Prioritization of SARS-CoV-2 epitopes using a pan-HLA and global population inference approach

Authors

Katie M. Campbell[1,6,7,8], Gabriela Steiner[2,6], Daniel K. Wells[2], Antoni Ribas[1,2,3,4], Anusha Kalbasi[3,4,5,7]

Affiliations

[1]Department of Medicine, Division of Hematology-Oncology, University of California, Los Angeles (UCLA), Los Angeles, CA, 90095, USA.

[2]Parker Institute for Cancer Immunotherapy, San Francisco, CA, 94129, USA.

[3]Department Surgery, Division of Surgical Oncology, University of California, Los Angeles, Los Angeles, CA, USA.

[4]Jonsson Comprehensive Cancer Center, Los Angeles, CA, USA.

[5]Department of Radiation Oncology, UCLA, CA, 90095, USA.

[6]These authors contributed equally to this work.

[7]Senior author

[8]Lead Contact

Lead Contact Footnote

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Katie Campbell (KatieCampbell@mednet.ucla.edu).

Contact Information

Katie Campbell, Ph.D., Department of Medicine, Division of Hematology-Oncology, 9-666 Factor Building, 700 Tiverton Avenue, Los Angeles, CA 90095. Email: KatieCampbell@mednet.ucla.edu

Anusha Kalbasi, M.D., Department of Radiation Oncology, Jonsson Comprehensive Cancer Center (JCCC), UCLA; B-262 Factor Building, 700 Tiverton Avenue, Los Angeles, CA 90095. Phone: (310) 267-4831; Email: AnushaKalbasi@mednet.ucla.edu

Summary

SARS-CoV-2 T cell response assessment and vaccine development may benefit from an approach that considers the global landscape of the human leukocyte antigen (HLA) proteins. We predicted the binding affinity between 9-mer and 15-mer peptides from the SARS-CoV-2 peptidome for 9,360 class I and 8,445 class II HLA alleles, respectively. We identified 368,145 unique combinations of peptide-HLA complexes (pMHCs) with a predicted binding affinity less than 500nM, and observed significant overlap between class I and II predicted pMHCs. Using simulated populations derived from worldwide HLA frequency data, we identified sets of epitopes predicted in at least 90% of the population in 57 countries. We also developed a method to prioritize pMHCs for specific populations. Collectively, this public dataset and accessible user interface (Shiny app: https://rstudio-connect.parkerici.org/content/13/) can be used to explore the SARS-CoV-2 epitope landscape in the context of diverse HLA types across global populations.

Keywords

SARS-CoV-2, T cells, epitope prediction, public resource

Introduction

Infection with SARS-CoV-2 can result in a spectrum of clinical phenotypes encompassed by COVID-19, from asymptomatic illness to a potentially lethal disease with hallmarks of acute respiratory distress syndrome (ARDS) (Yang et al., 2020). While some clinical demographics have been associated with a more severe disease course (Guan et al., 2020), the heterogeneity of clinical outcomes is otherwise poorly understood.

The range of clinical outcomes may at least in part be related to patient-specific antiviral T cell responses. T cells are crucial for viral clearance and development of immunologic memory (Wherry and Ahmed, 2004) and are plausible contributors to immunopathology following viral infection (Channappanavar and Perlman, 2017). Both SARS-CoV-2 reactive CD4 and CD8 T cells have been detected in patients with COVID-19 (Chour et al., 2020; Grifoni et al., 2020a; Weiskopf et al., 2020a), though early studies suggest the relationship between T cell responses and the severity of COVID-19 is complex (Mathew et al., 2020).

The heterogeneity in T cell responses to SARS-CoV-2 may be related to recognition of viral antigens in the context of class I and II human leukocyte antigen (HLA) proteins (Chour et al., 2020). Indeed, genetic susceptibilities to viral infection have been tied to variation in the major histocompatibility complex (MHC) genes that encode HLA proteins (Dutta et al., 2018; Hill, 2001). Meanwhile, functional differences in viral antigen-specific T cell responses in symptomatic and asymptomatic patients may also contribute to the biology of at-risk populations (Mathew et al., 2020; Weiskopf et al., 2020a).

Further understanding of virus-specific T cell responses may aid in designing and monitoring the impact of preventative SARS-CoV-2 T cell vaccines. In contrast to SARS-CoV-2 vaccines focused on generating antibody responses against the surface spike glycoprotein that facilitates viral entry into the cell (Thanh Le et al., 2020) T cell vaccines have the capacity to generate immune responses against the entire viral proteome (Gilbert, 2012). In fact, non-spike T cell responses may be associated with less severe COVID19 (Peng et al., 2020).

To evaluate patient-specific T cell responses, recent studies have used large pools of SARS-CoV-2 epitopes based on homology with SARS-CoV, or based on prediction of MHC class I- and class II-binding peptides across common HLA alleles in order to capture a broad population (Grifoni et al., 2020b; Smith et al., 2020). To facilitate a more comprehensive evaluation of anti-viral and vaccine-induced T cell responses, and to support region-specific and global vaccine design strategies, we generated a resource database with a corresponding user-friendly interface to facilitate exploration of predicted MHC-binding peptides across 9,360 and 8,445 class I and II HLA alleles, to account for the genetic diversity in the MHC gene complex across global populations.

Results

In silico predictions of SARS-CoV-2 antigens

We deployed binding predictions across the SARS-CoV-2 proteome (**Figure 1**) for 9,360 class I HLA alleles (2,987 HLA-A; 3,707 HLA-B; 2,666 HLA-C; 9-mers) and 8,445 class II HLA alleles (15-mers). The predicted binding affinity (in nanomolar [nM]) between peptides and HLA proteins (pMHCs) were summarized by the median predicted binding affinity across all algorithms (Median Score). The Median Score values were filtered to those less than 500nM, a common filter used in peptide binding predictions for the purpose of identifying T cell epitopes (Rajasagi et al., 2014; Sidney et al., 1999). There were 368,145 unique combinations of peptides and HLA alleles (pMHCs) with a predicted binding affinity of less than 500nM (**Table S2**), including 1,103 unique 9-mer and 2,547 15-mer peptides and 1,022 MHC class I and 3,481 MHC class II HLA proteins, respectively. Of note, 905 9-mers (82%) were nested within 1,789 15-mers (70%), indicating that a subset of peptides are predicted as both class I and II epitopes.

In order to better understand the predicted antigenic profile of SARS-CoV-2, we focused on the set of 368,145 pMHCs with predicted binding affinity of less than 500nM for the subsequent analyses. Both class I and class II antigens were predicted across 10 of the SARS-CoV-2 genes (**Figure 1B**), with the most derived from *Orf1ab* (n=690 9-mers; 1,589 15-mers), encoding the Orf1ab polyprotein. The number of peptides from each gene correlated with protein length (R^2 = 0.997, p=2.10e-11; **Figure S1**).

Confirmation of predicted SARS-CoV-2 antigens in published datasets

In order to assess the validity of the predictions in our dataset, we compared our predicted antigens to previously reported SARS-CoV-2 or SARS-CoV T cell epitopes. There were 9 nine-mer and 5 fifteen-mer peptides in our dataset that were previously validated experimentally as T cell epitopes and reported in IEDB from SARS (**Table 1**). Since our dataset was restricted to 9-mers and 15-mers, we expanded this search to include any IEDB epitopes that overlapped (i.e. either nested, or in overlapping positions) with our predicted peptides, which resulted in 81 additional epitopes (**Table S1**). Four of these total 95 epitopes were specifically associated with HLA-A*02:01, while HLA restrictions were not reported for the remaining 91 epitopes. Each of the 154 peptides from our dataset overlapping with the 95 epitopes reported in IEDB were each predicted to bind a median of 4 class I HLA proteins (range 1-49) and 35 class II HLA proteins (range 1-5,694), suggesting these experimentally validated epitopes may be relevant in multiple HLA contexts.

Grifoni et al. recently used the homology between the SARS-CoV and SARS-CoV-2 proteomes and existing annotated epitopes of SARS-CoV from IEDB to infer T cell epitopes derived from SARS-CoV-2 (Grifoni et al., 2020b). This pool of peptides was assessed in samples derived from COVID19 patients, resulting in the identification of SARS-CoV-2-associated CD4 and CD8 T cell responses in 100% and 70% of convalescent COVID19 patients, respectively (Grifoni et al., 2020a). Our dataset identified 271 nine-mer peptides and 331 fifteen-mer peptides that either overlapped or were nested in 241 CD8 and 628 CD4 T cell epitopes from this study, derived from 9 SARS-CoV-2 genes (**Table S1**). Still, there were 793 nine-mer and 2,139 fifteen-mer peptides in our dataset not included in the megapools experimentally evaluated in this study. Including these additional peptides in experimental validation may increase the sensitivity of detection of T cell responses in patients with SARS-CoV-2.

Accounting for regional and global relevance of predicted class I pMHCs

To address the regional and global relevance of our predicted class I pMHCs, we aggregated class I HLA frequency data from the Allele Frequency Net Database (AFND) (Gonzalez-Galarza et al., 2020), representing 77 countries from 11 global regions (**Table S2**). Simulated populations (n=100,000 individuals) were created for each individual country, as well as an additional "global" population, constructed by the weighted population frequency of HLA types across countries. Each simulated individual was mapped to their predicted epitope profile by matching their HLA types to their corresponding predicted pMHCs.

Restricting our analysis to HLA alleles with at least 5% frequency in each country-genepool, we observed that the set of predicted pMHCs differed greatly across countries. Per country, there was a median of 47 (range 1-127) predicted pMHCs, including a median of 6 (range 1-11) HLA alleles and a median of 45.5 (range 1-119)

unique peptides (**Table S3**). Still, we identified 20 nine-mer peptides shared by common HLA types across 30 of 77 countries (18 of these peptides correspond to HLA types prevalent in the United States) (**Figure 2**). These peptides spanned 5 genes, including *ORF1ab* (ORF1ab polyprotein, n=14), *S* (Spike glycoprotein, n=2), *M* (membrane protein, n=1), *N* (nucleocapsid protein, n=1), and *ORF3a* (Protein 3b, n=1). Notably, this approach excluded countries in Latin America (such as Brazil and Nicaragua) and in Africa (such as Rwanda and Libya), as HLA types prevalent in these countries do not correspond to this filtered list of peptides.

To improve the global reach of a putative peptide-based vaccine, we utilized a set cover algorithm to determine the smallest set of predicted antigens that covered the maximum number of individuals in each country's population. An individual was considered "covered" if their simulated class I HLA type was involved in at least one predicted pMHC, and these sets of peptides were denoted as the set cover solutions (SCSs) for the associated population. SCSs were calculated for 77 individual countries and for a "global" population, generated by pooling together the sample populations from all countries, and sampling from this combined pool (n=100,000) without replacement (**Figure S2, Table S3**).

Based upon our simulated presentations, SCSs were capable of summarizing predicted pMHCs in at least 90% of the population in 57 countries. Furthermore, in 45 of these 57 countries, SCSs included 30 or fewer peptides (**Figure 3A**). When we evaluated which viral genes were associated with peptides in SCSs, *Orf1ab* contributed the largest number of peptides across all countries. (**Figure 3B**). We filtered peptides to those included in SCSs for at least 30 countries, and identified 19 predicted peptides, spanning 9 genes (**Figure 3C**).

The constructed SCSs were also used to prioritize peptides of interest across geographic regions. Peptides were ranked within each SCS, based upon those associated with the largest cumulative percentage of the population. Evaluating the top ten ranked peptides within each SCS (n=95 unique peptides), each was associated with a mean of 7.73 country SCSs (range 1 - 63) (**Figure 3C**). Furthermore, 19 out of 95 top-ranked peptides were associated with SCSs for countries from at least 5 out of 11 global regions (**Figure 4D**). These peptides are of particular interest, as they may be relevant across disparate populations.

We compared the SCSs established from our predicted pMHCs to SCSs generated from a "reference" set of published peptide vaccine candidates ((Grifoni et al., 2020a; Smith et al., 2020; Weiskopf et al., 2020b) based upon highly prevalent HLA types in the United States and overlapping epitopes derived from both CD4 T cell, CD8 T cell, and B cell epitope predictions. The SCSs derived from the reference peptides were relevant in at least 90% of the population across 30 countries, including the United States (**Figure 4A**). Notably, the 14 epitopes that comprised these SCSs were associated with 15 HLA types that were prevalent (at least 5% allelic frequency) across an average of 25 country populations, including HLA-C*03:04, B*35:01, A*11:01, and A*2:01 (**Figure 4B**). In contrast, the SCSs from our dataset were relevant in at least 90% of the population for 57 countries (**Figure 4C**) by including 164 additional predicted peptides associated with 823 additional HLA types (**Table S3**). Thus, the inclusion of these additional HLA types or peptides in development may broaden the global applicability of vaccines.

Deployment of a user interface to explore the epitome of SARS-CoV-2

To make the predictions generated in this study publicly available and accessible to facilitate experimental validation, we established a user interface to explore the predicted SARS-CoV-2 epitopes in this dataset (https://rstudio-connect.parkerici.org/content/13/). Predicted T cell epitopes can be filtered by features described in this study, including viral gene or protein, peptide length, peptide sequence, HLA gene or specific type, and country (population) HLA allelic frequency. Furthermore, filtered predictions are mapped to other published datasets, including those validated or reported by other groups (Grifoni et al., 2020b; Smith et al., 2020). SCSs generated for this study are also made available through this interface. This tool will serve as a resource for the development of virus specific T cell assays or vaccine design, by considering the global landscape of HLA susceptibility in SARS-CoV-2.

Discussion

Our study was designed to evaluate the predicted epitope landscape with respect to the SARS-CoV-2 viral proteome across a globally representative set of HLA alleles. We aimed to establish a resource for the scientific

community, and have made the entirety of these data publicly available and accessible. This work expands upon recent studies that inferred the epitope landscape of SARS-CoV-2 to either interrogate T cell responses in infected individuals or develop vaccines (Chour et al., 2020; Grifoni et al., 2020a, 2020b; Smith et al., 2020; Weiskopf et al., 2020a). Our pan-HLA approach enabled identification of new HLA contexts for previously proposed and validated peptides, as well as the identification of additional peptides from less prevalent HLA types. Furthermore, the overlap between class I and II predicted pMHCs suggests that some epitopes may be presented to both CD4 and CD8 T cells.

The pan-HLA approach, the inclusion of the entire SARS-CoV-2 proteome, and integration of HLA frequency data from AFND allowed unique evaluation of the regional and global relevance of our predicted pMHC dataset. We establish a set-cover based approach to explore the relevance of our predicted pMHCs across distinct global populations, and use this to construct sets of predicted pMHCs that have putative relevance across 90% of the population in 57 countries. These set cover solutions were superior using our dataset, compared to previously published datasets of peptide-based vaccine candidates, due to the breadth of predicted pMHCs and HLA subtypes.

This dataset and analysis have limitations. Our analysis was restricted to pMHC complexes with predicted binding affinities of less than 500nM. Subsequent analysis did not treat the predicted binding affinities as a continuous variable (i.e. predicted values of 5nM and 400nM were treated similarly in the remaining analysis). In the absence of experimental validation, we did not try to over delineate the association between HLA diversity and the predicted binding affinity. Furthermore, utilizing a threshold of 500nM may result in underestimating the number of alleles associated with the predicted antigenic peptides. Our predictions were limited to 9-mers and 15-mers, which represent most but not all reported HLA class I and class II binding peptides. Our data also does not account for either the quantity or timing of viral protein expression in a host cell, both of which can impact the immunogenicity of predicted epitopes (Croft et al., 2019). Finally, analysis of global population frequencies was restricted to a limited number of HLA alleles and countries. While AFND is the most comprehensive database summarizing the population frequencies of HLA haplotypes, it is far from complete. Frequencies are reported for 73, 73, and 49 countries for genes *HLA-A, -B,* and *-C*, respectively. In addition, the number of alleles reported for each gene is variable across countries, ranging from 1-1,498.

In summary, our resource provides a pan-HLA tool for those seeking to study SARS-CoV-2 or vaccine-induced T cell responses. In addition, our strategy enables the identification of sets of class I peptides either within or across countries, an important consideration for vaccine design. For these reasons we have made our calculations available in full and have also developed a user-friendly web-app to enable exploration of these data at https://rstudio-connect.parkerici.org/content/13/. SARS-CoV-2 is an ongoing pandemic and these resources will be updated as further peptide validation becomes available.

Author Contributions

K.M.C., G.S., and A.K. conceived experiments. K.M.C. and A.K. supervised the study. K.M.C., G.S., D.K.W., A.R., and A.K. designed the experiments. K.M.C. and G.S. performed data processing and analysis. K.M.C., G.S., and A.K. wrote the manuscript, and all authors contributed to final revisions of the manuscript.

Acknowledgments

We are grateful to John Wherry (University of Pennsylvania, Philadelphia, PA) and Bonaventura Clotet, Julia Garcia Prado and Christian Brander (IrsiCaixa Foundation, Barcelona, Spain) for valuable feedback on the manuscript. The computational resources for this study were provided by the Parker Institute for Cancer Immunotherapy (PICI). K.M.C. is supported by the UCLA Tumor Immunology Training Grant (NIH T32CA009120) and the Cancer Research Institute (CRI) Irvington Postdoctoral Fellowship Program. A.K. is supported by the UCLA CTSI KL2 Award (NCATS TR001882) and Sarcoma Alliance for Research Through Collaboration Career Enhancement Program. A.R. is supported by R35 CA197633 and The Ressler Family Fund, and is a member researcher at PICI.

Declaration of Interests

K.M.C is a shareholder in Geneoscopy LLC. D.K.W. is a founder, equity holder and receives consulting fees from Immunai. A.R. is supported by the National Institute of Health (R35 CA197633), the Ressler Family Fund, the Agilent Thought Leader Award, a Stand Up to Cancer- Bristol-Meyer Squibb Catalyst Research Grant (Grant Number: SU2C-AACR-CT06-17). This research grant is administered by the American Association for Cancer Research, the scientific partner of SU2C. A.R. is a member researcher at the Parker Institute for Cancer Immunotherapy.

Figures and Figure Legends

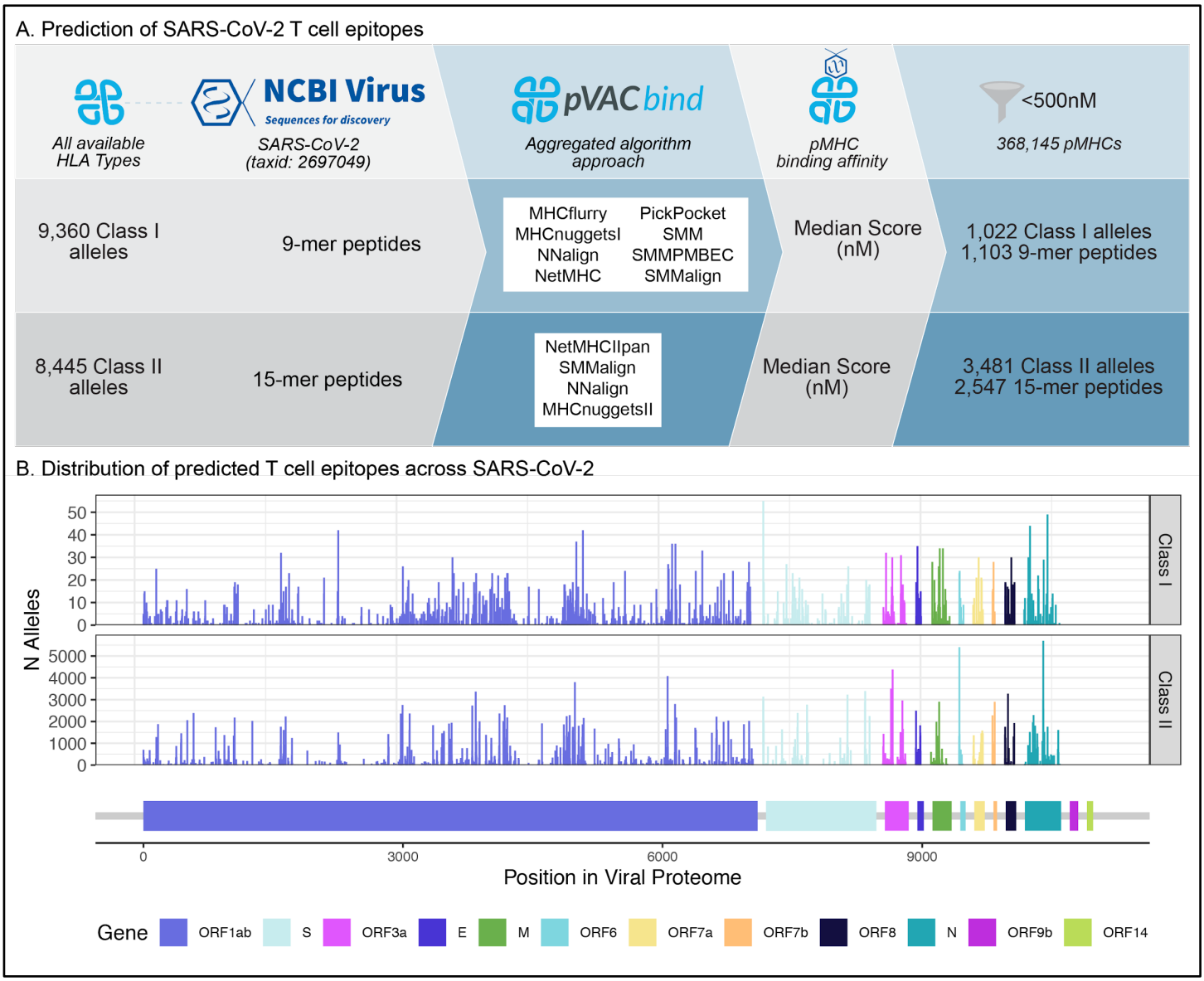

Figure 1. Peptide binding predictions for SARS-CoV-2

A. Overview of the analysis strategy. Class I and II HLA alleles, combined with 9-mer and 15-mer peptides spanning the viral proteome were used as inputs for an aggregated peptide binding prediction approach. A filter of peptides with a median score of 500nM was applied to summarise a set of peptide-MHC complexes (pMHCs) with predicted high binding affinity. B. The distribution of the number of HLA alleles (distinguished by Class I vs II) is shown, according to corresponding peptide, indicated by its starting position within the viral proteome (x-axis). Peptides are colored by their corresponding genes.

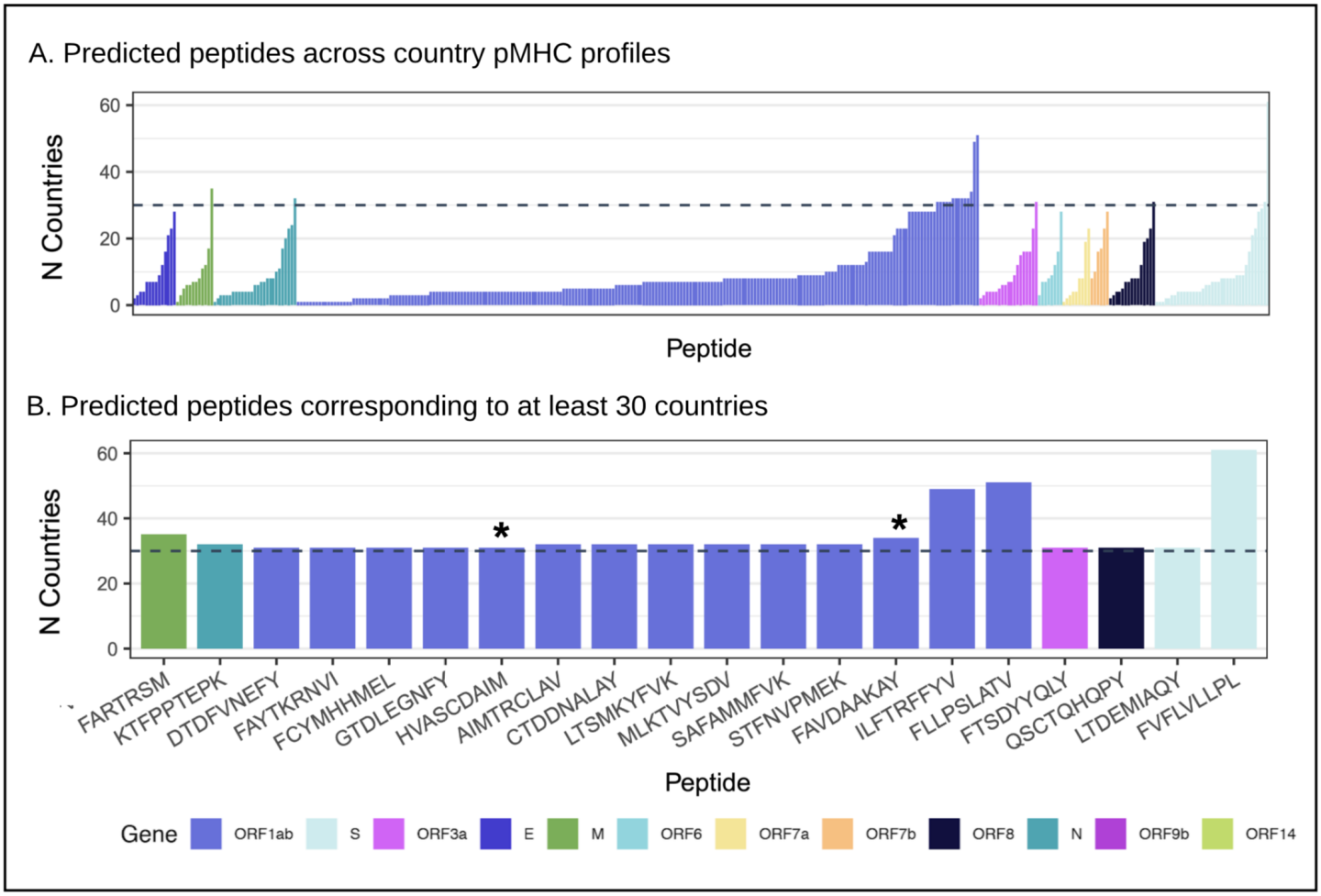

Figure 2. Peptide diversity in country pMHC profiles

Overview of country pMHC profiles, reflecting HLA frequency distributions reported by AFND. Frequency data was filtered to only include alleles with at least 5% frequency for each country. The y-axis indicates the number of country pMHC profiles that included each peptide along the x-axis. Two groups of peptides are shown, according to corresponding SARS-CoV-2 gene: A) peptides that appeared at least once in any country pMHC profile, and B) those that appeared in a minimum of 30 country pMHC profiles. (*) indicates that the peptide was not included in the pMHC profile of the United States.

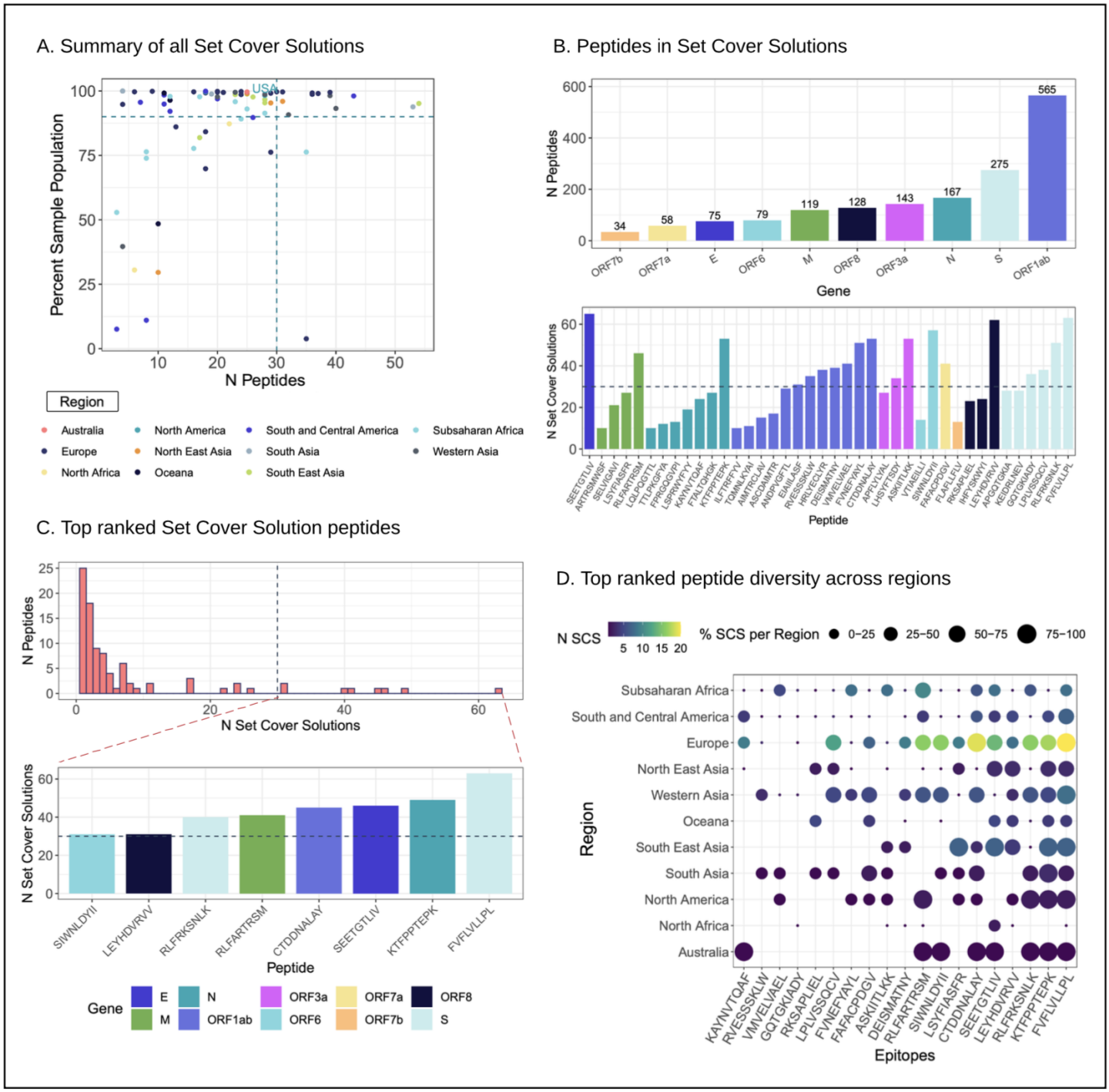

Figure 3. Set Cover Solution peptide summary

A. Summary of SCS results for all 77 countries. The percent of the sample population covered and the number of peptides involved in each SCS is shown, annotated by each country's corresponding region. The United States is also denoted by text. B. Overview of peptides comprising SCSs. The number of peptides each SARS-Cov-2 protein contributed is shown (top), as well as the number of SCSs individual peptides contributed to (bottom), filtered to show peptides that contributed to a minimum of 10 SCSs. C. Peptides were ranked within each SCS, based upon those associated with the largest cumulative percentage of the population. There were 95 unique peptides comprising the top 10 ranking of all SCSs, and are shown in this figure. The histogram (top panel) shows the number of SCSs associated with each of these peptides. The most recurrent peptides (present in over 30 SCSs) are further shown (bottom panel). D. Geographic distribution of top-ranked peptides, shown in C. Peptides were filtered to those associated with country SCSs spanning at least 5 regions are shown. Each tile

is colored by the number of SCSs (i.e. countries) within each global region (y-axis) corresponding to each peptide (x-axis).

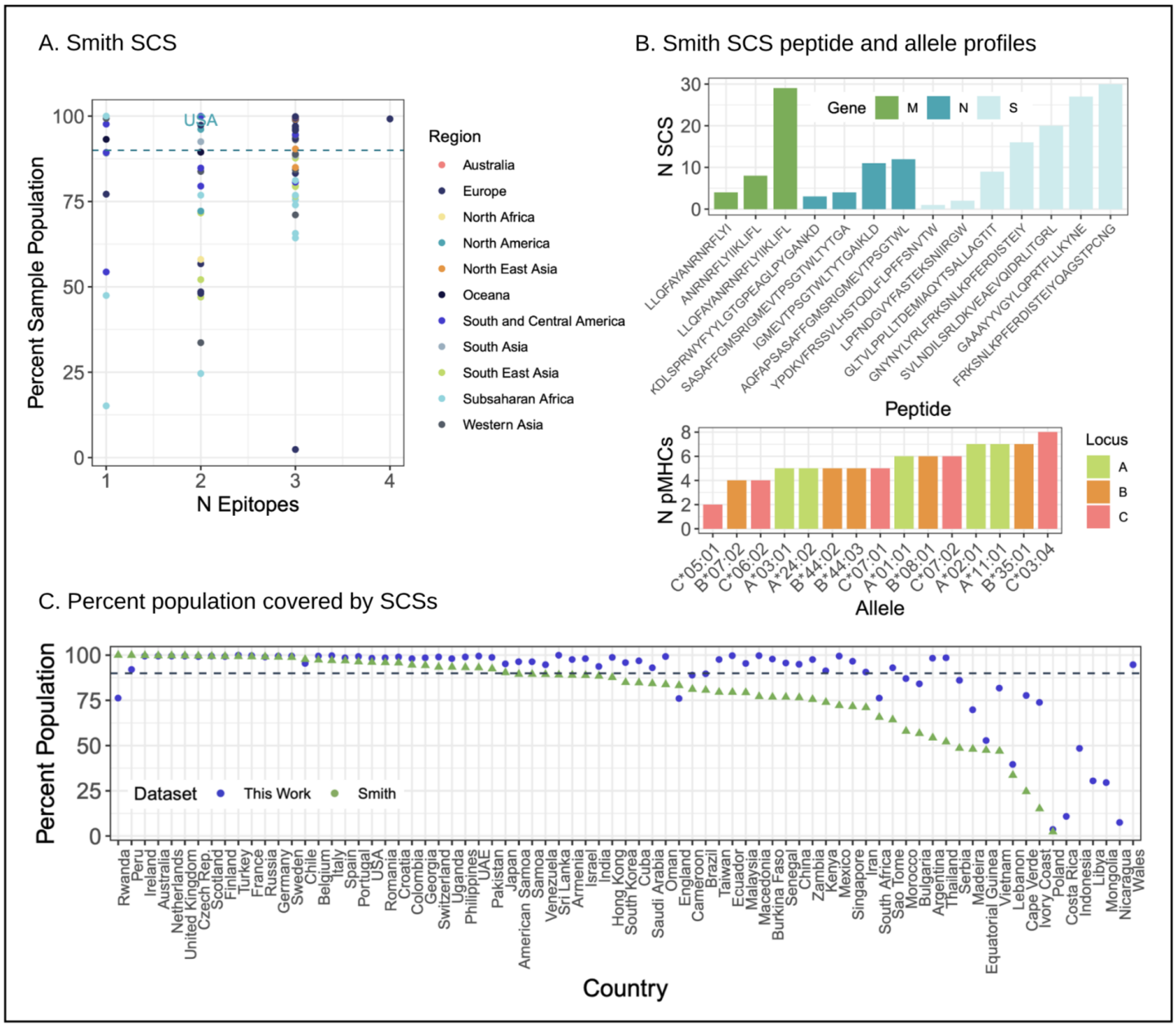

Figure 4. Smith et al. Set Cover Solution summary

A. The Smith et al. dataset was summarized across countries by previously reported epitopes and corresponding HLA types. The number of peptides included and the percent of the sample population covered by each SCS is shown, colored by each country's corresponding region. The United States is also denoted by text. B. The peptides comprising the Smith SCS are shown (top panel), along with the number of SCSs (y-axis) associated with each peptide and the corresponding SARS-CoV-2 gene (color). The HLA Alleles associated with these peptides is also shown (bottom), according to the number of pMHCs (i.e. peptides) predicted for the corresponding allele. C. The percent of each simulated country (x-axis) population covered by their respective SCSs is shown; SCS from this work is shown in blue, and the SCS results from Smith et al. is shown in green. A 90% cutoff is denoted by the horizontal, dashed line.

Tables

Table 1. Previously validated T cell epitopes in SARS-CoV from IEDB

*HLA Restriction and Experimental validation are taken from the annotation reported in IEDB. ICS: intracellular cytokine staining.

Peptide	Gene	HLA Restriction*	Predicted HLA types	T cell assays (Experimental validation)*
ALNTPKDHI	N	A*02:01	HLA-A*02:11	ELISPOT IFNg release
AQFAPSASAFFGMSR	N	HLA class II	DQA1*06:01, 06:02; DQB1*03:13, 06:03, 06:11, 06:14, 06:28, 06:31, 06:40, 06:41, 06:44	ELISA or ICS IFNg release
FIAGLIAIV	S	A2; A*02:01	A*02:03, 02:131, 02:150, 02:170, 02:179, 02:187, 02:196, 02:205, 02:214, 02:228, 02:238, 02:248, 02:257, 02:50, 02:69, 02:71, 02:85, 02:95	ELISA, ELISPOT, or ICS IFNg release; in vivo assay cytotoxicity; multimer/tetramer qualitative binding
GMSRIGMEV	N	A*02:01	A*02:03, 02:50	51 chromium or in vitro cytotoxicity; ELISA, ELISPOT, or ICS IFNg release
ILLNKHIDAYKTFPP	N	Mus musculus (BALB/c)	DPA1*02:02; DPB1*05:01	ELISPOT IFNg release
LLLDRLNQL	N	A*02:01	A*02:02, 02:03, 02:11, 02:13, 02:132, 02:141, 02:150, 02:16, 02:173, 02:181, 02:19, 02:196, 02:205, 02:214, 02:228, 02:238, 02:25, 02:262, 02:54, 02:70, 02:71, 02:73, 02:85, 02:95; B*08:22, 08:38, 08:41, 08:56; C*03:71	51 chromium or in vitro cytotoxicity; ELISA, ELISPOT, or ICS; in vivo pathogen burden after challenge
LPNNTASWFTALTQH	N	Mus musculus (BALB/c)	DQA1*01:01, 01:02, 01:03, 01:04, 01:05, 01:06, 01:07, 01:08, 01:09, 02:01; DQB1*03:01, 03:03, 03:04, 03:07, 03:08, 03:09, 03:10, 03:11, 03:12, 03:14, 03:15, 03:16, 03:17, 03:18, 03:19, 03:20, 03:21, 03:22, 03:23, 03:24, 03:26, 03:27, 03:28, 03:29, 03:30, 03:31, 03:32, 03:33, 03:34, 03:35, 03:36, 03:37, 03:38, 03:06, 03:13, 03:25, 04:03, 06:01, 06:14, 06:15, 06:16, 06:19, 06:32, 06:33, 06:35, 06:37, 06:43, 06:02, 06:04, 06:07, 06:24, 06:03, 06:09, 06:11, 06:22, 06:28, 06:29, 06:30, 06:31, 06:40, 06:41, 06:44; DRB1*14:01, 14:07, 14:10, 14:11, 14:16, 14:24, 14:26, 14:39, 14:45, 14:50, 14:54, 14:55, 14:58, 14:60, 14:68, 14:71, 14:75, 14:76, 14:79, 14:82, 14:86, 14:87, 14:88, 14:90, 14:93, 14:97	ELISA IFNg release
LQLPQGTTL	N	A*02:01	A*02:06, 02:14; B*15:01, 15:03, 15:103, 15:113, 15:127, 15:132, 15:179, 15:62,	ELISPOT IFNg release

			15:69, 15:75, 15:98, 39:23, 39:49, 40:07, 40:12, 40:13, 40:21, 40:46, 48:15, 48:21	
NLNESLIDL	S	A*02:01	A*02:02, 02:131, 02:141, 02:155, 02:16, 02:186, 02:19, 02:209, 02:22, 02:69, 02:90	51 chromium cytotoxicity; ELISPOT IFNg release
RLNEVAKNL	S	A*02:01	A*02:03, 02:11, 02:128, 02:171, 02:196, 02:230, 02:238, 02:253, 02:258, 02:99; B*27:20	51 chromium cytotoxicity; ELISPOT or ICS IFNg release; multimer/tetramer qualitative binding
SASAFFGMSRIGMEV	N	Mus musculus (BALB/c)	DRB1*11:04	ELISA IFNg release
SPRWYFYYLGTGPEA	N	Mus musculus (BALB/c; H2-d class II)	DPA1*01:03, DPB1*03:01, 14:01, 140:01, 141:01, 142:01, 143:01, 144:01, 145:01, 147:01, 148:01, 149:01; DRB1*04:03, 04:04, 13:03"	ELISA or ELISPOT IFNg release; ELISPOT IL-10 release; ELISPOT IL-2 release; ELISPOT IL-4 release
VLAWLYAAV	Orf1ab	A*02:01	A*02:11, 02:148, 02:22, 02:230, 02:253, 02:258	ICS IFNg release
VLNDILSRL	S	A*02:01	A*02:03, 02:11, 02:13, 02:132, 02:148, 02:151, 02:171, 02:186, 02:19, 02:196, 02:209, 02:22, 02:230, 02:238, 02:253, 02:258, 02:52, 02:54, 02:70, 02:71, 02:73, 02:85, 02:99; C*05:04, 05:23, 05:33	51 chromium cytotoxicity; ELISPOT IFNg release

STAR METHODS

Resource Availability

Lead Contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Katie Campbell (katiecampbell@mednet.ucla.edu).

Materials Availability

Data and Code Availability

The results of this study are available in a public Google bucket through the following link: https://console.cloud.google.com/storage/browser/pici-covid19-data-resources (gs://pici-covid19-data-resources). This bucket contains all of the unfiltered peptide binding predictions and the Supplemental Tables corresponding to this document. The filtered peptide binding predictions and set cover solutions can be explored using the interactive Shiny app at https://rstudio-connect.parkerici.org/content/13/. All filtered data can be exported from this web interface. The code for this manuscript and the Shiny App is available in the public github repository https://github.com/kcampbel/neocovid-app.

Method Details

Data acquisition

The NCBI Virus resource (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/) was used to obtain all annotated protein sequences for SARS-CoV-2 on March 15, 2020. This dataset spanned 166 genotypes, but protein

sequences were summarized by the corresponding UniProt annotation for the SARS-CoV-2 proteome (UP000464024). All possible 9-mer and 15-mer peptides were obtained from the entire viral proteome for Class I and Class II peptide binding predictions, respectively.

Peptide-MHC binding predictions

Peptide binding predictions were performed using the pVACbind tool from pVACtools and executed using the griffithlab/pvactools:1.5.7 (https://hub.docker.com/r/griffithlab/pvactools/) Docker image. Class I prediction algorithms included MHCflurry (v1.6.0) (O'Donnell et al., 2018), MHCnuggets (v2.3) (Shao et al., 2020), NetMHC (v4.0) (Andreatta and Nielsen, 2016), PickPocket (v1.1) (Zhang et al., 2009), SMM (v1.0) (Peters and Sette, 2005), and SMMPMBEC (v1.0) (Kim et al., 2009). Class II prediction algorithms included NetMHCIIpan (v4.0) (Reynisson et al., 2020), SMMalign (v1.1) (Nielsen et al., 2007), NNalign (v2.3) (Nielsen and Andreatta, 2017), and MHCnuggets (v2.3) (Shao et al., 2020).

HLA alleles were chosen by running the command `pvacseq valid_alleles` and filtering out any non-expressed or null HLA alleles (those ending with the "N" suffix), resulting in 9,360 Class I HLA proteins and 8,445 Class II HLA alleles. It is important to note that for Class II predictions, some algorithms include inputs of either individual HLA alleles (e.g. DPB1*01:01) or combinations of HLA alleles (e.g. DPA1*01:03-DPB1*01:01), since two HLA proteins pair together for Class II antigen presentation. All available individual (n=3,484) or combinations of (n=4,961) Class II HLA alleles were used for input.

Class I predictions were performed using the following command:

```
$ /opt/conda/bin/pvacbind/run ${fasta} ${hla} ${hla} MHCflurry MHCnuggetsI NetMHC PickPocket SMM SMMPMBEC tmp/ -e 9 --iedb-install-directory /opt/iedb --net-chop-method cterm --netmhc-stab
```

Class II predictions were performed using the following command:

```
$ /opt/conda/bin/pvacbind/run ${fasta} ${hla} ${hla} NetMHCIIpan SMMalign NNalign MHCnuggetsII tmp/ -e 15 --iedb-install-directory /opt/iedb --net-chop-method cterm --netmhc-stab
```

Where ${fasta} was the protein sequence fasta containing the SARS-CoV-2 proteome, obtained at NCBI Virus. The pVACbind tool was performed individually across the union of HLA alleles available for all algorithms, and each allele was specified by the ${hla} input in the command. The filtered results, containing peptide-MHC complexes with predicted Median Score (nM) less than 500nM, were aggregated for the final dataset.

Population frequencies of HLA types

Country Populations

Population frequencies of HLA alleles were obtained from the Allele Frequency Net Database (Gonzalez-Galarza et al., 2020). The database contains HLA Frequency data for Class I and Class II alleles across 1,028 distinct populations. Populations whose net frequency data exceeded 1 at a given allele were excluded from this analysis, as well as populations that did not report frequencies for alleles with 2 or 3 fields. Because these populations are highly granular (i.e. "USA San Francisco Caucasian"), we aggregated them into 98 populations by country; 77 had data for Class I alleles. This was done by (1) assigning each population to a country using the first word from each population name, (2) calculating the country "sample size" by summing the sample sizes of distinct populations, and (3) calculating HLA frequencies within each country population using **Formula 1.**

Formula 1. Country HLA Frequency Calculation

Country Frequency = sum(`Allele Frequency``Population Sample Size`)/(Country Sample Size)*

Sample populations were generated from this country-frequency data for 77 countries. This was done by sampling 2n alleles from each country gene pool for each Class I allele with reported data, where n represents the number of simulated individuals in the sample population (n = 100,000). The probability of selecting an allele for each sample population is equal to the frequency of that allele reported for each given country. The simulated populations used for analysis are available in the github repository (https://github.com/kcampbel/neocovid-app). These populations of simulated genotypes were then merged with

our Class I predictions to create a pMHC profile for each country, based on each country's reported allele frequencies.

Global Population

A simulated "global population" was generated by first aggregating all 77 country sample populations, and then sampling from this pool 2n times to create a sample population of n individuals (n = 100,000). This ensured that each country would be represented in this global population with equal probability regardless of sample size, such that the global population was not further biased towards the United States and European countries. It should be noted that consequently, this global population does not reflect true global HLA frequencies, which would require consideration of true country size.

Set Cover Solutions

Given a universal set of n elements (U), a collection of subsets of U (S), and the associated cost of each subset in S, the set cover problem is to identify I , the minimal subcollection of S, whose union equates to U and minimizes the total cost (Karp, 1972). The greedy algorithm addresses this problem by iteratively adding elements of U to I until all subsets in S are covered (Vazirani, 2013). This problem is NP-hard, so a logN approximate solution was used.

That is, for a simulated population X,

U = all individuals in X covered by at least one pMHC

S_i = all individuals in X covered by the pMHC with Epitope i

S = set of S_i whose union spans all individuals in U

cost(S_i) = 1 for all Epitopes i

The solution (I) represents the smallest set of epitopes whose union covers the largest portion of population X.

Quantification and Statistical Analysis

Data analysis and visualization was performed in R using the tidyverse packages [REF]. The neoCOVID Explorer application was developed using the Shiny R package [REF] and deployed using RStudio-Connect.

Key Resources Table

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
SARS-CoV-2 viral proteome sequence	NCBI Virus	taxid:2697049
SARS T cell epitopes	IEDB	taxid:694009
T cell epitopes shared by SARS and SARS-CoV-2	(Grifoni et al., 2020b)	Tables S3, S6
SARS-CoV-2 vaccine candidates	(Smith et al., 2020)	Table S6
Population frequencies of HLA types	Allele Frequency Net Database	http://www.allelefrequencies.net
Software and Algorithms		
pVACtools v1.5.7	https://pvactools.readthedocs.io/	https://hub.docker.com/r/griffithlab/pvactools/
MHCflurry (v1.6.0)	(O'Donnell et al., 2018)	

MHCnuggets (v2.3)	(Shao et al., 2020)	
NetMHC (v4.0)	(Andreatta and Nielsen, 2016)	
PickPocket (v1.1)	(Zhang et al., 2009)	
SMM (v1.0)	(Peters and Sette, 2005)	
SMMPMBEC (v1.0)	(Kim et al., 2009)	
NetMHCIIpan (v4.0)	(Reynisson et al., 2020)	
SMMalign (v1.1)	(Nielsen et al., 2007),	
NNalign (v2.3)	(Nielsen and Andreatta, 2017)	
R v4.0.0	https://cran.r-project.org/	
tidyverse v1.3.0	https://www.tidyverse.org/	
Shiny v1.4.0.2	https://shiny.rstudio.com/	
Set Cover Solution	This study	https://github.com/kcampbel/neocovid-app
neoCOVID Explorer Shiny App	This study	https://rstudio-connect.parkerici.org/content/13/

REFERENCES

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

Channappanavar, R., and Perlman, S. (2017). Pathogenic human coronavirus infections: causes and consequences of cytokine storm and immunopathology. Semin. Immunopathol. *39*, 529–539.

Chour, W., Xu, A.M., Ng, A.H.C., Choi, J., Xie, J., Yuan, D., Lee, J.K., Delucia, D.C., Edmark, R., Jones, L., et al. (2020). Shared Antigen-specific CD8+ T cell Responses Against the SARS-COV-2 Spike Protein in HLA A*02:01 COVID-19 Participants (medRxiv).

Croft, N.P., Smith, S.A., Pickering, J., Sidney, J., Peters, B., Faridi, P., Witney, M.J., Sebastian, P., Flesch, I.E.A., Heading, S.L., et al. (2019). Most viral peptides displayed by class I MHC on infected cells are immunogenic. Proc. Natl. Acad. Sci. U. S. A. *116*, 3112–3117.

Dutta, M., Dutta, P., Medhi, S., Borkakoty, B., and Biswas, D. (2018). Polymorphism of HLA class I and class II alleles in influenza A(H1N1)pdm09 virus infected population of Assam, Northeast India. J. Med. Virol. *90*, 854–860.

Gilbert, S.C. (2012). T-cell-inducing vaccines - what's the future. Immunology *135*, 19–26.

Gonzalez-Galarza, F.F., McCabe, A., Santos, E.J.M.D., Jones, J., Takeshita, L., Ortega-Rivera, N.D., Cid-Pavon, G.M.D., Ramsbottom, K., Ghattaoraya, G., Alfirevic, A., et al. (2020). Allele frequency net database (AFND) 2020 update: gold-standard data classification, open access genotype data and new query tools. Nucleic Acids Res. *48*, D783–D788.

Grifoni, A., Weiskopf, D., Ramirez, S.I., Mateus, J., Dan, J.M., Moderbacher, C.R., Rawlings, S.A., Sutherland, A., Premkumar, L., Jadi, R.S., et al. (2020a). Targets of T cell responses to SARS-CoV-2 coronavirus in humans with COVID-19 disease and unexposed individuals. Cell.

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

Guan, W.-J., Ni, Z.-Y., Hu, Y., Liang, W.-H., Ou, C.-Q., He, J.-X., Liu, L., Shan, H., Lei, C.-L., Hui, D.S.C., et al. (2020). Clinical Characteristics of Coronavirus Disease 2019 in China. N. Engl. J. Med. *382*, 1708–1720.

Hill, A.V. (2001). Immunogenetics and genomics. Lancet *357*, 2037–2041.

Karp, R.M. (1972). Reducibility among Combinatorial Problems. In Complexity of Computer Computations: Proceedings of a Symposium on the Complexity of Computer Computations, Held March 20–22, 1972, at the IBM Thomas J. Watson Research Center, Yorktown Heights, New York, and Sponsored by the Office of Naval Research, Mathematics Program, IBM World Trade Corporation, and the IBM Research Mathematical Sciences Department, R.E. Miller, J.W. Thatcher, and J.D. Bohlinger, eds. (Boston, MA: Springer US), pp. 85–103.

Kim, Y., Sidney, J., Pinilla, C., Sette, A., and Peters, B. (2009). Derivation of an amino acid similarity matrix for peptide: MHC binding and its application as a Bayesian prior. BMC Bioinformatics *10*, 394.

Mathew, D., Giles, J.R., Baxter, A.E., Greenplate, A.R., Wu, J.E., Alanio, C., Oldridge, D.A., Kuri-Cervantes, L., Betina Pampena, M., D'Andrea, K., et al. (2020). Deep immune profiling of COVID-19 patients reveals patient heterogeneity and distinct immunotypes with implications for therapeutic interventions.

Nielsen, M., and Andreatta, M. (2017). NNAlign: a platform to construct and evaluate artificial neural network

models of receptor–ligand interactions. Nucleic Acids Res. *45*, W344–W349.

Nielsen, M., Lundegaard, C., and Lund, O. (2007). Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. BMC Bioinformatics *8*, 238.

O'Donnell, T.J., Rubinsteyn, A., Bonsack, M., Riemer, A.B., Laserson, U., and Hammerbacher, J. (2018). MHCflurry: Open-Source Class I MHC Binding Affinity Prediction. Cell Syst *7*, 129–132.e4.

Peng, Y., Mentzer, A.J., Liu, G., Yao, X., Yin, Z., Dong, D., Dejnirattisai, W., Rostron, T., Supasa, P., Liu, C., et al. (2020). Broad and strong memory CD4 + and CD8 + T cells induced by SARS-CoV-2 in UK convalescent COVID-19 patients. bioRxiv.

Peters, B., and Sette, A. (2005). Generating quantitative models describing the sequence specificity of biological processes with the stabilized matrix method. BMC Bioinformatics *6*, 132.

Rajasagi, M., Shukla, S.A., Fritsch, E.F., Keskin, D.B., DeLuca, D., Carmona, E., Zhang, W., Sougnez, C., Cibulskis, K., Sidney, J., et al. (2014). Systematic identification of personal tumor-specific neoantigens in chronic lymphocytic leukemia. Blood *124*, 453–462.

Reynisson, B., Alvarez, B., Paul, S., Peters, B., and Nielsen, M. (2020). NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. Nucleic Acids Res. *48*, W449–W454.

Shao, X.M., Bhattacharya, R., Huang, J., Sivakumar, I.K.A., Tokheim, C., Zheng, L., Hirsch, D., Kaminow, B., Omdahl, A., Bonsack, M., et al. (2020). High-Throughput Prediction of MHC Class I and II Neoantigens with MHCnuggets. Cancer Immunol Res *8*, 396–408.

Sidney, J., Southwood, S., Oseroff, C., Guercio, M., Sette, A., and Grey, H.M. (1999). Measurement of MHC/Peptide Interactions by Gel Filtration. Curr. Protoc. Immunol. *31*, 26.1.

Smith, C.C., Entwistle, S., Willis, C., Vensko, S., Beck, W., Garness, J., Sambade, M., Routh, E., Olsen, K., Kodysh, J., et al. (2020). Landscape and Selection of Vaccine Epitopes in SARS-CoV-2.

Thanh Le, T., Andreadakis, Z., Kumar, A., Gómez Román, R., Tollefsen, S., Saville, M., and Mayhew, S. (2020). The COVID-19 vaccine development landscape. Nat. Rev. Drug Discov. *19*, 305–306.

Vazirani, V.V. (2013). Approximation Algorithms (Springer Science & Business Media).

Weiskopf, D., Schmitz, K.S., Raadsen, M.P., Grifoni, A., Okba, N.M.A., Endeman, H., van den Akker, J.P.C., Molenkamp, R., Koopmans, M.P.G., van Gorp, E.C.M., et al. (2020a). Phenotype and kinetics of SARS-CoV-2-specific T cells in COVID-19 patients with acute respiratory distress syndrome. Science Immunology *5*.

Weiskopf, D., Schmitz, K.S., Raadsen, M.P., Grifoni, A., Okba, N.M.A., Endeman, H., van den Akker, J.P.C., Molenkamp, R., Koopmans, M.P.G., van Gorp, E.C.M., et al. (2020b). Phenotype of SARS-CoV-2-specific T-cells in COVID-19 patients with acute respiratory distress syndrome (medRxiv).

Wherry, E.J., and Ahmed, R. (2004). Memory CD8 T-cell differentiation during viral infection. J. Virol. *78*, 5535–5545.

Yang, X., Yu, Y., Xu, J., Shu, H., Xia, J. 'an, Liu, H., Wu, Y., Zhang, L., Yu, Z., Fang, M., et al. (2020). Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study. Lancet Respir Med *8*, 475–481.

Zhang, H., Lund, O., and Nielsen, M. (2009). The PickPocket method for predicting binding specificities for

receptors based on receptor pocket similarities: application to MHC-peptide binding. Bioinformatics *25*, 1293–1299.