per-country allele and haplotype frequencies, weighted by each country's population in 2005.

ACKNOWLEDGEMENTS

We thank Drs. Christopher Loo and Jeffrey Barnet for their critical reading of the manuscript. We thank Drs. Jonah Sacha and Paul Spellman for their helpful discussions.

APPENDICES

Appendix 1 contains graphical depictions of sequence alignments for 5 protein sequences (ORF1ab polyprotein, spike, envelope, membrane, and nucleocapsid) across 34 diverse coronavirus proteomes (including all 7 human coronaviruses). Appendix 2 contains a series of world maps displaying the estimated global distributions and frequencies across 145 different HLA alleles, as well as the distribution of predicted SARS-CoV-2 peptide binding and global haplotype frequencies across all haplotypes containing that HLA allele. Appendix 3 contains Supplementary Tables S1-S9~~S8~~ as follows: list of SARS-CoV-2 peptides and their predicted binding affinity across 145 HLA alleles (Supplementary Table S1), SARS-CoV-2 peptides conserved across diverse coronavirus sequences (Supplementary Table S2), presence of 8- to 12-mer peptides across four human coronavirus sequences (Supplementary Table S3), prediction of validated SARS-CoV peptides

from the Immune Epitope Database (Supplementary Table S4), coronavirus taxonomy and sequence accession numbers for conserved coronavirus proteins (Supplementary Table ~~S4~~S5), list of 145 HLA alleles used for binding affinity predictions (Supplementary Table ~~S5~~S6), list of SARS-CoV peptides and their predicted binding affinity across 145 HLA alleles (Supplementary Table ~~S6~~S7), list of SARS-CoV-2 peptides and their predicted MHCnuggets and MHCflurry binding affinities across 145 HLA alleles (Supplementary Table ~~S7~~S8), list of SARS-CoV peptides and their predicted MHCnuggets and MHCflurry binding affinities across 145 HLA alleles (Supplementary Table ~~S8~~S9). Appendix 4 contains the full list of SARS-CoV-2 8- to 12-mers with their individual netMHCpan v4.0 predicted binding affinities across all 145 HLA alleles.

DISCLAIMER

The contents do not represent the views of the U.S. Department of Veterans Affairs or the United States Government.

FUNDING

RFT was supported by the U.S. Department of Veterans Affairs under award number 1IK2CX002049-01, and the Sunlin & Priscilla Chou Foundation.

REFERENCES

1. Lin M, Tseng H-K, Trejaut JA, Lee H-L, Loo J-H, Chu C-C, Chen P-J, Su Y-W, Lim KH, Tsai Z-U, Lin R-Y, Lin R-S, Huang C-H. 2003. Association of HLA class I with severe acute respiratory syndrome coronavirus infection. BMC Med Genet 4:9.

2. World Health Organization. 2020. Statement on the second meeting of the International Health Regulations (2005) Emergency Committee regarding the outbreak of novel coronavirus (2019-nCoV). WHO | World Health Organization.

3. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, Zhao X, Huang B, Shi W, Lu R, Niu P, Zhan F, Ma X, Wang D, Xu W, Wu G, Gao GF, Tan W. 2020. A Novel Coronavirus from Patients with Pneumonia in China, 2019. New England Journal of Medicine.

4. Max Roser HR, Ortiz-Ospina E. 2020. Coronavirus Disease (COVID-19). Our World in Data.

5. Caramelo F, Ferreira N, Oliveiros B. 2020. Estimation of risk factors for COVID-19 mortality-preliminary results. medRxiv.

6. Jain V, Yuan J-M. Systematic review and meta-analysis of predictive symptoms and comorbidities for severe COVID-19 infection.

7. Yang X, Yu Y, Xu J, Shu H, Xia J ’an, Liu H, Wu Y, Zhang L, Yu Z, Fang M, Yu T, Wang Y, Pan S, Zou X, Yuan S, Shang Y. 2020. Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study. The Lancet Respiratory Medicine.

8. Wang D, Hu B, Hu C, Zhu F, Liu X, Zhang J, Wang B, Xiang H, Cheng Z, Xiong Y, Zhao Y, Li Y, Wang X, Peng Z. 2020. Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus–Infected Pneumonia in Wuhan, China. JAMA.

9. Zhou F, Yu T, Du R, Fan G, Liu Y, Liu Z, Xiang J, Wang Y, Song B, Gu X, Guan L, Wei Y, Li H, Wu X, Xu J, Tu S, Zhang Y, Chen H, Cao B. 2020. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. The Lancet.

10. Guan W-J, Liang W-H, Zhao Y, Liang H-R, Chen Z-S, Li Y-M, Liu X-Q, Chen R-C, Tang C-L, Wang T, Others. 2020. Comorbidity and its impact on 1,590 patients with COVID-19 in China: A Nationwide Analysis. medRxiv.

11. Novel Coronavirus Pneumonia Emergency Response Epidemiology Team. 2020. [The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19) in China]. Zhonghua Liu Xing Bing Xue Za Zhi 41:145–151.

12. Lau JTF, Lau M, Kim JH, Wong E, Tsui H-Y, Tsang T, Wong TW. 2004. Probable Secondary Infections in Households of SARS Patients in Hong Kong. Emerging Infectious Diseases.

13. Denison MR. 2004. Severe acute respiratory syndrome coronavirus pathogenesis, disease and vaccines: an update. Pediatr Infect Dis J 23:S207–14.

14. Thabet F, Chehab M, Bafaqih H, AlMohaimeed S. 2015. Middle East respiratory syndrome coronavirus in children. Saudi Medical Journal.

15. Al-Tawfiq JA, Kattan RF, Memish ZA. 2016. Middle East respiratory syndrome coronavirus disease is rare in children: An update from Saudi Arabia. World J Clin Pediatr 5:391–396.

16. Cao Q, Chen Y-C, Chen C-L, Chiu C-H. 2020. SARS-CoV-2 infection in children: Transmission dynamics and clinical characteristics. Journal of the Formosan Medical Association.

17. Tang A, Xu W, Shen M, Chen P, Li G, Liu Y, Liu L. A retrospective study of the clinical characteristics of COVID-19 infection in 26 children.

18. Lu X, Zhang L, Du H, Zhang J, Li YY, Qu J, Zhang W, Wang Y, Bao S, Li Y, Wu C, Liu H, Liu D, Shao J, Peng X, Yang Y, Liu Z, Xiang Y, Zhang F, Silva RM, Pinkerton KE, Shen K, Xiao H, Xu S, Wong GWK, Chinese Pediatric Novel Coronavirus Study Team. 2020. SARS-CoV-2 Infection in Children. N Engl J Med.

19. Bi Q, Wu Y, Mei S, Ye C, Zou X, Zhang Z, Liu X, Wei L, Truelove SA, Zhang T, Gao W, Cheng C, Tang X, Wu X, Wu Y, Sun B, Huang S, Sun Y, Zhang J, Ma T, Lessler J, Feng T. Epidemiology and Transmission of COVID-19 in Shenzhen China: Analysis of 391 cases

and 1,286 of their close contacts.

20. Dong Y, Mo X, Hu Y, Qi X, Jiang F, Jiang Z, Tong S. 2020. Epidemiological Characteristics of 2143 Pediatric Patients With 2019 Coronavirus Disease in China. Pediatrics.

21. Ashour HM, Elkhatib WF, Rahman MM, Elshabrawy HA. 2020. Insights into the Recent 2019 Novel Coronavirus (SARS-CoV-2) in Light of Past Human Coronavirus Outbreaks. Pathogens 9.

22. Yang Y, Peng F, Wang R, Guan K, Jiang T, Xu G, Sun J, Chang C. 2020. The deadly coronaviruses: The 2003 SARS pandemic and the 2020 novel coronavirus epidemic in China. Journal of Autoimmunity.

23. Ge Y, Tian T, Huang S, Wan F, Li J, Li S, Yang H, Hong L, Wu N, Yuan E, Cheng L, Lei Y, Shu H, Feng X, Jiang Z, Chi Y, Guo X, Cui L, Xiao L, Li Z, Yang C, Miao Z, Tang H, Chen L, Zeng H, Zhao D, Zhu F, Shen X, Zeng J. A data-driven drug repositioning framework discovered a potential therapeutic agent targeting COVID-19.

24. Ceraolo C, Giorgi FM. 2020. Genomic variance of the 2019-nCoV coronavirus. J Med Virol 92:522–528.

25. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, Wang W, Song H, Huang B, Zhu N, Bi Y, Ma X, Zhan F, Wang L, Hu T, Zhou H, Hu Z, Zhou W, Zhao L, Chen J, Meng Y, Wang J, Lin Y, Yuan J, Xie Z, Ma J, Liu WJ, Wang D, Xu W, Holmes EC, Gao GF, Wu G, Chen W, Shi W, Tan W. 2020. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet 395:565–574.

26. Prompetchara E, Ketloy C, Palaga T. 2020. Immune responses in COVID-19 and potential

vaccines: Lessons learned from SARS and MERS epidemic. Asian Pac J Allergy Immunol.

27. Zheng Z, Monteil VM, Maurer-Stroh S, Yew CW, Leong C, Mohd-Ismail NK, Arularasu SC, Chow VTK, Pin RLT, Mirazimi A, Hong W, Tan Y-J. Monoclonal antibodies for the S2 subunit of spike of SARS-CoV cross-react with the newly-emerged SARS-CoV-2.

28. Yang, Yang Y, Shen C, Li J, Yuan J, Yang M, Wang F, Li G, Li Y, Xing L, Peng L, Wei J, Cao M, Zheng H, Wu W, Zou R, Li D, Xu Z, Wang H, Zhang M, Zhang Z, Liu L, Liu Y. Exuberant elevation of IP-10, MCP-3 and IL-1ra during SARS-CoV-2 infection is associated with disease severity and fatal outcome.

29. Ahmed SF, Quadeer AA, McKay MR. 2020. Preliminary Identification of Potential Vaccine Targets for the COVID-19 Coronavirus (SARS-CoV-2) Based on SARS-CoV Immunological Studies. Viruses.

30. Li G, Fan Y, Lai Y, Han T, Li Z, Zhou P, Pan P, Wang W, Hu D, Liu X, Zhang Q, Wu J. 2020. Coronavirus infections and immune responses. J Med Virol 92:424–432.

31. Lokugamage KG, Schindewolf C, Menachery VD. SARS-CoV-2 sensitive to type I interferon pretreatment.

32. Wang C, Li W, Drabek D, Okba NMA, van Haperen R, Albert D M, van Kuppeveld FJM, Haagmans BL, Grosveld F, Bosch B-J. A human monoclonal antibody blocking SARS-CoV-2 infection.

33. Lv H, Wu NC, Tsang OT-Y, Yuan M, Ranawaka A P, Leung WS, So RTY, Chan JMC, Yip GK, Chik TSH, Wang Y, Choi CYC, Lin Y, Ng WW, Zhao J, Poon LLM, Malik Peiris JS, Wilson IA, Mok CKP. Cross-reactive antibody response between SARS-CoV-2 and

SARS-CoV infections.

34. Yuan M, Wu NC, Zhu X, Lee C-CD, So RTY, Lv H, Mok CKP, Wilson IA. A highly conserved cryptic epitope in the receptor-binding domains of SARS-CoV-2 and SARS-CoV.

35. Tan Y, Schneider T, Leong M, Aravind L, Zhang D. Novel Immunoglobulin Domain Proteins Provide Insights into Evolution and Pathogenesis Mechanisms of SARS-Related Coronaviruses.

36. Tetro JA. 2020. Is COVID-19 receiving ADE from other coronaviruses? Microbes Infect 22:72–73.

37. Breadth of concomitant immune responses underpinning viral clearance and patient recovery in a non-severe case of COVID-19.

38. Zhu J, Kim J, Xiao X, Wang Y, Luo D, Chen R, Xu L, Zhang H, Xiao G, Zhan X, Wang T, Xie Y. Profiling the Immune Vulnerability Landscape of the 2019 Novel Coronavirus.

39. Dong C, Ni L, Ye F, Chen M-L, Feng Y, Deng Y-Q, Zhao H, Wei P, Ge J, Li X, Sun L, Wang P, Liang P, Guo H, Wang X, Qin C-F, Chen F. Characterization of anti-viral immunity in recovered individuals infected by SARS-CoV-2.

40. Okba NMA, Muller MA, Li W, Wang C, GeurtsvanKessel CH, Corman VM, Lamers MM, Sikkema RS, de Bruin E, Chandler FD, Yazdanpanah Y, Le Hingrat Q, Descamps D, Houhou-Fidouh N, Chantal B E, Bosch B-J, Drosten C, Koopmans MPG, Haagmans BL. SARS-CoV-2 specific antibody responses in COVID-19 patients.

41. Chen G, Wu D, Guo W, Cao Y, Huang D, Wang H, Wang T, Zhang X, Chen H, Yu H, Zhang X, Zhang M, Wu S, Song J, Chen T, Han M, Li S, Luo X, Zhao J, Ning Q. 2020.

Clinical and immunologic features in severe and moderate Coronavirus Disease 2019. J Clin Invest.

42. Designing of a next generation multiepitope based vaccine (MEV) against SARS-COV-2: Immunoinformatics and in silico approaches.

43. Fast E, Chen B. Potential T-cell and B-cell Epitopes of 2019-nCoV.

44. Abdelmageed MI, Abdelmoneim AH, Mustafa MI, Elfadol NM, Murshed NS, Shantier SW, Makhawi AM. Design of multi epitope-based peptide vaccine against E protein of human COVID-19: An immunoinformatics approach.

45. Baruah V, Bose S. 2020. Immunoinformatics-aided identification of T cell and B cell epitopes in the surface glycoprotein of 2019-nCoV. J Med Virol 92:495–500.

46. Robson B. 2020. Computers and viral diseases. Preliminary bioinformatics studies on the design of a synthetic vaccine and a preventative peptidomimetic antagonist against the SARS-CoV-2 (2019-nCoV, COVID-19) coronavirus. Computers in Biology and Medicine.

47. Dhama K, Sharun K, Tiwari R, Dadar M, Malik YS, Singh KP, Chaicumpa W. 2020. COVID-19, an emerging coronavirus infection: advances and prospects in designing and developing vaccines, immunotherapeutics, and therapeutics. Hum Vaccin Immunother 1–7.

48. Grifoni A, Sidney J, Zhang Y, Scheuermann RH, Peters B, Sette A. 2020. A Sequence Homology and Bioinformatic Approach Can Predict Candidate Targets for Immune Responses to SARS-CoV-2. Cell Host Microbe.

49. Campbell KM, Steiner G, Wells DK, Ribas A, Kalbasi A. Prediction of SARS-CoV-2

epitopes across 9360 HLA class I alleles.

50. MacDonald KS, Fowke KR, Kimani J, Dunand VA, Nagelkerke NJ, Ball TB, Oyugi J, Njagi E, Gaur LK, Brunham RC, Wade J, Luscher MA, Krausa P, Rowland-Jones S, Ngugi E, Bwayo JJ, Plummer FA. 2000. Influence of HLA supertypes on susceptibility and resistance to human immunodeficiency virus type 1 infection. J Infect Dis 181:1581–1589.

51. Stephens HAF, Klaythong R, Sirikong M, Vaughn DW, Green S, Kalayanarooj S, Endy TP, Libraty DH, Nisalak A, Innis BL, Rothman AL, Ennis FA, Chandanayingyong D. 2002. HLA-A and -B allele associations with secondary dengue virus infections correlate with disease severity and the infecting viral serotype in ethnic Thais. Tissue Antigens.

52. González-Galarza FF, Takeshita LYC, Santos EJM, Kempson F, Maia MHT, da Silva ALS, André Luiz Teles, Ghattaoraya GS, Alfirevic A, Jones AR, Middleton D. 2015. Allele frequency net 2015 update: new features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations. Nucleic Acids Research.

53. Sette A, Grey H, Oseroff C, Peters B, Moutaftsi M, Crotty S, Assarsson E, Greenbaum J, Kim Y, Kolla R, Tscharke D, Koelle D, Johnson RP, Blum J, Head S, Sidney J. 2009. Definition of epitopes and antigens recognized by vaccinia specific immune responses: their conservation in variola virus sequences, and use as a model system to study complex pathogens. Vaccine 27 Suppl 6:G21–6.

54. Sibener LV, Fernandes RA, Kolawole EM, Carbone CB, Liu F, McAffee D, Birnbaum ME, Yang X, Su LF, Yu W, Dong S, Gee MH, Jude KM, Davis MM, Groves JT, Goddard WA 3rd, Heath JR, Evavold BD, Vale RD, Garcia KC. 2018. Isolation of a Structural Mechanism for Uncoupling T Cell Receptor Signaling from Peptide-MHC Binding. Cell

174:672–687.e27.

55. Gálvez J, Gálvez JJ, García-Peñarrubia P. 2019. Is TCR/pMHC Affinity a Good Estimate of the T-cell Response? An Answer Based on Predictions From 12 Phenotypic Models. Frontiers in Immunology.

56. Zehn D, Lee SY, Bevan MJ. 2009. Complete but curtailed T-cell response to very low-affinity antigen. Nature.

57. Vatti A, Monsalve DM, Pacheco Y, Chang C, Anaya J-M, Eric Gershwin M. 2017. Original antigenic sin: A comprehensive review. Journal of Autoimmunity.

58. Park MS, Kim JI, Park S, Lee I, Park M-S. 2016. Original Antigenic Sin Response to RNA Viruses and Antiviral Immunity. Immune Network.

59. Lambert P-H, Liu M, Siegrist C-A. 2005. Can successful vaccines teach us how to induce efficient protective immune responses? Nature Medicine.

60. Singh RAK, Rodgers JR, Barry MA. 2002. The Role of T Cell Antagonism and Original Antigenic Sin in Genetic Immunization. The Journal of Immunology.

61. Weiskopf D, Angelo MA, de Azeredo EL, Sidney J, Greenbaum JA, Fernando AN, Broadwater A, Kolla RV, De Silva AD, de Silva AM, Mattia KA, Doranz BJ, Grey HM, Shresta S, Peters B, Sette A. 2013. Comprehensive analysis of dengue virus-specific responses supports an HLA-linked protective role for CD8 T cells. Proceedings of the National Academy of Sciences.

62. Klenerman P, Zinkernagel RM. 1998. Original antigenic sin impairs cytotoxic T lymphocyte

responses to viruses bearing variant epitopes. Nature.

63. Rothman AL. 2011. Immunity to dengue virus: a tale of original antigenic sin and tropical cytokine storms. Nature Reviews Immunology.

64. Shen Z, Xiao Y, Kang L, Ma W, Shi L, Zhang L, Zhou Z, Yang J, Zhong J, Yang D, Guo L, Zhang G, Li H, Xu Y, Chen M, Gao Z, Wang J, Ren L, Li M. 2020. Genomic diversity of SARS-CoV-2 in Coronavirus Disease 2019 patients. Clin Infect Dis.

65. Tang X, Wu C, Li X, Song Y, Yao X, Wu X, Duan Y, Zhang H, Wang Y, Qian Z, Cui J, Lu J. 2020. On the origin and continuing evolution of SARS-CoV-2. National Science Review.

66. Hoffmann M, Kleine-Weber H, Schroeder S, Krüger N, Herrler T, Erichsen S, Schiergens TS, Herrler G, Wu N-H, Nitsche A, Müller MA, Drosten C, Pöhlmann S. 2020. SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. Cell.

67. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44:D733–45.

68. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert

M, Söding J, Thompson JD, Higgins DG. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7:539.

69. Baranov PV, Henderson CM, Anderson CB, Gesteland RF, Atkins JF, Howard MT. 2005. Programmed ribosomal frameshifting in decoding the SARS-CoV genome. Virology 332:498–510.

70. Taiaroa G, Rawlinson D, Featherstone L, Pitt M, Caly L, Druce J, Purcell D, Harty L, Tran T, Roberts J, Scott N, Catton M, Williamson D, Coin L, Duchene S. 2020. Direct RNA sequencing and early evolution of SARS-CoV-2. biorxiv;2020.03.05.976167v2. Microbiology. bioRxiv.

71. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. 2009. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics 25:1189–1191.

72. Wood MA, Paralkar M, Paralkar MP, Nguyen A, Struck AJ, Ellrott K, Margolin A, Nellore A, Thompson RF. 2018. Population-level distribution and putative immunogenicity of cancer neoepitopes. BMC Cancer 18:414.

73. Jurtz V, Paul S, Andreatta M, Marcatili P, Peters B, Nielsen M. 2017. NetMHCpan-4.0: Improved Peptide–MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. The Journal of Immunology.

74. Shao XM, Bhattacharya R, Huang J, Sivakumar IKA, Tokheim C, Zheng L, Hirsch D, Kaminow B, Omdahl A, Bonsack M, Riemer AB, Velculescu VE, Anagnostou V, Pagel KA, Karchin R. 2020. High-Throughput Prediction of MHC Class I and II Neoantigens with MHCnuggets. Cancer Immunol Res 8:396–408.

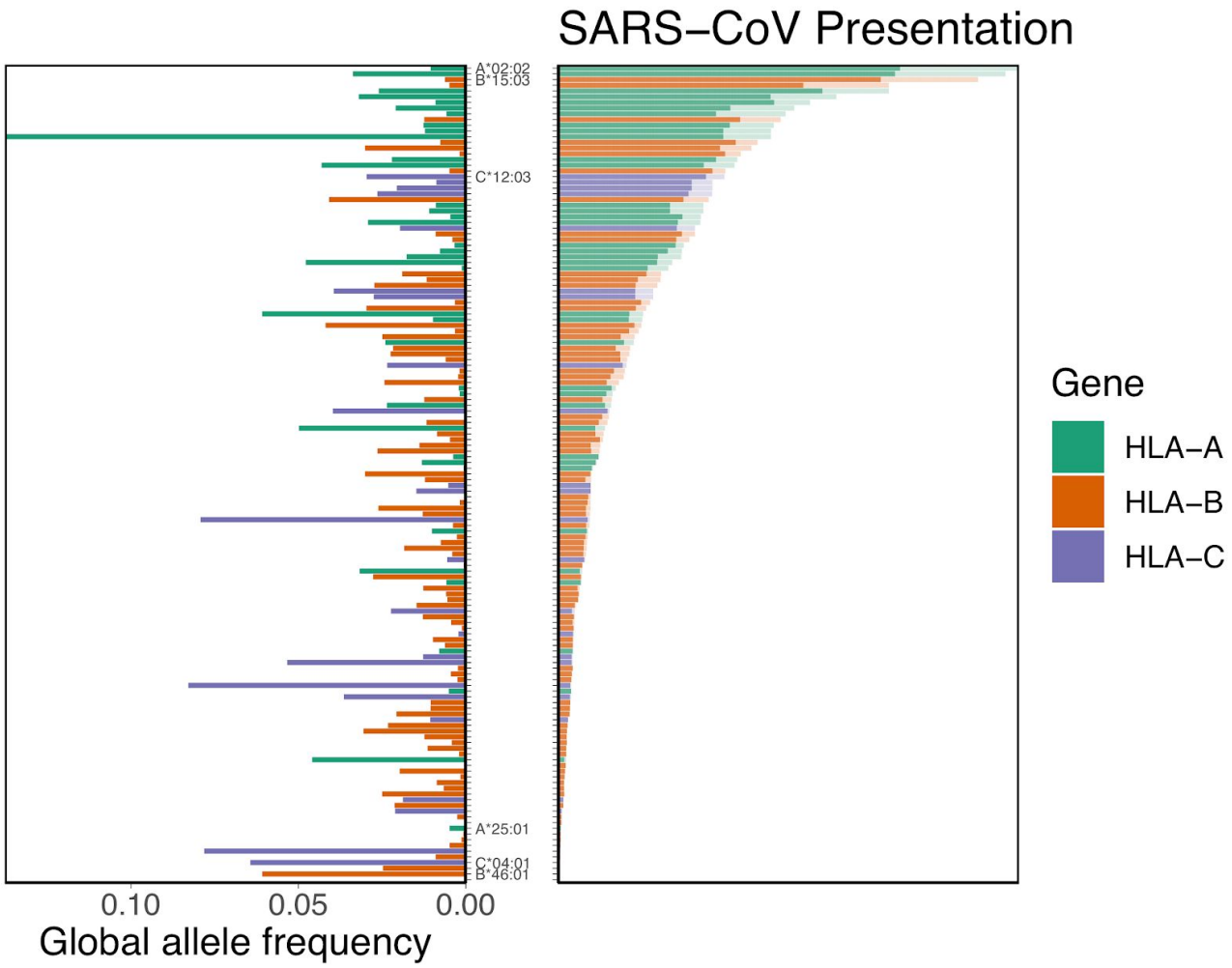
75. O'Donnell TJ, Rubinsteyn A, Bonsack M, Riemer AB, Laserson U, Hammerbacher J. 2018. MHCflurry: Open-Source Class I MHC Binding Affinity Prediction. Cell Syst 7:129–132.e4.

76. Vita R, Mahajan S, Overton JA, Dhanda SK, Martini S, Cantrell JR, Wheeler DK, Sette A, Peters B. 2019. The Immune Epitope Database (IEDB): 2018 update. Nucleic Acids Res 47:D339–D343.

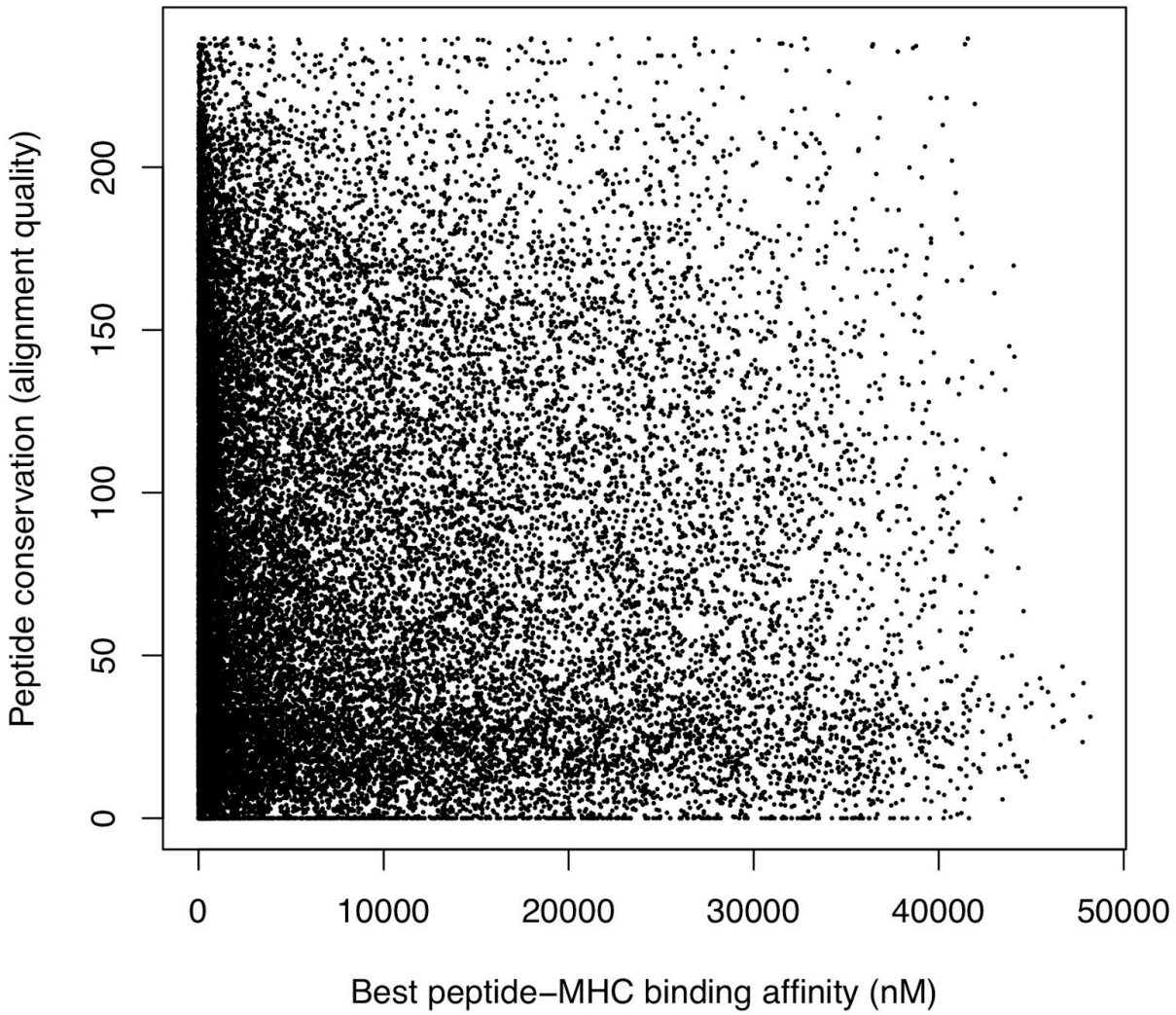
77. Nielsen M, Lundegaard C, Lund O, Keşmir C. 2005. The role of the proteasome in generating cytotoxic T-cell epitopes: insights obtained from improved predictions of proteasomal cleavage. Immunogenetics 57:33–41.

78. South A. 2011. rworldmap : a new R package for mapping global data. R J 3:35.

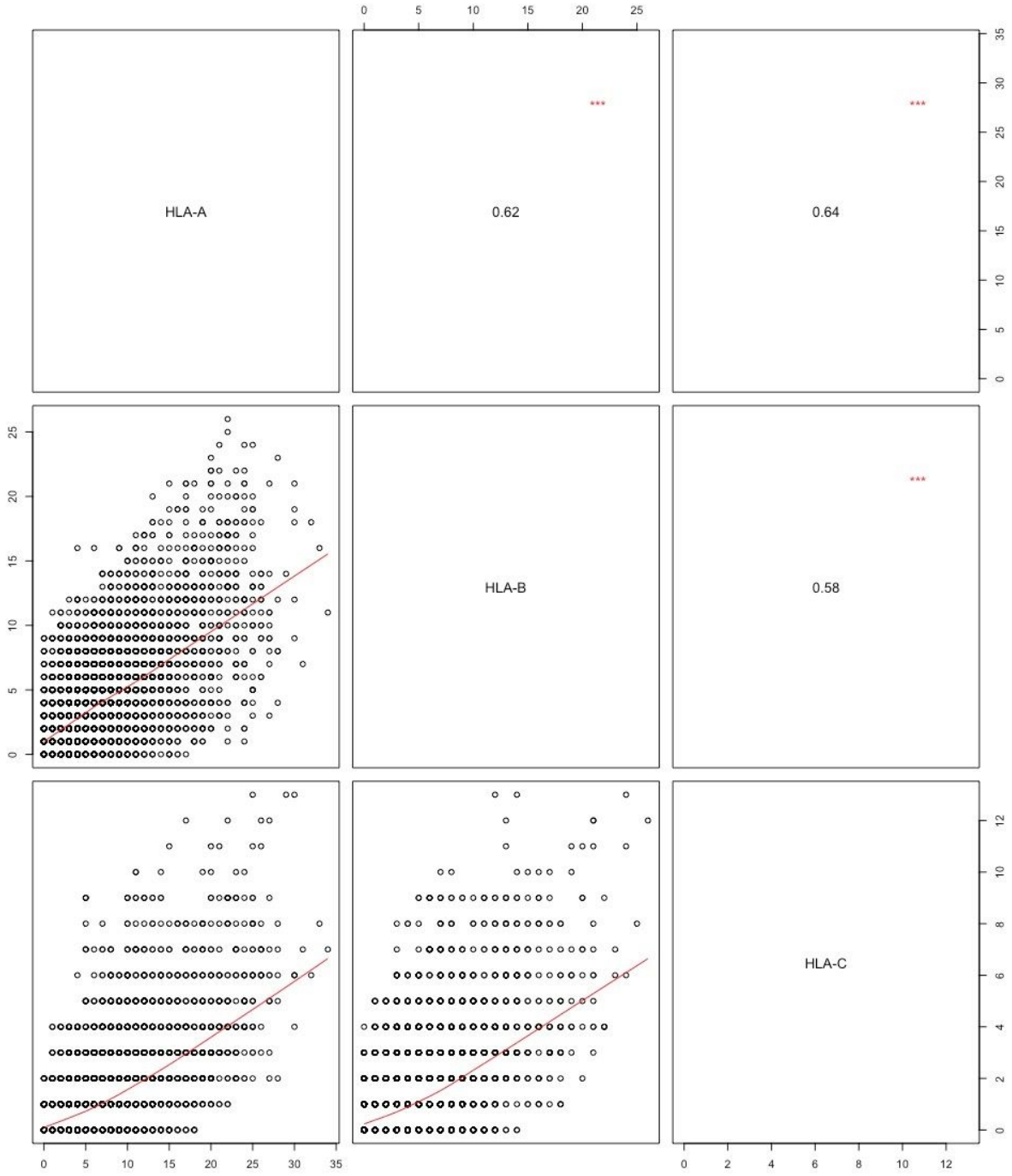
SUPPLEMENTARY DATA



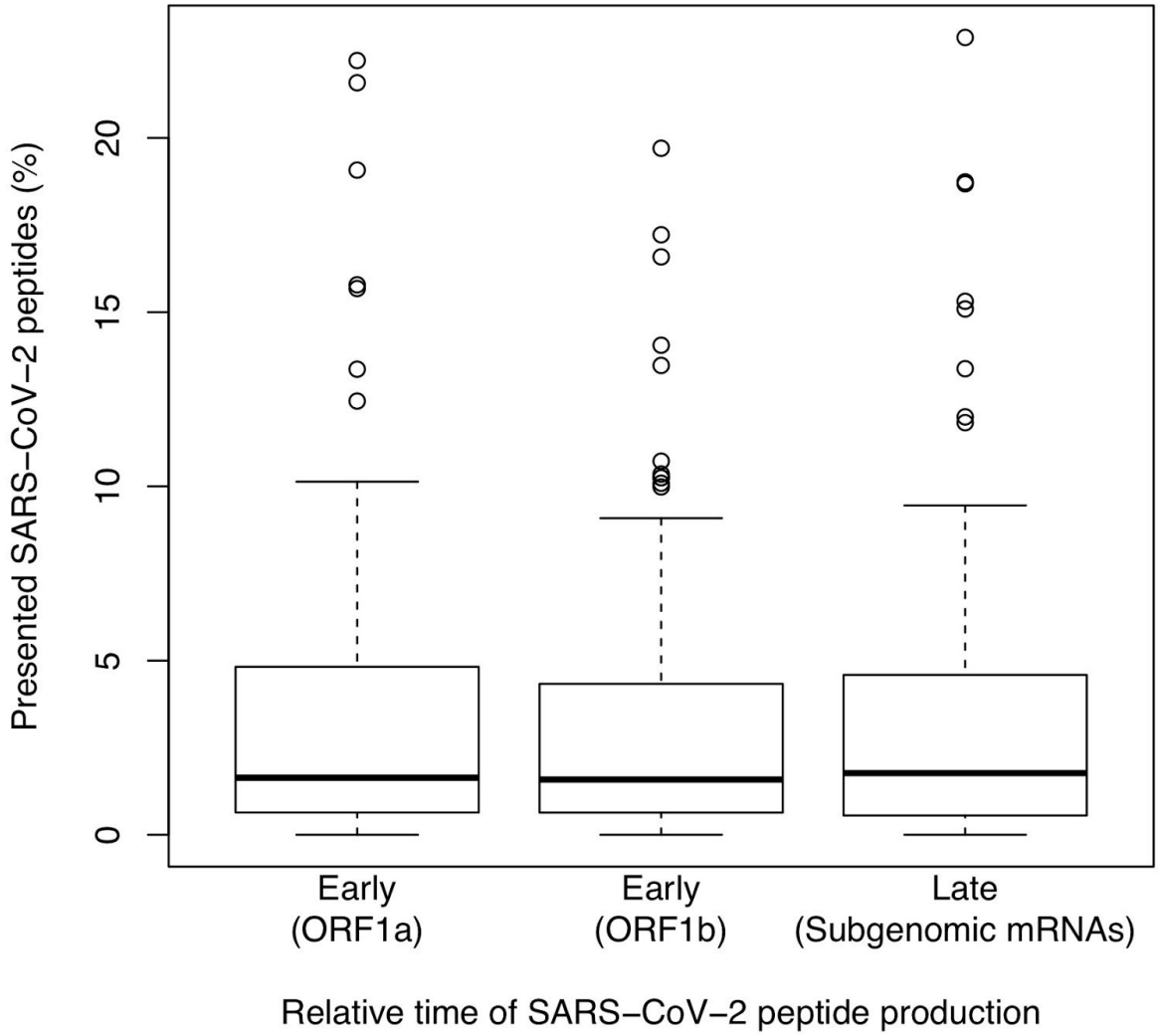
Supplementary Figure S1: Distribution of HLA allelic presentation of 8-12mers from the SARS-CoV proteome (see Supplementary Table S6). At right, the number of peptides that putatively bind to each of 145 HLA alleles is shown as a series of horizontal bars, with dark and light shading indicating the number of tightly (<50nM) and loosely (<500nM) binding peptides respectively, and with green, orange, and purple colors representing HLA-A, -B, and -C alleles, respectively. Alleles are sorted in descending order based on the number of peptides they bind (<500nM). The corresponding estimated allelic frequency in the global population is also shown (to left), with length of horizontal bar indicating absolute frequency in the population.



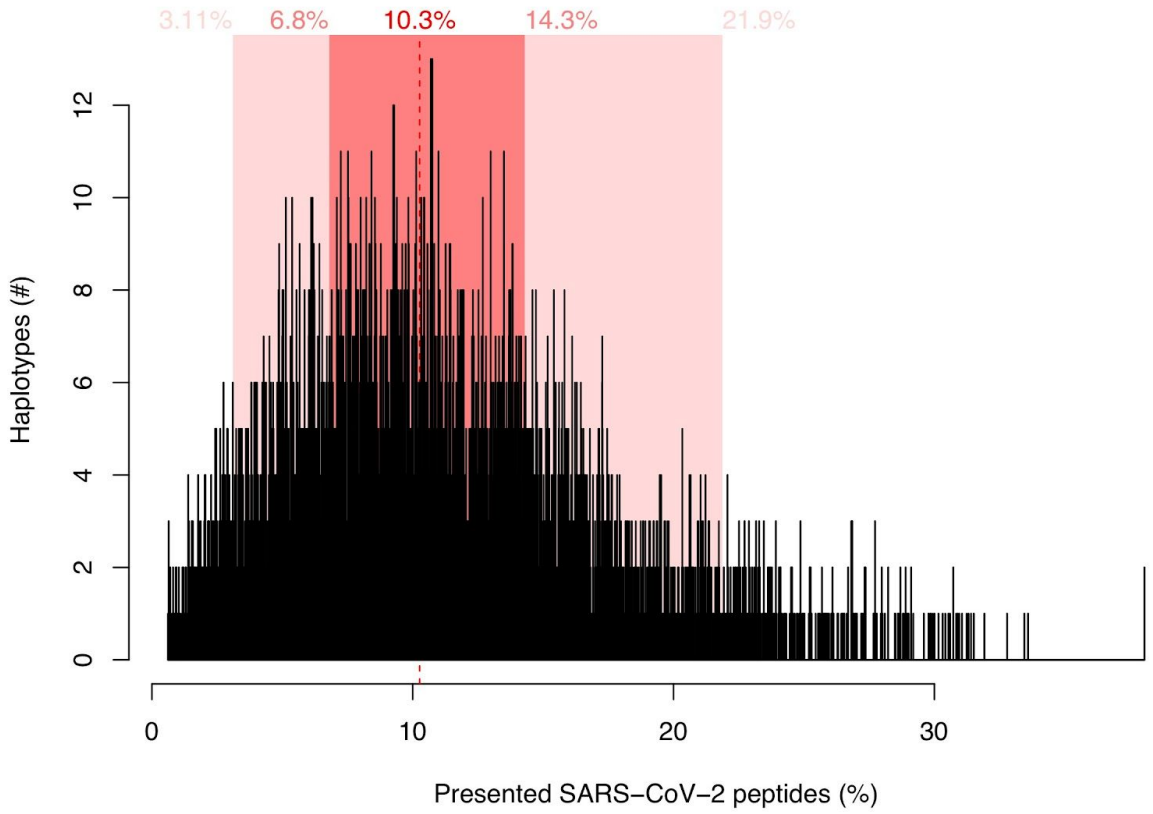
Supplementary Figure S2: Relationship between predicted peptide-MHC binding affinity and peptide conservation across coronaviruses. Every point represents a single unique peptide covering, together, the entirety of the SARS-CoV-2 proteome. The best predicted MHC binding affinity scores across 145 different HLA alleles are shown for each peptide along the x-axis. Sequence conservation (Clustal Omega alignment score) is shown for each peptide along the y-axis.



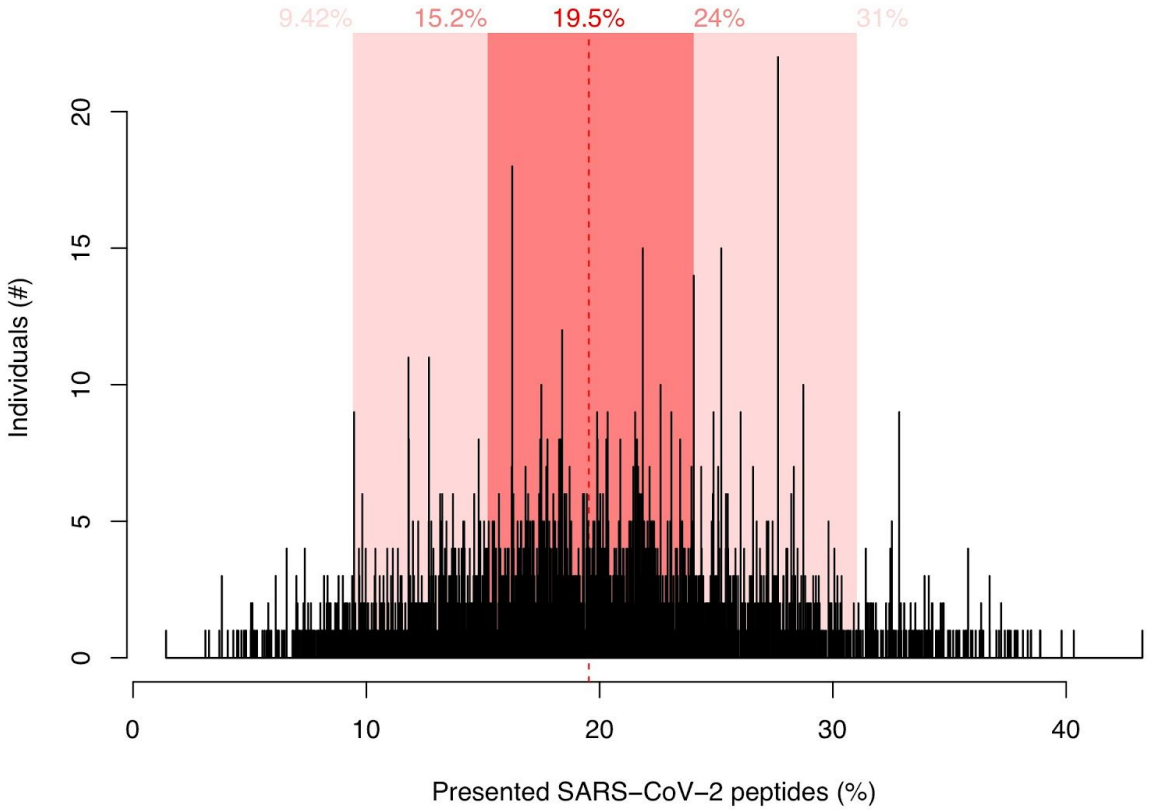
*Supplementary Figure S3: Pairwise relationship of peptide presentation between HLA-A, -B, and -C. In the bottom left three panels, every point represents the pairwise comparison of the number of peptide-allele interactions for all position coordinates. Taken together, the position coordinates cover the entirety of the SARS-CoV-2 proteome. The top right three panels show the quantitative correlation scores between each pair of HLA type comparisons (*** indicates statistical significance).*



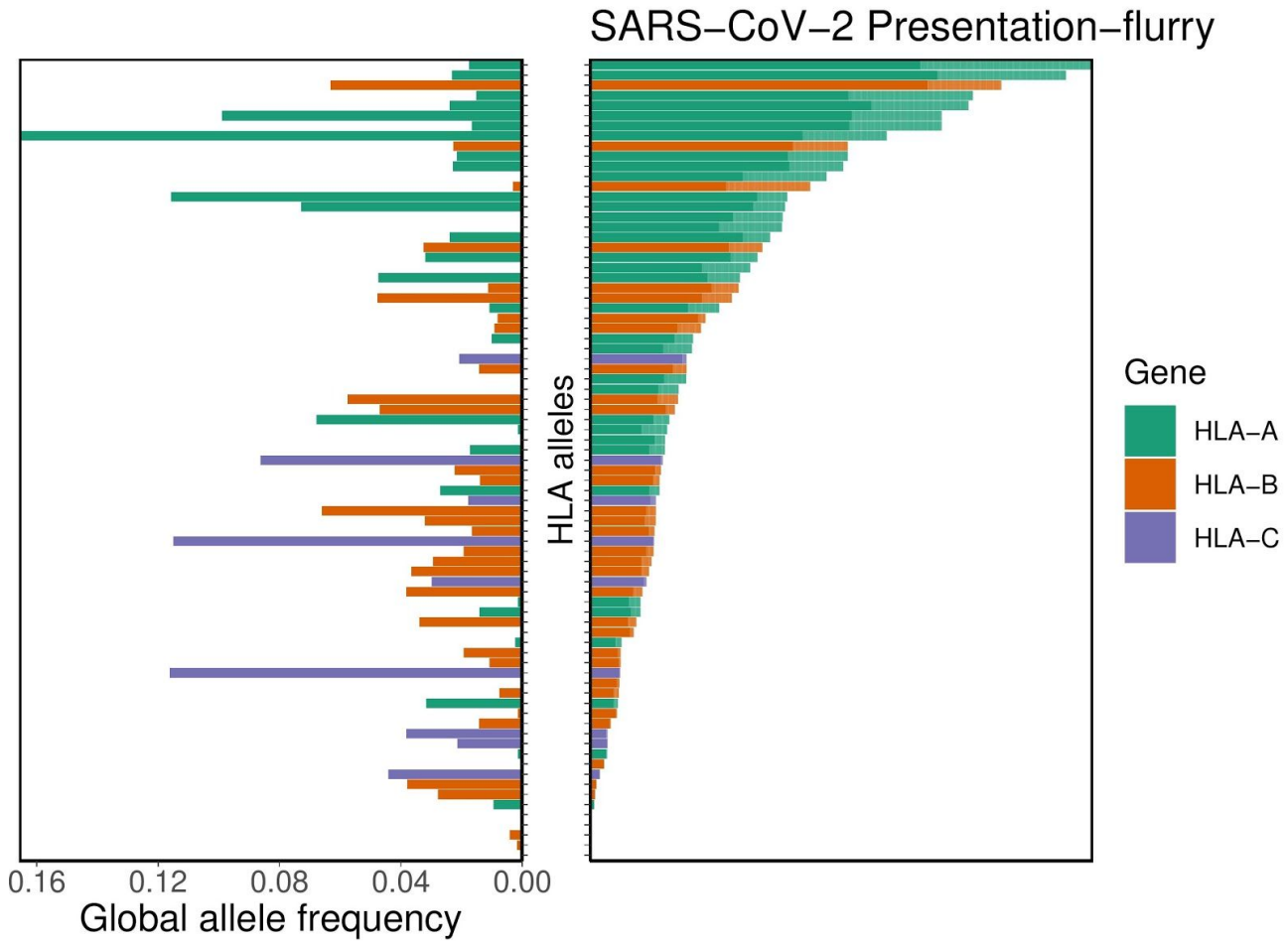
Supplementary Figure S4: Boxplot distributions of estimated epitope presentation across 145 HLA alleles for early and late SARS-CoV-2 peptides. Capacity for peptide presentation is shown along the y-axis for 145 distinct HLA alleles, for three non-overlapping sets of peptides produced at different timepoints in the viral life cycle as indicated (x-axis). Y-axis percentiles are calculated as the number of peptides from the indicated compartment of the SARS-CoV-2 proteome divided by the total number of presentable 8- to 12-mer peptides from that compartment of the proteome. Dark black lines represent median values, with boxes indicating the 25% and 75% quantiles, with whiskers representing the 25% and 75% quantiles minus or plus the interquartile range, respectively, and with additional outliers shown as open circles.



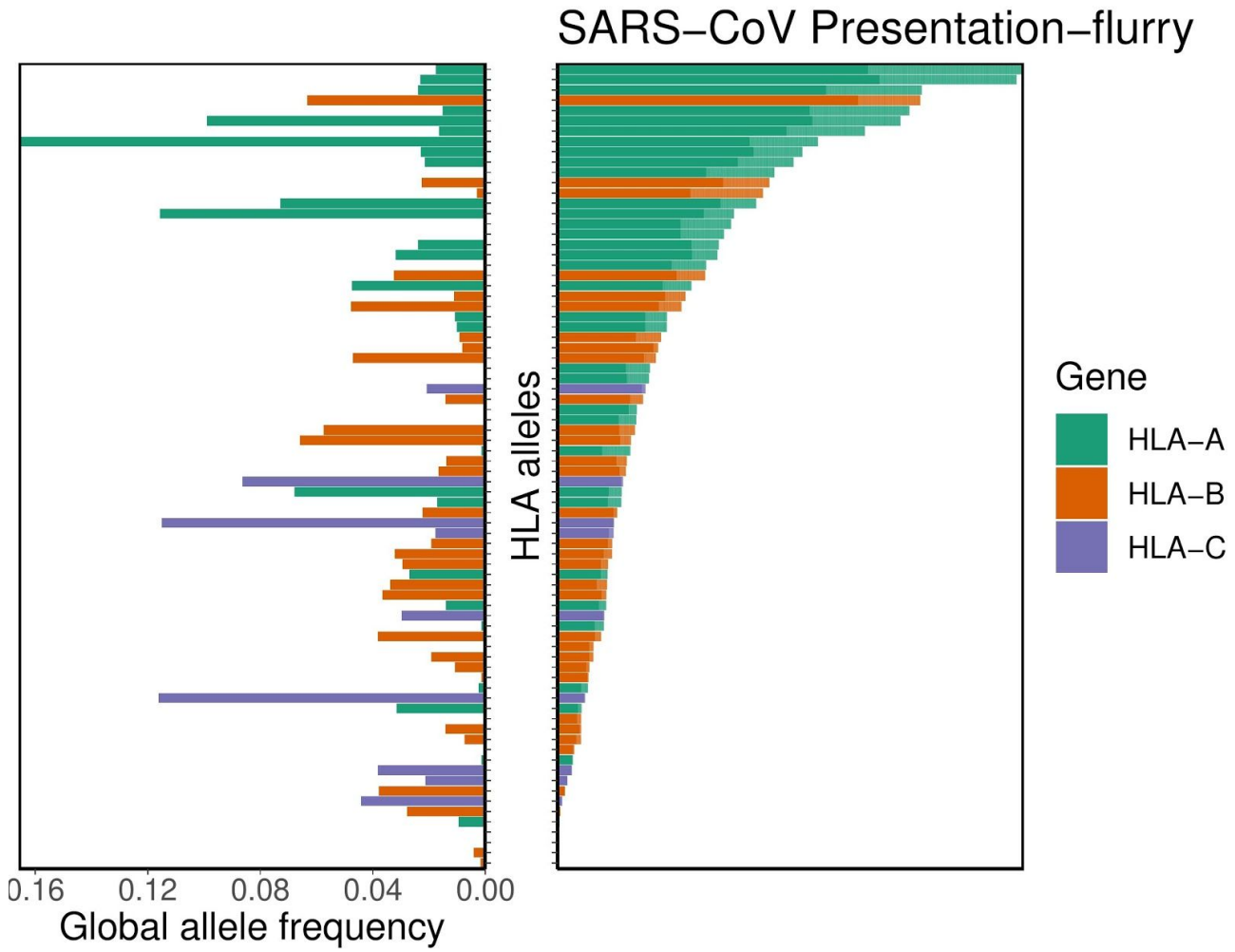
Supplementary Figure S5: Histogram of SARS-CoV-2 peptide presentation for 5,905 distinct HLA-A/B/C haplotypes. Number of haplotypes are counted along the y-axis, corresponding to their individual capacity (aggregated across all their three component HLA types) to present peptides from the SARS-CoV-2 proteome, shown along the x-axis (percentile is calculated as number of unique peptides presented divided by the total number of presentable 8- to 12-mer peptides from the SARS-CoV-2 proteome). Dashed red line corresponds to the median presentation capacity, while dark and light pink highlighted regions correspond to the 25/75% and 5/95% quantiles, respectively, with numerical values shown in the upper aspect of the plotting region.



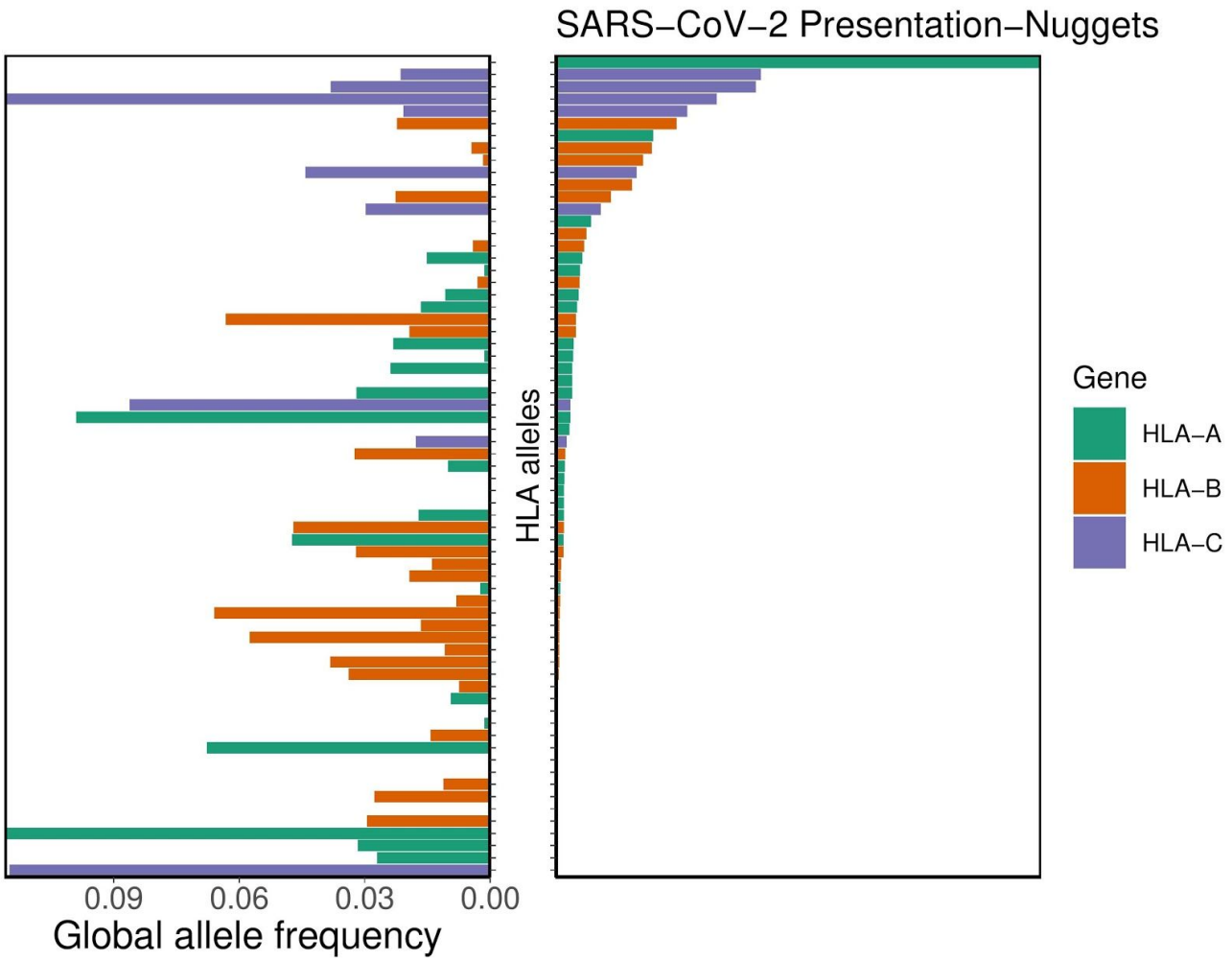
Supplementary Figure S6: Histogram of SARS-CoV-2 peptide presentation for 3,382 individuals' full HLA repertoires. Individuals are counted along the y-axis, corresponding to their individual capacity (aggregated across all 6 of their HLA types) to present peptides from the SARS-CoV-2 proteome, shown along the x-axis (percentile is calculated as number of unique peptides presented divided by the total number of presentable 8- to 12-mer peptides from the SARS-CoV-2 proteome). Dashed red line corresponds to the median presentation capacity, while dark and light pink highlighted regions correspond to the 25/75% and 5/95% quantiles, respectively, with numerical values shown in the upper aspect of the plotting region.



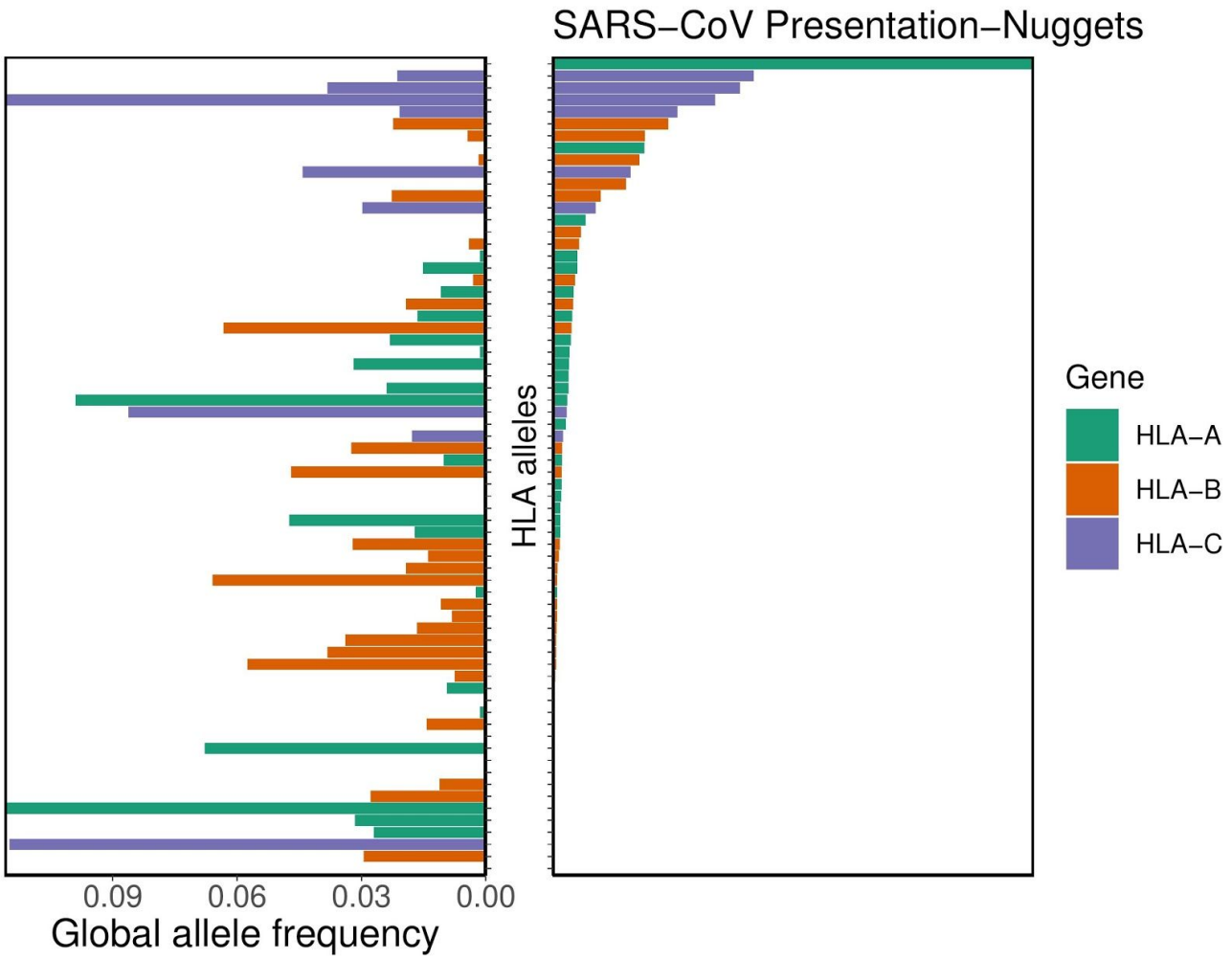
Supplementary Figure S7: Distribution of HLA allelic presentation of 8- to 12-mers from the SARS-CoV-2 proteome using the tool MHCflurry. At right, the number of peptides (see Supplementary Table S7) that putatively bind to each of 66 HLA alleles is shown as a series of horizontal bars, with dark and light shading indicating the number of tightly (<50nM) and loosely (<500nM) binding peptides respectively, and with green, orange, and purple colors representing HLA-A, -B, and -C alleles, respectively. Alleles are sorted in descending order based on the number of peptides they bind (<500nM). The corresponding estimated allelic frequency in the global population is also shown (to left), with length of horizontal bar indicating absolute frequency in the population.



Supplementary Figure S8: Distribution of HLA allelic presentation of 8- to 12-mers from the SARS-CoV proteome using the tool MHCflurry. At right, the number of peptides (see Supplementary Table S8) that putatively bind to each of 66 HLA alleles is shown as a series of horizontal bars, with dark and light shading indicating the number of tightly (<50nM) and loosely (<500nM) binding peptides respectively, and with green, orange, and purple colors representing HLA-A, -B, and -C alleles, respectively. Alleles are sorted in descending order based on the number of peptides they bind (<500nM). The corresponding estimated allelic frequency in the global population is also shown (to left), with length of horizontal bar indicating absolute frequency in the population.



Supplementary Figure S9: Distribution of HLA allelic presentation of 8- to 12-mers from the SARS-CoV-2 proteome using the tool MHCnuggets. At right, the number of peptides (see Supplementary Table S7) that putatively bind to each of 66 HLA alleles is shown as a series of horizontal bars, with dark and light shading indicating the number of tightly (<50nM) and loosely (<500nM) binding peptides respectively, and with green, orange, and purple colors representing HLA-A, -B, and -C alleles, respectively. Alleles are sorted in descending order based on the number of peptides they bind (<500nM). The corresponding estimated allelic frequency in the global population is also shown (to left), with length of horizontal bar indicating absolute frequency in the population.



Supplementary Figure S10: Distribution of HLA allelic presentation of 8- to 12-mers from the SARS-CoV proteome using the tool MHCnuggets. At right, the number of peptides (see Supplementary Table S8) that putatively bind to each of 66 HLA alleles is shown as a series of horizontal bars, with dark and light shading indicating the number of tightly (<50nM) and loosely (<500nM) binding peptides respectively, and with green, orange, and purple colors representing HLA-A, -B, and -C alleles, respectively. Alleles are sorted in descending order based on the number of peptides they bind (<500nM). The corresponding estimated allelic frequency in the global population is also shown (to left), with length of horizontal bar indicating absolute frequency in the population.