TOTAL PREDICTED MHC-I EPITOPE LOAD IS INVERSELY ASSOCIATED WITH MORTALITY FROM SARS-COV-2

Eric A. Wilson
School of Molecular Sciences
Arizona State University

Gabrielle Herneise
School of Life Sciences
Arizona State University

Abhishek Singharoy*
School of Molecular Sciences
Arizona State University

Karen S. Anderson*
Biodesign Institute
Arizona State University

May 9, 2020

ABSTRACT

Polymorphism in MHC-I protein sequences across human populations significantly impacts their binding to viral peptides and alters T cell immunity to infection. Prioritization of MHC-I restricted viral epitopes remains a fundamental challenge for understanding adaptive immunity to SARS-CoV-2. Here, we present a consensus MHC-I binding prediction model, EnsembleMHC, based on the biochemical and structural basis of peptide presentation to aid the discovery of SARS-CoV-2 MHC-I peptides. We performed immunopeptidome predictions of SARS-CoV-2 proteins across 52 common MHC-I alleles identifying 658 high confidence peptides. Analysis of the resulting peptide-allele assignment distribution demonstrated significant variation across the allele panel up to an order of magnitude. Using MHC-I population-based allele frequencies, we estimated the average SARS-CoV-2 peptide population binding capacity across 21 individual countries. We have discovered a strong inverse association between the predicted population SARS-CoV-2 peptide binding capacity and overall mortality. Furthermore, we found that the consideration of only structural proteins produced a stronger association with observed death rate, highlighting their importance in protein-targeted immune responses. The 108 predicted SARS-CoV-2 structural protein peptides were shown to be derived from enriched regions in the originating protein, and present minimal risk for disruption by mutation. These results suggest that the immunologic fitness of both individuals and populations to generate class I-restricted T cell immunity to SARS-CoV-2 infection may impact clinical outcome from viral infection.

Keywords SARS-CoV-2 · EnsembleMHC · MHC-I · risk-model

*corresponding author

1 Introduction

In December 2019, the novel coronavirus, SARS-CoV-2 was identified from a cluster of cases of pneumonia in Wuhan, China[1,2]. With over 3.5 million cases and over 250,000 deaths, the viral spread has been declared a global pandemic by the World Health Organization[3]. Prior to the emergence of SARS-CoV-2, there had been six strains of coronaviruses detected in humans including severe acute respiratory syndrome coronavirus (SARS) and Middle East respiratory syndrome (MERS)[4]. The rapid emergence of SARS-CoV-2 is in part due to a long asymptomatic incubation period preceding observable symptoms[5–7]. Symptoms common to each of the human-inhabited coronaviruses include shortness of breath, cough, and fever, progressing in some individuals to pneumonia or acute respiratory distress syndrome (ARDS)[8,9]. Due to its high rate of transmission and the current lack of a vaccine, there is an immediate demand for information on T cell immunity to SARS-CoV-2 both for monitoring the infection and developing immunotherapies.

Coronaviruses (CoVs) are enveloped positive-sense, single-stranded RNA viruses inhabiting birds and mammals[10]. SARS, MERS, and SARS-CoV-2 all fall into the betacoronavirus Clade[11,12]. These viruses encode an abnormally large RNA genome (approx. 30 kb) with a high recombination rate, as shown by the differentiation of the L-type and the S-type SARS-CoV-2 strains[13]. Coronaviruses are made up of four primary structural proteins (spike, envelope, membrane, and nucleocapsid), eight accessory proteins, and 13-16 nonstructural proteins[4,14]. The spike protein is responsible for binding the virus to the host receptor binding domains (RBDs) and fusing the membrane of the virus and the host cell, while the envelope and membrane proteins interact to form a virus particle for replication and eventual egress from the host cell[15]. Different strains of coronavirus have spike proteins that bind to different RBDs on the host cell – SARS-CoV-1 and SARS-CoV-2 binds to the ACE2 domain while MERS binds to the DPP4 domain[16,17]. It remains unknown how these different RBDs dictate pathogenicity for coronavirus strains, but it is known that neutralizing antibodies prevent the spike protein from binding to the host RBD. The nucleocapsid protein is multi-functional – it maintains the structure of the virion by containing the RNA within the nucleocapsid it encodes, it can suppress RNA silencing in mammalian cells, and it dictates the formation of the replicase complex within the host cell[18,19].

SARs-CoV-2 strains have 88% sequence homology with SARS and SARS-like viruses (namely bat-SL-CoVZC45 and bat-SL-CoVZXC21), with approximately 380 amino acid substitutions[4,20]. This homology includes 27 amino acid substitutions in the spike protein, which could impact binding to host RBD[4]. Although no substitutions were noted in the RBD motif responsible for binding to the ACE2 receptor, six mutations, of unknown significance, are present in other parts of the RBD sequence[4]. Less is known on how these substitutions impact T cell immunity; however, several T cell epitopes in the structural proteins are conserved between SARS-CoV-1 and SARS-CoV-2, offering potential therapeutic targets for a vaccine[21].

There is limited information on immunogenic MHC-I restricted T cell epitopes for SARS-CoV-2, but existing studies have characterized the immunogenicity of peptides derived from SARS-CoV and MERS-CoV. For SARS-CoV, immunogenic T cell epitopes have been identified in the S (Spike), N (Nucleocapsid), M (Membrane), and E (Envelope) protein following the 2002-03 outbreak[22]. The majority of these immunogenic targets were HLA-A2 restricted $CD8^+$ T cell epitopes located in the spike protein, with additional epitopes being identified in the nucleocapsid protein[23,24]. It is generally considered that epitopes in the M and E protein are less immunogenic and in lower frequency than that of the S and N protein[25], however studies have been limited. Peptides derived from the non-structural polyprotein 1a have been used to generate IFN-γ-producing memory $CD8^+$ T cells from patients with SARS-CoV, and immunogenic epitopes from other non-structural or accessory proteins have been investigated as vaccine targets[26,27].

In this study, we developed a novel consensus MHC-I binding and processing prediction workflow called EnsembleMHC. This prediction workflow integrates seven different prediction algorithms that have been parameterized on high quality mass spectrometry data. Additionally, by calculating the underlying precision of each component algorithm, we are able to assign confidence levels to the identified peptides, a feature currently unavailable in other prediction platforms. We apply this workflow to predict all potential 8-14mer SARS-CoV-2 peptides within a set of 52 common MHC-I alleles, and characterize the disparity in the peptide allele assignment distribution. The identified allele-peptide distribution was then used to assign a country-wide EnsembleMHC population score, a metric based on the predicted total population SARS-CoV-2 presentation capacity as a function of endemic allele frequency. We observe a strong inverse correlation of the EnsembleMHC population score and observed mortality rate, suggesting that population fitness towards SARS-CoV-2 may be shaped by overall presentation of SARS-CoV-2 peptides to the $CD8^+$ T cells. We identified that the correlation between EnsembleMHC population score and death rate may be particular sensitive to the presentation of SARS-CoV-2 structural proteins. Accordingly, we identify 108 peptides derived from SARS-CoV-2 structural proteins that are potential high value targets for T cell vaccine development, based on their predictive binding, expression, and sequence conservation in isolates.

2 Methods

2.1 EnsembleMHC prediction workflow

EnsembleMHC source binding and processing prediction algorithms

EnsembleMHC incorporates MHC-I binding and processing predictions from 7 publicly available prediction algorithms: MHCflurry-affinity[28], MHCflurry-presentation[28], netMHC-4.0[29], netMHCpan-4.0-EL[30], netMHCstabpan[31], PickPocket[32], and MixMHCpred[33]. These algorithms were chosen based on the criteria of providing a free academic license, bash command line integration, and demonstrated accuracy for predicting experimentally validated SARS-CoV-2 MHC-I peptides[34].

Each of the selected algorithms cover components of MHC-I binding and antigen processing that roughly fall into two categories: ones based primarily on MHC-I Binding affinity predictions and others that model antigen presentation. To this end, MHCflurry-affinity, netMHC-4.0, PickPocket, and netMHCstabpan predict binding affinity based on quantitative peptide binding affinity measurements. netMHCstabpan also incorporates peptide-MHC stability measurements and PickPocket performs prediction based on binding pocket structural extrapolation. To model the effects of antigen presentation, MixMHCpred, netMHCpan-EL, and MHCflurry-presentation are trained on naturally eluted MHC-I ligands. Additionally, MHCflurry-presentation incorporates an antigen processing term.

Parameterization of EnsembleMHC using mass spectrometry data

EnsembleMHC combines multiple disparate MHC-I binding and processing algorithms in order to improve accuracy and confidence of peptide calls unattainable by the use of any single method. Such integration is accomplished by parameterizing allele and algorithm specific score thresholds based on observed performance on a comprehensive and high-quality mass spectrometry (MS) dataset of naturally presented MHC-I peptides[35]. This particular dataset was selected as it is the largest single laboratory MS-based characterization of MHC-I peptides derived from monoallelic cell lines. This aspect significantly reduces the number of artifacts introduced by differences in peptide isolation methods, mass spectrometry acquisition, and convolution of peptides in multiallelic cell lines. An overview of the EnsembleMHC parameterization is provided in supplemental figures **(SI A1)**.

Fifty-two common MHC-I alleles were selected for parameterization based on the criteria that they were characterized in *Sarkizova et al.* dataset and that all 7 prediction methods could be applied to that allele. Each target peptide (observed in the MS dataset) was paired with 100 length-matched randomly sampled decoy peptides (not observed in the MS dataset) derived from the same source proteins, minimizing the incidence of false negatives in the testing set. If a protein was less than 100 amino acids in length, then every potential peptide from that protein was extracted. Subsequently, Each of the seven algorithms were then independently applied to each of the 52 allele datasets. For each allele dataset, the score threshold required for a particular algorithm to achieve 50% recall of target peptides (the point at which 50% of the target peptides were identified within the selected threshold score) was recorded. Additionally, the expected accuracy of each algorithm at each allele was assessed by calculating the observed false detection rate or FDR (the fraction of identified peptides that were decoy peptides) for each algorithm using its allele specific scoring threshold. The parameterization process was repeated 1000 times for each allele through bootstrap sampling of half of the peptides in the total dataset. The final FDR and score threshold for each algorithm at each allele was determined by taking the median value of the bootstrap sampling.

Application of EnsembleMHC for the prediction of SARS-CoV-2 MHC-I peptides

MHC-I peptide predictions for the SARS-CoV-2 proteome were performed using the reference sequence MN908947.3 (https://www.ncbi.nlm.nih.gov/genbank/sars-cov-2-seqs/). All potential 8-14mer peptides (n= 67,207) were derived from the open reading frames in the reported proteome, and each peptide was evaluated by the EnsembleMHC workflow. All peptides are initially filtered by the criteria of needing to fall within the allele specific score threshold for at least one algorithm. The remaining peptides are then aggregated, and the confidence level of each peptide call, $peptide^{FDR}$, is calculated as the product of the predetermined allele specific FDRs for each of the algorithms that detected a given peptide. This relationship is given by equation 1,

$$peptide^{FDR} = \prod_{i=1, i \neq ND}^{N} algorithm_i^{FDR} \tag{1}$$

, where N is the number of MHC-I binding and processing algorithms, ND represents an algorithm that did not detect a given peptide, and $algorithm^{FDR}$ represents the allele specific FDR of the Nth algorithm.

The $peptide^{FDR}$ represents the joint probability that all of the MHC-I binding and processing algorithms that detected a particular peptide did so in error, and therefore returns a probability of false detection. Peptides that were assigned a false detection probability of less than or equal to 5% were selected for inclusion in the predicted peptide set. An overview of the ap-

plication of EnsembleMHC for the prediction of SARS-CoV-2 peptides is shown in **Figure 1A**.

SARS-CoV-2 polymorphism analysis and Protein structure visualizations

Polymorphism analysis of SARS-CoV-2 structural proteins were performed using 4,455 full length protein sequences obtained from the National Center for Biotechnology Information Virus database[36]. Proteins were visualized in VMD[37] using the solved structures for the E (5x29) and S (6vxx) proteins (http://www.rcsb.org/)[38] and predicted structures for the M and N proteins[39].

2.2 Application of EnsembleMHC to determine population fitness against SARS-CoV-2

The peptides identified by the EnsembleMHC workflow were used to assess the fitness of a given population against the SARS-CoV-2 virus by considering the observed SARS-CoV-2 binding capacity of a given MHC-I allele as function of regional expression of that alleles within a given population. This workflow is summarized in **SI A2**.

Population-wide MHC-I frequency estimates by country

The selection of countries included in the EnsembleMHC population fitness assessment was based on several criteria regarding the underlying MHC-I allele data for that country **(SI A2)**. The MHC-I allele frequency data used in our model was obtained from the Allele Frequency Net Database (AFND)[40], and the frequencies were aggregated by country. However, the currently available population-based MHC-I frequency data has specific limitations and variances, which we have addressed as follows:

Quality of MHC data within countries. We define MHC-typing breadth as the diversity of identified MHC-I alleles within a population of communities, and its depth as the ability to accurately achieve 4-digit MHC-I genotype resolution. High variability was observed in both the MHC-I genotyping breadth and depth **(SI A2 inset)**. Consequently, additional filter-measures were introduced to capture potential sources of variance within the analyzed cohort of countries. The thresholds for filtering the country-wide MHC-I allele data were set based on meeting two inclusion criteria: 1) MHC genotyping of at least 1000 individuals have been performed in that population, avoiding skewing of allele frequencies due to small sample size. 2) MHC-I allele frequency data for at least 95% of the 52 MHC-I alleles for which the EnsembleMHC was parameterized to predict, ensuring full power of the EnsembleMHC workflow.

Ethnic communities within countries. In instances where the MHC-I allele frequencies would pertain to more than one community, the reported frequencies were counted towards both contributing groups. For example, the MHC-I frequency data pertaining to the Chinese minority in Germany would be factored into the population MHC-I frequencies for both China and Germany. In doing so, this treatment resolves both ancestral and demographic MHC-I allele frequencies.

Normalization of HLA data

Analogous to past work on HIV[41], a focus of this work was to uncover potential differences in SARS-CoV-2 MHC-I peptide presentation dynamics induced by the 52 selected alleles within a population. Accordingly, the MHC-I allele frequency data was carefully processed in order to maintain important differences in the expression of selected alleles, while minimizing the effect of confounding variables.

The MHC-I allele frequency data for a given population was first filtered to the 52 selected alleles. These allele frequencies were then converted to the theoretical total number of copies of that allele within the population (allele count) following

$$allele\ count = allele_{freq} \times 2 \times n \tag{2}$$

where $allele_{freq}$ is the observed allele frequency in a population and n is the sample size, both available at AFND. The allele count is then normalized with respect to the total allele count of selected 52 alleles within that population using the following relationship,

$$norm\ allele\ count_i = \frac{allele\ count_i}{\sum_{i=1}^{52} allele\ count_i} \tag{3}$$

,

Where i is one of the 52 selected alleles. This normalization is required to overcome the potential bias towards *hidden alleles* (alleles that are either not well characterized or not supported by EnsembleMHC) as would be seen using alternative allele frequency accounting techniques (e.g. sample weighted mean of selected allele frequencies or normalization with respect to all observed alleles with a population **(SI A3)**). The SARS-CoV-2 binding capacity of these *hidden alleles* cannot be accurately determined using the EnsembleMHC workflow, and therefore important potential relationships would be observed.

EnsembleMHC population score

The predicted ability of a given population to present SARS-CoV-2 derived peptides was assessed by calculating the EnsembleMHC Population (EMP) score. After the MHC-I allele frequency data filtering steps, 21 countries were included in the analysis. The calculation of the EnsembleMHC population

score is as follows:

$$EMP\ score = \frac{\sum_{i=1}^{52} peptide_{frac} \times norm\ allele\ count_i}{N_{norm\ allele\ count \neq 0}} \quad (4)$$

Where $norm\ allele\ count$ is the observed normalized allele count for a given allele in a population, $N_{norm\ allele\ count\ \neq 0}$ is the number of the 52 select alleles detected in a given population, and $peptide_{frac}$ is the peptide fraction or the fraction of total predicted peptides expected to be presented by that allele within the total set of predicted peptides.

Death rate-presentation correlation

The correlation between the EMP score and the observed deaths per million within the cohort of selected countries was calculated as a function of time. SARS-Cov-2 data covering the time dependent global evolution of the SARS-CoV-2 pandemic was obtained from Johns Hopkins University Center for Systems Science and Engineering[42] covering the time frame of January 22nd to April 9th. The temporal variations in occurrence of community spread observed in different countries were accounted for by rescaling the time series data relative to when a certain death threshold was met in a country. For example, if the analyzed death threshold was 10 deaths, then day 0 for all considered countries would be when that country met or surpassed 10 deaths. This analysis was performed for thresholds of 1-100 total deaths by day 0, and correlations were calculated at each day sequentially from day 0 until there were fewer than 6 countries remaining at that time point. The upper-limit of 100-deaths was chosen to ensure availability of death-rate data on at least 50% of the countries for a minimum of 7 days starting from day 0. Additionally, a steep decline in average statistical power is observed with day 0 death thresholds greater than 100 deaths **(SI A4)**.

The time death correlation was computed using Spearman's rank correlation coefficient. This method was chosen due to the small sample size and non-normality of the underlying data **(SI A5)**. The reported correlations of EMP score and deaths per million using other correlation methods can be seen in supplemental figures **(SI A6)**.

The low observed statistical power of the obtained correlations, due to the small sample size, was accounted for by calculating the Positive Predictive Value (PPV) using the following equation[43]:

$$PPV = \frac{1 - \beta \times R}{1 - \beta \times R + \alpha} \quad (5)$$

Where $1 - \beta$ is the statistical power of a given correlation, R is the pre-study odds, and α is the significance level. A PPV value of $\geq 95\%$ is analogous to a p value of ≤ 0.05. Due to an unknown pre-study odd (probability that probed effect is truly non-null), R was set to 1 in the reported correlations. The proportion of reported correlations with a PPV of 95% at different R values can be seen in supplemental figures **(SI A7)**. The significance of partitioning high risk and low risk countries based on median EMP score was determined using Mann-Whitney U-test.

3 Results

3.1 EnsembleMHC workflow

Existing literature has established the benefits of ensemble algorithms towards improving the quality of MHC-I binding predictions[44]. Accordingly, EnsembleMHC employs a query-by-committee approach, using each of the 7 contributing algorithms separately to predict MHC-I peptides. Due to the observed variations in binding capacity across different MHC-I alleles, EnsembleMHC uses allele and algorithm specific binding affinity thresholds, a method that has been shown to improve peptide identification[45,46]. These allele and algorithm specific binding affinity thresholds were set for 52 MHC-I alleles based on the minimum score required to produce 50% recall of peptides detected in the MS-based immunopeptidome characterization[35]. In addition to the determination of specific score thresholds, the empirical false detection rate (FDR) was determined for each algorithm at each allele. This allowed for the assessment of confidence level in each peptide call, $peptide^{FDR}$, by taking the product of the empirical FDRs of each algorithm that detected a given peptide. These qualities of EnsembleMHC produce two desirable traits. First, it determines an allele specific score threshold for each algorithm at which a known quantity of peptides can be expected to be successfully presented on the cell surface. Second, it allows for confidence level assignment of each peptide call made by each algorithm **(Methods)**.

The results of EnsembleMHC parameterization are illustrated in **Figure 1**. The measured allele- and algorithm-specific FDR is shown in **Figure 1A**. The average deviation between FDRs for each algorithm across all alleles was found to 0.12 with C*4:01 showing the highest FDR deviation of 0.291 and A*01:01 showing the lowest deviation of 0.033. Overall, all 7 algorithms exhibited a similar distribution of FDR values **(Figure 1B)**, however, analysis of individual peptide score correlations between algorithm indicated only a moderate level of score correlation (mean = 0.603) **(Figure 1C)**. This indicated that, while the overall performance of each of the included algorithms was comparable, there was a diversity in individual peptide calls by each algorithm, supporting an integrated ap-

proach to peptide selection. An overview of the EnsembleMHC pipeline is presented in **Figure 1D**.

3.2 EnsembleMHC predictions of the SARS-CoV-2 MHC-I peptides reveals unequal peptide-allele distributions between the SARS-CoV-2 proteome and SARS-CoV-2 viral capsid proteins.

The MHC-I peptides derived from the SARS-CoV-2 proteome were predicted and prioritized using EnsembleMHC. A total of 67,207 potential 8-14mer viral peptides were evaluated for each of the considered MHC-I alleles. After filtering the pool of candidate peptides at the 5% peptide FDR threshold, the number potential peptides were reduced to 971 (658 unique peptides) **(SI A8, table 1)**. Illustrated in **Figure 2A**, the identified peptides were assigned to their respective alleles. There was an average of 18.6 peptides per allele with a maximum of 47 peptides (HLA-A*24:02), a minimum of 3 peptides (HLA-A*02:05), and a standard deviation (SD) of 11.3 peptides/per allele. In support of the quality of the identified peptides, the predicted peptides adhere to expected MHC-I peptide length distributions[47] **(SI A9)** and reflect known peptide binding motifs[35,48] **(SI A10)**.

The high expression, relative conservation, and reduced search space of SARS-CoV-2 viral capsid structural proteins (S, E, M, and N) makes MHC-I binding peptides derived from these proteins high-value targets for T cell-based vaccine development. **Figure 2B** describes the distribution of peptide-allele assignments originating from the four structural proteins. This assignment analysis markedly reduces the number of total predicted peptides to 160 (108 unique peptides) peptides (SI table 1). The average number of peptides per allele for specific SARS-CoV-2 structural proteins was found to be 3.1 with a maximum of 12 peptides (HLA-B*53:01), a minimum of 0 (HLA*15:02, B*35:03,B*38:01,C*03:03,C*15:02), and a SD of 2.6 peptides.

The peptide-allele distributions for SARS-CoV-2 structural proteins demonstrated a considerable decrease in the total number assigned peptides. To determine if this reduction significantly altered the overall SARS-CoV-2 presentation landscape, the relative changes in peptide-allele assignment between the full SARS-CoV-2 proteome and specifically structural proteins was visualized **(Figure 2C)**. 9 alleles demonstrated change of greater than one SD in relative peptide count between the two protein sets. The greatest decrease in the number of predicted peptides for a given allele was observed for A*25:01 (1.85 SDs), and the greatest increase was seen with A*31:01 (1.81 SDs). Furthermore, the SARS-CoV-2 structural protein allele-peptide distribution was found to be more variable with a coefficient of variance (SD / mean) of 0.83 compared to 0.61 for all SARS-CoV-2 proteins.

The current results offer two key insights. First, there is an uneven distribution of high confidence predicted SARS-CoV-2 MHC-I peptides across a diverse panel of 52 common alleles. Second, there is a significant rearrangement of the peptide-allele distribution of predicted MHC-I peptides originating from SARS-CoV-2 structural proteins producing a more variable distribution. Taken together, these results provide preliminary evidence of MHC-I allele bias in the presentation of SARS-CoV-2 peptides that is more pronounced for structural proteins.

3.3 Total population epitope load inversely correlates with reported death rates from SARS-CoV-2

The high variability in total epitope load per allele has several clinical implications. In cancer immunology, total epitope load (the number of novel potential MHC-I binding epitopes in a tumor) is strongly associated with the response to immunotherapy, and to the presence of pre-existing cytotoxic T cell immunity[49–51]. For viral immunity, certain MHC-I alleles are strongly associated with long term control of chronic viruses such as HIV[41,52]. We observed an uneven distribution of peptide-allele assignments for predicted SARS-CoV-2 MHC-I ppeptides across the 52 common MHC-I alleles. To determine if the described inequities in the peptide-allele assignment could be predictive of population fitness against SARS-CoV-2, the correlation of the EnsembleMHC population score (EMP) with the reported deaths per million for 21 countries was analyzed as a function of time **(Figure 3AB, Methods)**.

The EMP score for a given country was calculated based on the average predicted SARS-CoV-2 presentation capacity of 52 MHC-I alleles weighted by the normalized expression of those alleles in a given population **(Methods)**. Therefore, the EMP score represents the average predicted SARS-CoV-2 binding capacity in a given population. The individual country EMP scores were then correlated to observed deaths per million as a function of time (days) from when a country met a certain minimum death threshold, correcting for temporal variance in the occurrence of SARS-CoV-2 community spread.

Figure 3A shows results of the correlation analysis for the EMP score based on the entire SARS-Cov-2 proteome **(left panel)** or restricted to only SARS-Cov-2 structural proteins **(right panel)**. Both analyses demonstrated an overall inverse correlation between EMP score and deaths per million that strengthened as time progressed with a mean correlation of -0.61 and -0.71 for structural protein population score and all proteome population score respectively. Significance testing of each correlation revealed that 66% of correlations attained a p-value of ≤ 0.05. Furthermore, correlations based on the structural protein EMP score demonstrated a 30% higher proportion of statistically significant correlations compared to the

all SARS-CoV-2 protein EMP score (76% *vs* 57%). Due to relatively low statistical power of the obtained correlations (63.8% of correlations below 80% power), the positive predictive value for each correlation (equation 5) was calculated **(SI A7)**. The resulting proportions of correlations with a positive predictive value of $\geq$ 95% were similar to the observed significant p-value proportions with 62% of all correlations, 69% of structural protein EMP correlations, and 55% all SARS-CoV-2 protein EMP correlations. This similarity in results for p-values and PPV analysis supports that an overall true correlation is being captured.

To further capture the dynamics of EMP score based on the presentation of SARS-CoV-2 structural proteins and the observed deaths per million, **Figure 3 BC** depicts the correlation between the structural protein based EMP score and deaths per million at the median death threshold (50 deaths) for days 1, 5, 10, and 15 days. **Figure 3B** shows the correlation at the selected time points with all days excluding day 1 reaching statistical significance. To assess the ability for the EMP score to stratify high and low risk groups, countries were partitioned at each time point based on whether their assigned EMP score was greater or less than the median observed EMP score **(Figure 3C)**. Similarly, with the exception of day 1, all other days showed a significant difference in observed deaths per million the high and low EMP score groups. The robustness of the reported correlation calculations was determined through bootstrap sampling of half of the countries using the 50-death threshold and reevaluating the correlation between between observed deaths per million and EMP score **(SI A11)**. The significantly reduced sample size decreased the proportion of bootstrapped correlations with PPV values of $\geq$ 95% to 11%. However, the correlations attained from the EMP score based on structural proteins produced 10-fold higher number of significant correlations than randomly assigned EMP scores and 2.5-fold higher than the correlations based on the all SARS-CoV-2 proteins EMP.

In summary, we make three important observations. First, there is evidence of a statistically significant inverse correlation of EMP score and observed deaths per million. Second, there is evidence that this relationship is primarily driven by the presentation of SARS-CoV-2 structural proteins. Finally, there is the potential to separate high and low risk populations based on EMP score.

3.4 Peptides identified by the presentation score function identify high value target regions

The EMP score based on structural proteins indicated that observed deaths per million may be primarily shaped by the presentation of MHC-I peptides derived from SARS-Cov-2 structural proteins. To gain additional insight into these predicted peptides, the identified structural peptides were mapped back into their originating protein sequence revealing regions that were enriched for predicted MHC-I peptides **(Figure 4A)**. The potential for predicted peptides to be disrupted by polymorphisms was assesed by aligning 4,455 full length protein coding sequences for SARS-Cov-2 structural proteins, and calculating the number of unique polymorphisms at each position **(figure 2B)**. This analysis revealed that MHC-I peptides derived from structural proteins were unlikely to be impacted by polymorphism with the average number of polymorphic residues between all sequences being 6.2% and overall mutation frequency being rare (99.99% sequence conservation). The structural implications of these MHC-I peptide hotspots were explored by mapping predicted MHC-I peptides onto existing protein structures for the envelope and spike protein **(Figure 4C,F)** and the predicted structures for the nucleocapsid and membrane protein **(Figure 4D,E)**. Of particular note, the peptides mapped to the envelope protein revealed an enrichment of peptide located in the transmembrane portion of the channel. This localization indicates that peptides predicted from the envelope protein are less likely to allow an avenue for viral escape due to the necessitated invariance of such regions for function[53]. In terms of peptide localization, density, and sequence invariance, the membrane protein appears as the most appealing target. However, further interpretation of the structural impact and importance of these peptides will require solved structures.

4 Discussion

The current SARS-CoV-2 pandemic has affected the global population in ways that have rarely been observed in human history, prompting an unparalleled community effort to develop targeted antiviral therapies. One manifestation of this effort has been the development of targeted therapeutics for symptomatic or high-risk patients. In parallel, there is a global effort to develop immune-based therapies for induction of lasting immunity and disease prevention. Previously, SARS-CoV-1 has demonstrated that long lasting immunity is largely contingent on T cell mediated immunity with memory T cells being detected up to 11 years post infection and memory B cells showing a rapid decline 1-2 year post infection[54,55]. Furthermore, elevated levels of $CD8^+$ T cell are predictive of patient outcome and duration of viral positivity in SARS-CoV-2 infections[56–58]. Therefore, further investigations for inducing T cell immunity are warranted.

The goal of the presented work was to uncover the potential relationship between the presentation of SARS-CoV-2 peptides and patient outcome. However, due to the lack of large scale genomic assessments of individual MHC genotype and outcome associations, we created a simplified paradigm in

which every country was assigned a population SARS-CoV-2 presentation capacity (EnsembleMHC population) score based on the individual binding capacities and frequencies of endemic alleles. This approach has several limitations. First, this model assumes the fidelity of reported SARS-CoV-2 deaths and MHC-I allele frequencies. Second, the presented model does not account for other epidemiologic factors that impact mortality rates from the virus (e.g. social distance, governmental policies, comorbidities, healthcare infrastructure, age). Due to nature of the EnsembleMHC parameterization, only a subset of MHC-I alleles are considered, but the selected alleles represent some of the most common global MHC-I alleles. While there is large observed variance in the MHC-I protein sequence, the variation in unique peptide binding motifs is much lower. Future iterations of EnsembleMHC will be expanded to a larger set of alleles with structure-based clustering of MHC alleles. Finally, the model operates under the assumption that the identified peptides are both naturally processed and immunogenic. While evidence supports that most presented viral peptides are immunogenic[59], future biochemical validation with mass spectrometry of infected cells will be needed.

The ability to mount an effective $CD8^+$ T cell response is reliant on the MHC-I driven presentation of viral peptides. EnsembleMHC aims to improve peptide call confidence through three mechanisms. First, EnsembleMHC leverages 7 disparate **(Figure 1C)** publicly available MHC-I binding and processing algorithms to perform ensemble-based predictions, a method which improves the accuracy of MHC-I predictions[44,46,60]. Next, allele- and algorithm-specific binding thresholds were used to identify peptides. One potential source of error in MHC-I binding predictions, especially in the identification of immunogenic epitopes, is the use of global binding affinity thresholds (e.g. selecting peptides with a predicted binding affinity of $\leq$ 500nm) across diverse allele sets[45,46]. Finally, EnsembleMHC is designed to permit cross-allele predictions by benchmarking allele score thresholds to known mass spectrometry data and evaluating each peptide with an assigned $peptide^{FDR}$. This quantity represents the probability that the identified peptide is a false positive based on the measured empirical FDR of each detecting algorithm as determined during parameterization.

The algorithm FDR assessment was shown to be robust even in best case scenarios for some of the component algorithms. MHCflurry, netMHCpan-EL-4.0 and MixMHCpred had been trained on subsets of MS data used for parameterization. This could potentially lead to unfair weighting of these algorithms in the $peptide^{FDR}$ calculation. However, these algorithms still produced modest FDRs **(Figure 1B)**, likely due to the in-house generation of decoy peptides. The predicted peptides were identified in a highly cooperative manner with all peptides identified by at least 3 algorithms, and 60% of peptides being identified by 5 or more algorithms.

The EnsembleMHC workflow was used to predict 8-14mer peptides derived from SARS-CoV-2 proteins for a panel of 52 common MHC-I alleles resulting in the identification of 658 unique peptides. Analysis of the peptide-allele assignment distribution shows a wide variation in the number of peptides assigned to each allele, in particular for structural proteins, indicating a potential presentation capacity hierarchy for SARS-CoV-2 peptides **(Figure 2)**.

The relationship between MHC-I allele and clinical outcome of infection was indirectly measured by calculating the correlation of the EnsembleMHC population score and observed mortality rates in 21 countries as a function of time. Since mortality was inversely correlated with the EnsembleMHC population score, this supports the hypothesis that enhanced presentation of SARS-CoV-2 could lead to improved T cell immunity **(Figure 3)**. The correlation was strongest when using the predicted epitope load from the SARS-CoV-2 structural proteins, suggesting that protective SARS-CoV-2 CD8+ T cell responses are predominantly driven by the presentation of peptides derived from structural proteins, many of which have limited sequence variation **(Figure 4)**. In addition, the correlation between EnsembleMHC population score and mortality rates were stronger when starting the correlation analysis from higher mortality thresholds and later time points, suggesting that the impact of T cell immunity is more likely to shape an observed death rate per million once the virus becomes embedded in a population.

The influence of MHC genotype to shape clinical outcome has been well studied in the context of HIV infections[41]. For coronaviruses, there have been several studies of MHC association with disease susceptibility. A study of a Taiwanese and Hong Kong cohort of patients with SARS-CoV found that HLA-B*07:03 and HLA-B*46:01 were linked to increased susceptibility while HLA-Cw*15:02 was linked to increased resistance[61–63]. However, such associations did not remain after statistical correction and it is still unclear if MHC-outcome associations reported for SARS-CoV-1 are applicable to SARS-CoV-2[64]. Recently, a comprehensive prediction of SARS-CoV-2 MHC-I peptides indicated a depletion of a high affinity binding peptides for HLA-B*46:01, potentially supporting a similar association in SARS-CoV-2. However, due to the use of global binding affinity-based thresholds, it remains unclear if reported results represent a true depletion of targetable peptides or an artifact due to variation in binding capacity between diverse alleles[45]. Accordingly, when using allele specific binding thresholds, an obvious depletion of peptides for HLA-B*46:01 was not observed in our study **(Figure 2AB)**. These conflicting results are likely a product of the underlying complexity of $CD8^+$ immunity and methodology of epitope prediction. While

overall epitope load has been associated with viral control[65,66], the quality of the presented peptide is also an important factor[52]. Individual risk based on full MHC-I genotype will be critical for validation of these predictions. Since evolutionary divergence of individual patient MHC-I genotype has shown to be predictive of response to immune checkpoint therapy in cancer[67], we predict that a similar relationship would exist for the presentation of diverse SARS-CoV-2 peptides.

Another important factor to consider is the overall system dynamics of immune response to SARS-CoV-2 infections. Overstimulation of the innate immune response has been implicated as the driving cause of mortality through the induction of cytokine storms, which reduces T cell quantity and functional markers[56]. As cytokine dysregulation diminishes with robust T cell responses, we predict that an early and diverse CD8+ T cell response would limit the progression to cytokine toxicity[68,69].

In summary, we identify a set of high confidence SARS-CoV-2 peptides that provide a valuable starting point for experimental validation. We show that the predicted peptides form a variable distribution across a diverse panel of 52 MHC-I alleles, and that a population score function based on the binding capacity and regional frequencies of individual alleles provides a strong and statistically significant inverse correlation with observed mortality within countries. Furthermore, we highlight the potential importance of peptides derived from viral structural proteins and show that these peptide originate from conserved regions of the virus. These epitopes may be targets for vaccine design and T cell immunotherapies.

5 Acknowledgments

We would like to thank Dr. Diego Chowell and John Vant for critical feedback.

References

[1] Zi Yue Zu et al. "Coronavirus disease 2019 (COVID-19): a perspective from China". In: *Radiology* (2020), p. 200490.

[2] Qun Li et al. "Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia". In: *New England Journal of Medicine* (2020).

[3] Yan-Rong Guo et al. "The origin, transmission and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak–an update on the status". In: *Military Medical Research* 7.1 (2020), pp. 1–10.

[4] Aiping Wu et al. "Genome composition and divergence of the novel coronavirus (2019-nCoV) originating in China". In: *Cell host & microbe* (2020).

[5] David E Swayne et al. "Domestic poultry and SARS coronavirus, southern China". In: *Emerging infectious diseases* 10.5 (2004), p. 914.

[6] Salah T Al Awaidy and Faryal Khamis. "Middle East Respiratory Syndrome Coronavirus (MERS-CoV) in Oman: Current Situation and Going Forward". In: *Oman medical journal* 34.3 (2019), p. 181.

[7] Zhiliang Hu et al. "Clinical characteristics of 24 asymptomatic infections with COVID-19 screened among close contacts in Nanjing, China". In: *Science China Life Sciences* (2020), pp. 1–6.

[8] H Keipp B Talbot et al. "Coronavirus infection and hospitalizations for acute respiratory illness in young children". In: *Journal of medical virology* 81.5 (2009), pp. 853–856.

[9] Chaomin Wu et al. "Risk factors associated with acute respiratory distress syndrome and death in patients with coronavirus disease 2019 pneumonia in Wuhan, China". In: *JAMA internal medicine* (2020).

[10] Anthony R Fehr and Stanley Perlman. "Coronaviruses: an overview of their replication and pathogenesis". In: *Coronaviruses*. Springer, 2015, pp. 1–23.

[11] Biao He et al. "Identification of diverse alphacoronaviruses and genomic characterization of a novel severe acute respiratory syndrome-like coronavirus from bats in China". In: *Journal of virology* 88.12 (2014), pp. 7070–7082.

[12] Catrin Sohrabi et al. "World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19)". In: *International Journal of Surgery* (2020).

[13] Xiaolu Tang et al. "On the origin and continuing evolution of SARS-CoV-2". In: *National Science Review* (2020).

[14] Fang Li. "Structure, function, and evolution of coronavirus spike proteins". In: *Annual review of virology* 3 (2016), pp. 237–261.

[15] Yushun Wan et al. "Receptor recognition by the novel coronavirus from Wuhan: an analysis based on decade-long structural studies of SARS coronavirus". In: *Journal of virology* 94.7 (2020).

[16] Zhixin Liu et al. "Composition and divergence of coronavirus spike proteins and host ACE2 receptors predict potential intermediate hosts of SARS-CoV-2". In: *Journal of medical virology* 92.6 (2020), pp. 595–601.

[17] Nianshuang Wang et al. "Structure of MERS-CoV spike receptor-binding domain complexed with human receptor DPP4". In: *Cell research* 23.8 (2013), p. 986.

[18] Wei-Chen Hsin et al. "Nucleocapsid protein-dependent assembly of the RNA packaging signal of Middle East respiratory syndrome coronavirus". In: *Journal of biomedical science* 25.1 (2018), p. 47.

[19] Lei Cui et al. "The nucleocapsid protein of coronaviruses acts as a viral suppressor of RNA silencing in mammalian cells". In: *Journal of virology* 89.17 (2015), pp. 9029–9043.

[20] Roujian Lu et al. "Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding". In: *The Lancet* 395.10224 (2020), pp. 565–574.

[21] Syed Faraz Ahmed, Ahmed A Quadeer, and Matthew R McKay. "Preliminary identification of potential vaccine targets for the COVID-19 coronavirus (SARS-CoV-2) based on SARS-CoV immunological studies". In: *Viruses* 12.3 (2020), p. 254.

[22] Hsueh-Ling Janice Oh et al. "Understanding the T cell immune response in SARS coronavirus infection". In: *Emerging microbes & infections* 1.1 (2012), pp. 1–6.

[23] Yue-Dan Wang et al. "T-cell epitopes in severe acute respiratory syndrome (SARS) coronavirus spike protein elicit a specific T-cell immune response in patients who recover from SARS". In: *Journal of virology* 78.11 (2004), pp. 5612–5618.

[24] William J Liu et al. "T-cell immunity of SARS-CoV: Implications for vaccine development against MERS-CoV". In: *Antiviral research* 137 (2017), pp. 82–92.

[25] Li-Tao Yang et al. "Long-lived effector/central memory T-cell responses to severe acute respiratory syndrome coronavirus (SARS-CoV) S antigen in recovered SARS patients". In: *Clinical immunology* 120.2 (2006), pp. 171–178.

[26] Shunsuke Kohyama et al. "Efficient induction of cytotoxic T lymphocytes specific for severe acute respiratory syndrome (SARS)-associated coronavirus by immunization with surface-linked liposomal peptides derived from a non-structural polyprotein 1a". In: *Antiviral research* 84.2 (2009), pp. 168–177.

[27] Alba Grifoni et al. "A sequence homology and bioinformatic approach can predict candidate targets for immune responses to SARS-CoV-2". In: *Cell host & microbe* (2020).

[28] Timothy O'Donnell, Alex Rubinsteyn, and Uri Laserson. "Improved predictive models for peptide presentation on MHC I". In: *BioRxiv* (2020).

[29] Massimo Andreatta and Morten Nielsen. "Gapped sequence alignment using artificial neural networks: application to the MHC class I system". In: *Bioinformatics* 32.4 (2016), pp. 511–517.

[30] Vanessa Jurtz et al. "NetMHCpan-4.0: improved peptide–MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data". In: *The Journal of Immunology* 199.9 (2017), pp. 3360–3368.

[31] Michael Rasmussen et al. "Pan-specific prediction of peptide–MHC class I complex stability, a correlate of T cell immunogenicity". In: *The Journal of Immunology* 197.4 (2016), pp. 1517–1524.

[32] Hao Zhang, Ole Lund, and Morten Nielsen. "The PickPocket method for predicting binding specificities for receptors based on receptor pocket similarities: application to MHC-peptide binding". In: *Bioinformatics* 25.10 (2009), pp. 1293–1299.

[33] Michal Bassani-Sternberg et al. "Deciphering HLA-I motifs across HLA peptidomes improves neo-antigen predictions and identifies allostery regulating HLA specificity". In: *PLoS computational biology* 13.8 (2017), e1005725.

[34] Marek Prachar et al. "COVID-19 Vaccine Candidates: Prediction and Validation of 174 SARS-CoV-2 Epitopes". In: *bioRxiv* (2020).

[35] Siranush Sarkizova et al. "A large peptidome dataset improves HLA class I epitope prediction across most of the human population". In: *Nature Biotechnology* 38.2 (2020), pp. 199–209.

[36] Eneida L Hatcher et al. "Virus Variation Resource–improved response to emergent viral outbreaks". In: *Nucleic acids research* 45.D1 (2017), pp. D482–D490.

[37] William Humphrey, Andrew Dalke, Klaus Schulten, et al. "VMD: visual molecular dynamics". In: *Journal of molecular graphics* 14.1 (1996), pp. 33–38.

[38] Helen M Berman et al. "The protein data bank". In: *Nucleic acids research* 28.1 (2000), pp. 235–242.

[39] Chengxin Zhang et al. "Protein structure and sequence re-analysis of 2019-nCoV genome refutes snakes as its intermediate host or the unique similarity between its spike protein insertions and HIV-1". In: *Journal of proteome research* (2020).

[40] Faviel F González-Galarza et al. "Allele frequency net 2015 update: new features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations". In: *Nucleic acids research* 43.D1 (2015), pp. D784–D788.

[41] The International HIV Controllers Study et al. "The major genetic determinants of HIV-1 control affect HLA class I peptide presentation". In: *Science (New York, NY)* 330.6010 (2010), p. 1551.

[42] Ensheng Dong, Hongru Du, and Lauren Gardner. "An interactive web-based dashboard to track COVID-19 in real time". In: *The Lancet infectious diseases* (2020).

[43] Katherine S Button et al. "Power failure: why small sample size undermines the reliability of neuroscience". In: *Nature Reviews Neuroscience* 14.5 (2013), pp. 365–376.

[44] Edita Karosiene et al. "NetMHCcons: a consensus method for the major histocompatibility complex class I predictions". In: *Immunogenetics* 64.3 (2012), pp. 177–186.

[45] Sinu Paul et al. "HLA class I alleles are associated with peptide-binding repertoires of different size, affinity, and immunogenicity". In: *The Journal of Immunology* 191.12 (2013), pp. 5831–5839.

[46] Maria Bonsack et al. "Performance evaluation of MHC class-I binding prediction tools based on an experimentally validated MHC-peptide binding dataset". In: *Cancer immunology research* (2019).

[47] Thomas Trolle et al. "The length distribution of class I–restricted T cell epitopes is determined by both peptide supply and MHC allele–specific binding preference". In: *The Journal of Immunology* 196.4 (2016), pp. 1480–1487.

[48] Nicolas Rapin et al. "The MHC motif viewer: a visualization tool for MHC binding motifs". In: *Current protocols in immunology* 88.1 (2010), pp. 18–17.

[49] Scott D Brown et al. "Neo-antigens predicted by tumor genome meta-analysis correlate with increased patient survival". In: *Genome research* 24.5 (2014), pp. 743–750.

[50] Diego Chowell et al. "Patient HLA class I genotype influences cancer response to checkpoint blockade immunotherapy". In: *Science* 359.6375 (2018), pp. 582–587.

[51] Marta Łuksza et al. "A neoantigen fitness model predicts tumour response to checkpoint blockade immunotherapy". In: *Nature* 551.7681 (2017), pp. 517–520.

[52] Gaurav D Gaiha et al. "Structural topology defines protective CD8+ T cell epitopes in the HIV proteome". In: *Science* 364.6439 (2019), pp. 480–484.

[53] Stephan Wickles et al. "A structural model of the active ribosome-bound membrane protein insertase YidC". In: *Elife* 3 (2014), e03035.

[54] Oi-Wing Ng et al. "Memory T cell responses targeting the SARS coronavirus persist up to 11 years post-infection". In: *Vaccine* 34.17 (2016), pp. 2008–2014.

[55] Rudragouda Channappanavar, Jincun Zhao, and Stanley Perlman. "T cell-mediated immune response to respiratory coronaviruses". In: *Immunologic research* 59.1-3 (2014), pp. 118–128.

[56] Bo Diao et al. "Reduction and functional exhaustion of T cells in patients with coronavirus disease 2019 (COVID-19)". In: *Medrxiv* (2020).

[57] Li Tan et al. "Lymphopenia predicts disease severity of COVID-19: a descriptive and predictive study". In: *Signal transduction and targeted therapy* 5.1 (2020), pp. 1–3.

[58] Aifen Lin et al. "Early risk factors for the duration of SARS-CoV-2 viral positivity in COVID-19 patients". In: *Clinical Infectious Diseases* (2020).

[59] Nathan P Croft et al. "Most viral peptides displayed by class I MHC on infected cells are immunogenic". In: *Proceedings of the National Academy of Sciences* 116.8 (2019), pp. 3112–3117.

[60] Sri Krishna et al. "Human papilloma virus specific immunogenicity and dysfunction of CD8+ T cells in head and neck cancer". In: *Cancer research* 78.21 (2018), pp. 6159–6170.

[61] Marie Lin et al. "Association of HLA class I with severe acute respiratory syndrome coronavirus infection". In: *BMC Medical Genetics* 4.1 (2003), p. 9.

[62] Sheng-Fan Wang et al. "Human-leukocyte antigen class I Cw 1502 and class II DR 0301 genotypes are associated with resistance to severe acute respiratory syndrome (SARS) infection". In: *Viral immunology* 24.5 (2011), pp. 421–426.

[63] Margaret HL Ng et al. "Association of human-leukocyte-antigen class I (B* 0703) and class II (DRB1* 0301) genotypes with susceptibility and resistance to the development of severe acute respiratory syndrome". In: *Journal of Infectious Diseases* 190.3 (2004), pp. 515–518.

[64] MH Ng et al. "Immunogenetics in SARS: a case-control study." In: *Hong Kong medical journal= Xianggang yi xue za zhi* 16.5 Suppl 4 (2010), p. 29.

[65] Christof Geldmacher et al. "CD8 T-cell recognition of multiple epitopes within specific Gag regions is associated with maintenance of a low steady-state viremia in human immunodeficiency virus type 1-seropositive patients". In: *Journal of virology* 81.5 (2007), pp. 2440–2448.

[66] Morgane Rolland et al. "Broad and Gag-biased HIV-1 epitope repertoires are associated with lower viral loads". In: *PloS one* 3.1 (2008).

[67] Diego Chowell et al. "Evolutionary divergence of HLA class I genotype impacts efficacy of cancer immunotherapy". In: *Nature medicine* 25.11 (2019), pp. 1715–1720.

[68] Kwang Dong Kim et al. "Adaptive immune cells temper initial innate responses". In: *Nature medicine* 13.10 (2007), pp. 1248–1252.

[69] Rudragouda Channappanavar and Stanley Perlman. "Pathogenic human coronavirus infections: causes and consequences of cytokine storm and immunopathology".

In: *Seminars in immunopathology*. Vol. 39. 5. Springer.
2017, pp. 529–539.

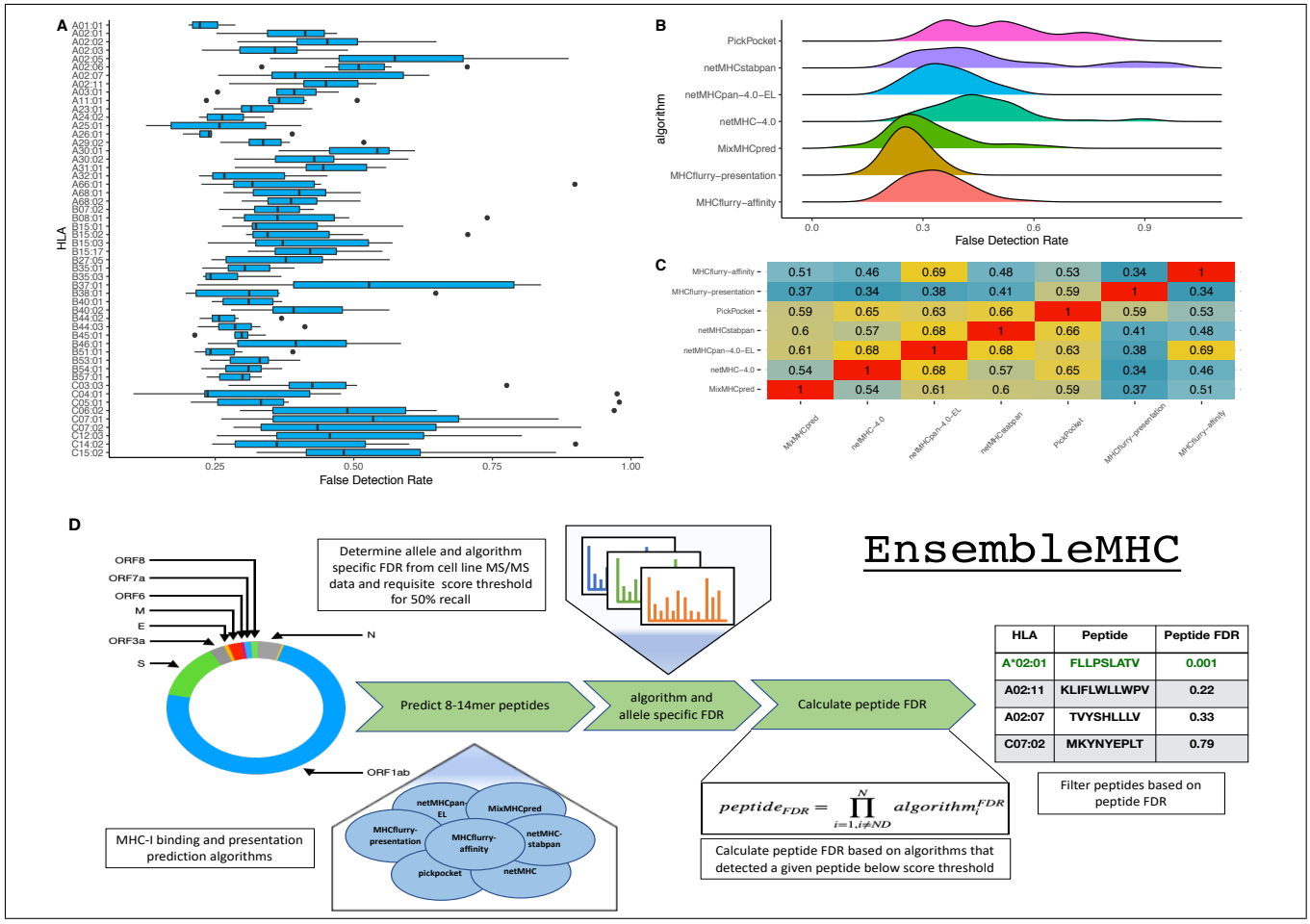

Figure 1: **EnsembleMHC prediction workflow. (A).** The EnsembleMHC score algorithm was parameterized using MHC-I peptides observed in mass spectrometry datasets and 100 randomly generate length and protein matched decoy peptides. The observed false detection rate (FDR) distribution at 50% recall for algorithms relative to each HLA. **(B).** The distribution of observed FDRs for each algorithm across all alleles. **(C).** The correlation between individual peptide scores for each algorithm across all alleles. **(D).** The EnsembleMHC workflow for the prediction of SARS-CoV-2 peptides (SI A1).

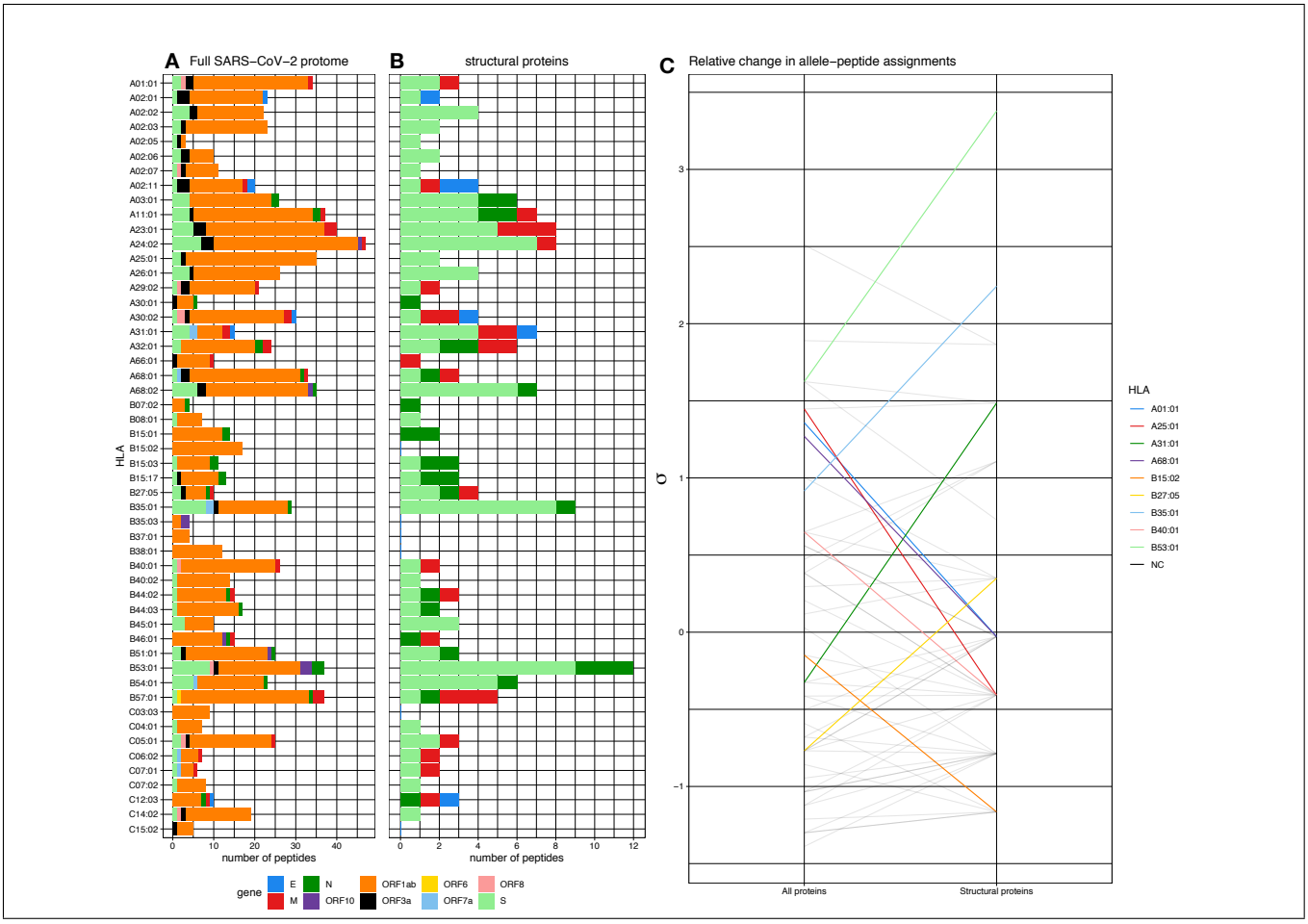

Figure 2: **Prediction of SARS-CoV-2 peptides across 52 common MHC-I alleles** The EnsembleMHC workflow was used to predict 8-14mer MHC-I peptides for 52 alleles from the entire SARS-CoV-2 proteome **(A)** or specifically SARS-CoV-2 structural proteins **(B)**. **(C)** Both distributions were individually standardized and the relative change in the binding capacity of each allele was calculated by taking the absolute difference of the Z-scores of allele binding capacity with respect to all SARS proteins or SARS structural proteins. Alleles showing a greater than 1 standard deviation increase or decrease change in binding capacity are highlighted in color.

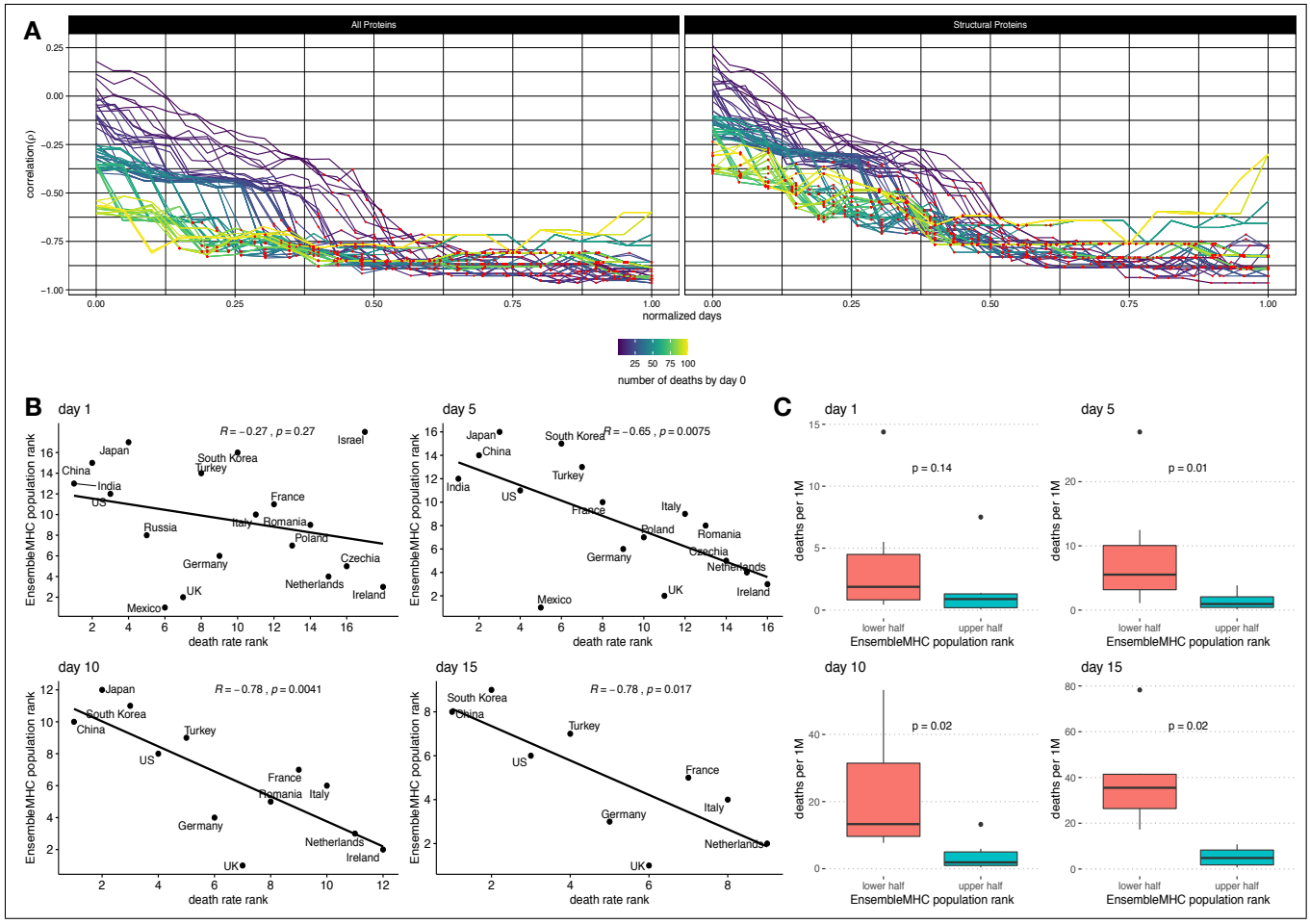

Figure 3: **predicted total epitope load within a population inversely correlates with mortality.** **(A)** The correlation between EnsembleMHC population score with respect to all SARS-CoV-2 proteins (left panel) or structural proteins and deaths per million were calculated at each day starting from the day a country passed a particular death milestone ranging from 1 reported death to 100 reported death (line color). The days from each start point were normalized, and correlations that were shown to be statistically significant are colored with a red point. **(B-C)** The correlations for between the EnsembleMHC score based on structural proteins and death rate were shown for countries meeting the 50 confirmed death threshold. **(B)** The correlation between deaths per million rank (min rank = least number of deaths max rank = most deaths) and EnsembleMHC population score rank (min rank = lowest score max rank = highest score) at days 1, 5, 10, and 15. Correlation coefficients and p values were assigned using spearman's rank correlation **(C)** The countries at each time point were partitioned into a upper or lower half based on the observed EnsembleMHC population score. P values were determined by Mann-Whitney U test.

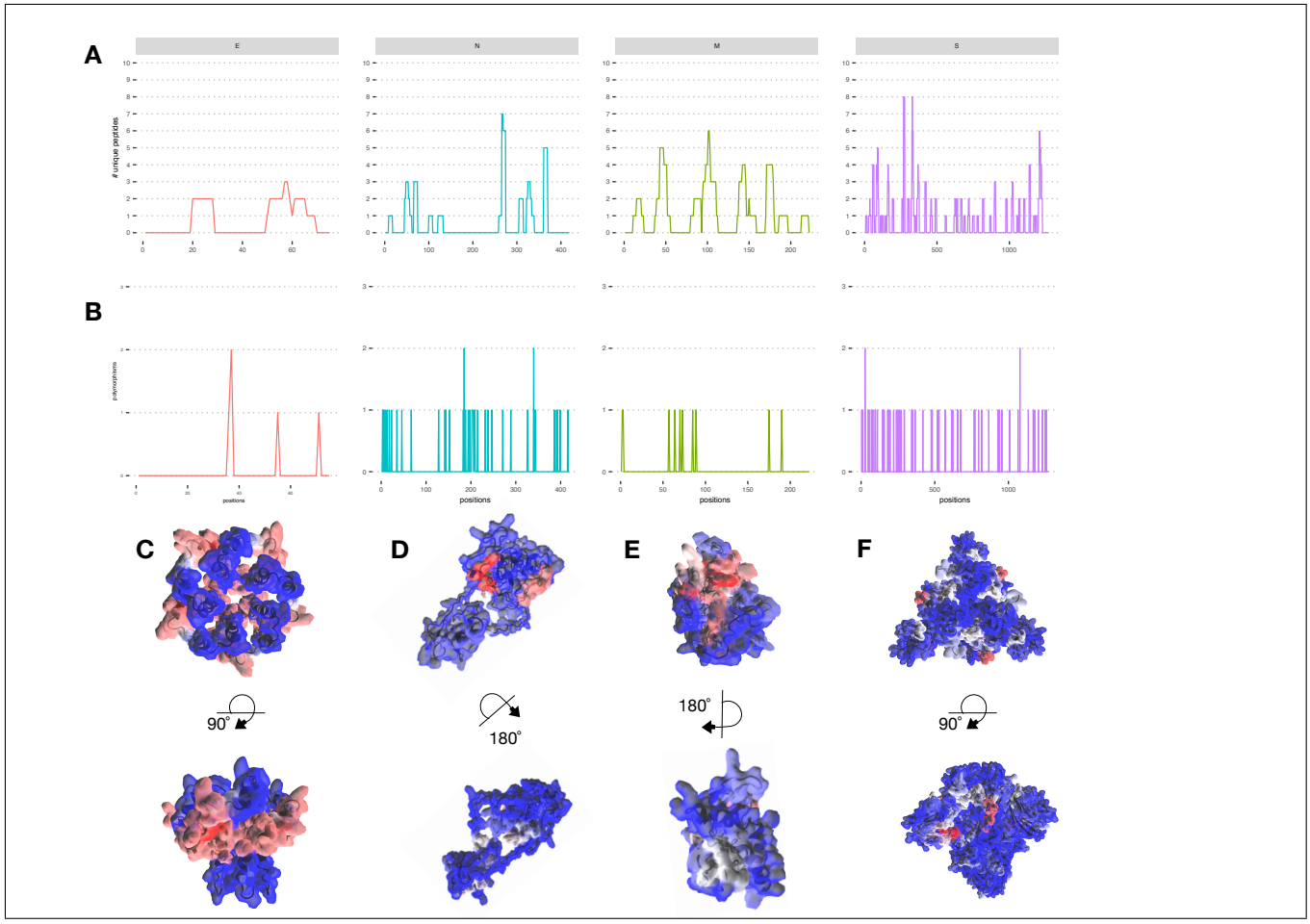

Figure 4: **Protein origin of predicted SARS-CoV-2 peptides.** The localization of predicted MHC-I peptides derived from SARS-CoV-2 structural proteins was determined by mapping the peptides back to the reference sequence. **(A).** The frequency of each amino acid for each of the four SARS-CoV-2 structural proteins appearing in one of the 160 predicted peptides. **(B)** The number of polymorphisms appearing at each position in the structural sequences determined from the alignment of 104 reported SARS-CoV-2 sequences. **(C).** The predicted peptides were mapped onto the solved structures for the envelope **(C)** and spike **(F)** proteins, and the predicted structures for the nucleocapsid **(D)** and membrane **(E)** proteins. Red regions indicate an enrichment of predicted peptides and blue regions indicate a depletion of predicted peptides.

A Supplemental figures and methods

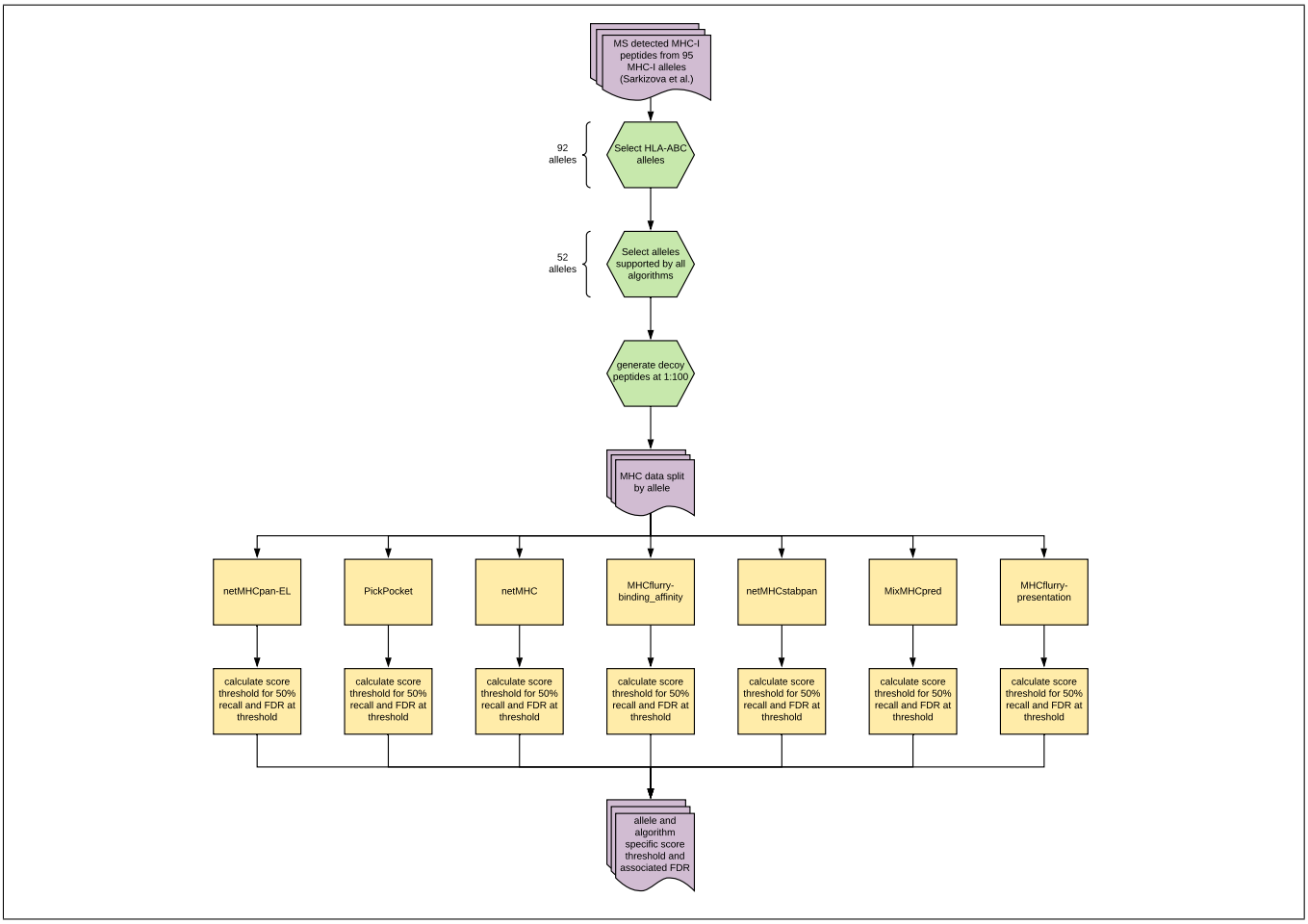

Figure A.1: **EnsembleMHC Parameterization overview**

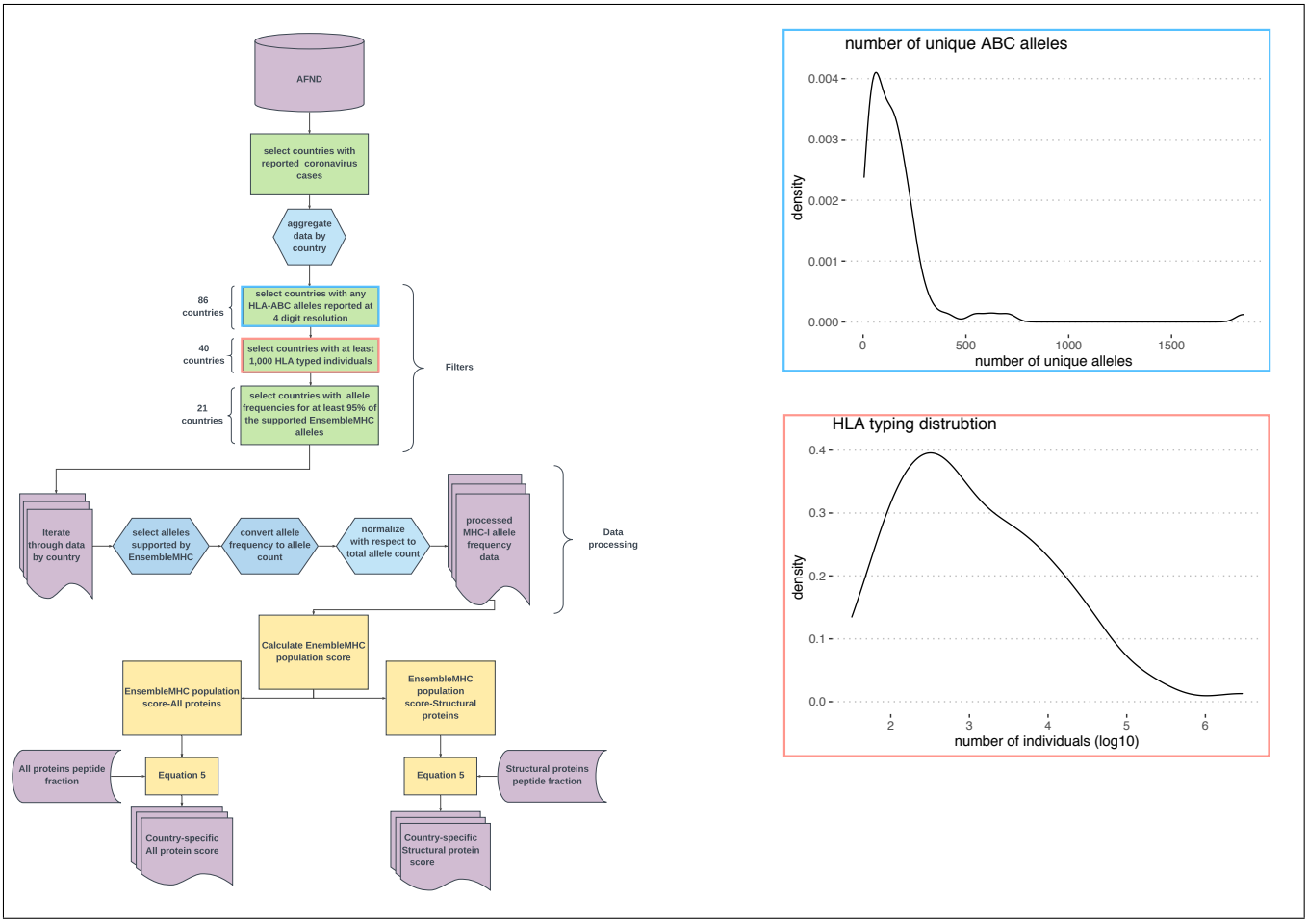

Figure A.2: **Data processing EnsembleMHC population score calculation.** The overview of the data processing steps for the global MHC-I allele frequency data and its application in the calculation the EnsembleMHC population score with respect to all SARS-CoV-2 proteins and SARS-CoV-2 structural proteins. **(inset plots)** The blue inset plot shows the distribution of the number of unique MHC-I alleles at 4 digit resolution in the set of countries with at least 1 reported coronavirus case. The red inset plot shows the distribution of the number of number of HLA-typed individuals in the set of countries with at least 1 reported coronavirus case. **AFND = Allele Frequency Net Database**

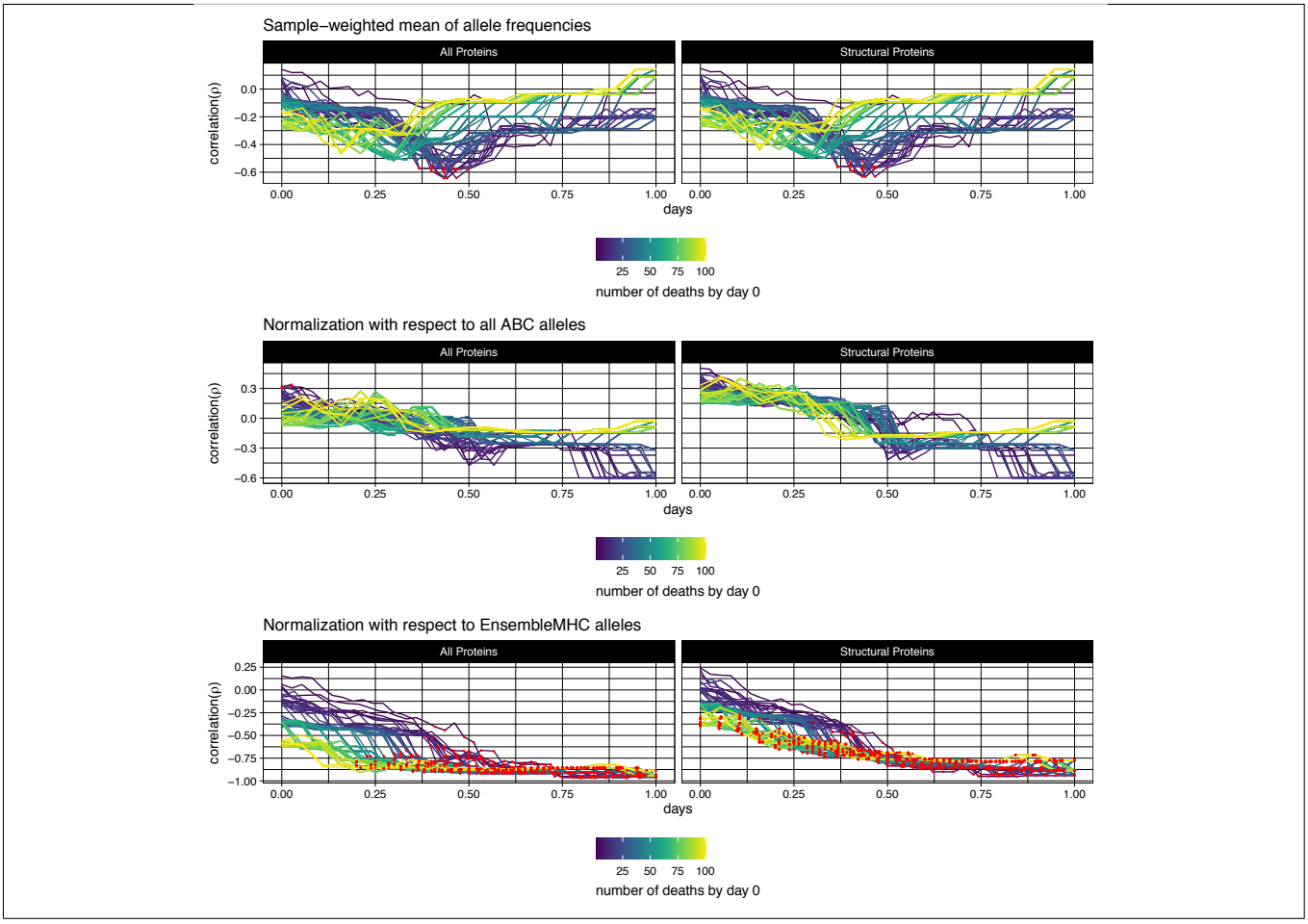

Figure A.3: **EnsembleMHC population score and deaths per million correlation using different allele frequency accounting methods.** The effect on the reported correlation with respect to alternative allele frequency accounting methods **(methods)**. **(Top panel)** The aggregation of allele frequencies within a particular country by taking the sample weighted mean of reported frequencies for a given allele. **(Middle panel)** Normalizing the allele count with respect to all detected alleles in a given population. **(Bottom panel)** Normalizing allele count with respect to only alleles supported by EnsembleMHC.

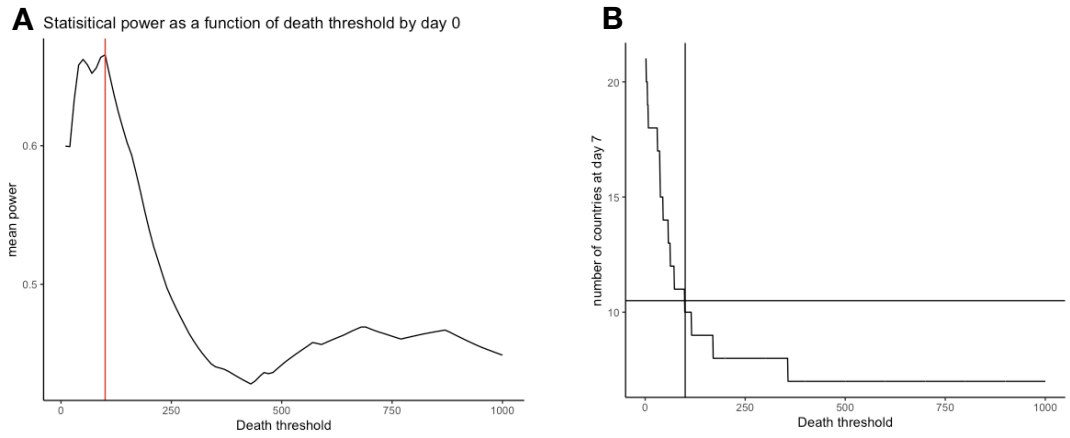

Figure A.4: **Justification of upper limit for death threshold.** **(A)** The mean statistical power of the resulting correlation of EnsembleMHC population at different death thresholds by day zero. The red line indicates the selected upper limit of analysis of 100 deaths by day 0. **(B)** The number of countries remaining at day seven from different death thresholds.

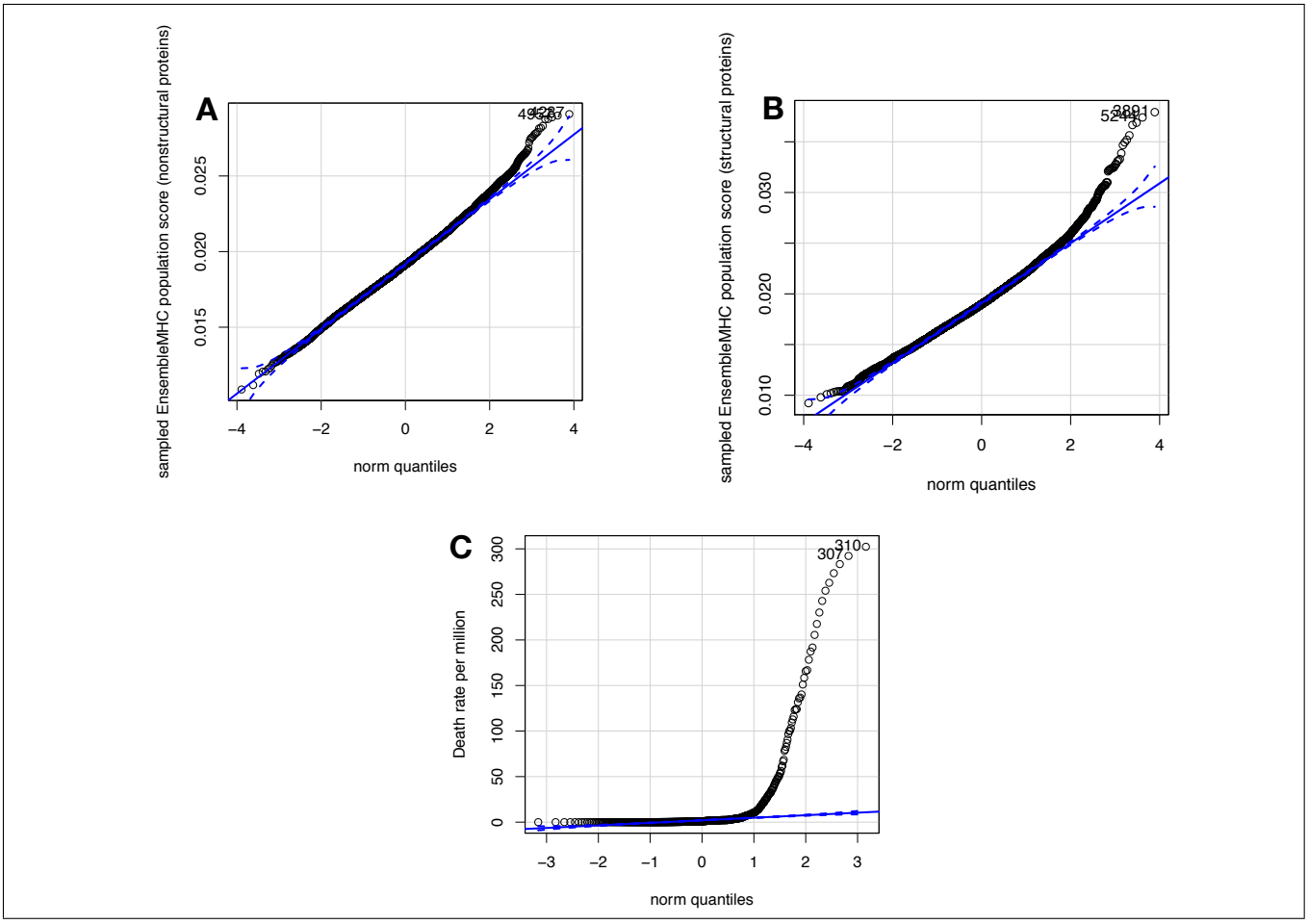

Figure A.5: **Justification of nonparametric correlation analysis.** The use of nonparametric correlation analysis, namely spearman's rho, is justified by the non-normality of the underlying data. EnsembleMHC population scores for all SARS-CoV-2 proteins and structural proteins were calculated for 10,000 simulated countries. Allele frequencies for simulated countries were generate by randomly sampling an observed allele frequencies for each of the considered 52 alleles and renormalizing to ensure the sum of allele frequencies were equal to one. Q-Q plots for all SARS-CoV-2 proteins EnsembleMHC population score **(A)** and structural protein EnsembleMHC population score **(B)** demonstrate positive skew. Similarly, the QQ plot for all reported deaths per Million indicates a very strong positive skew.

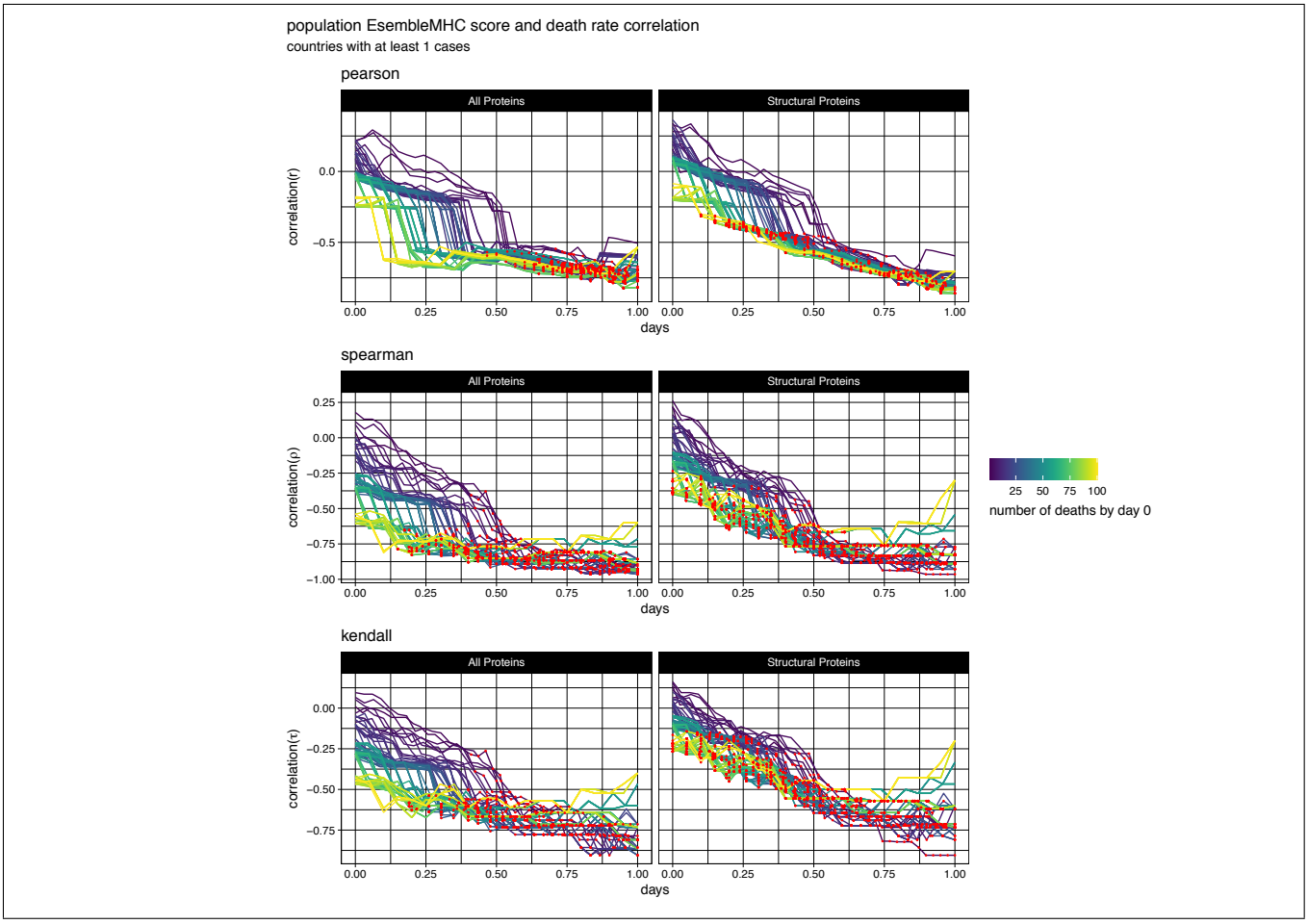

Figure A.6: **The effect of different correlation methods on the relationship between EnsembleMHC score and deaths per million.** The correlation between EnsembleMHC population score with respect to all SARS-CoV-2 proteins (left panel) or structural proteins and deaths per million were calculated at each day starting from the day a country passed a particular death milestone using Pearson's correlation (top), Spearman's rho (middle), and Kendall's tau (bottom). The days from each start point were normalized, and correlations that were shown to be statistically significant are colored with a red point.

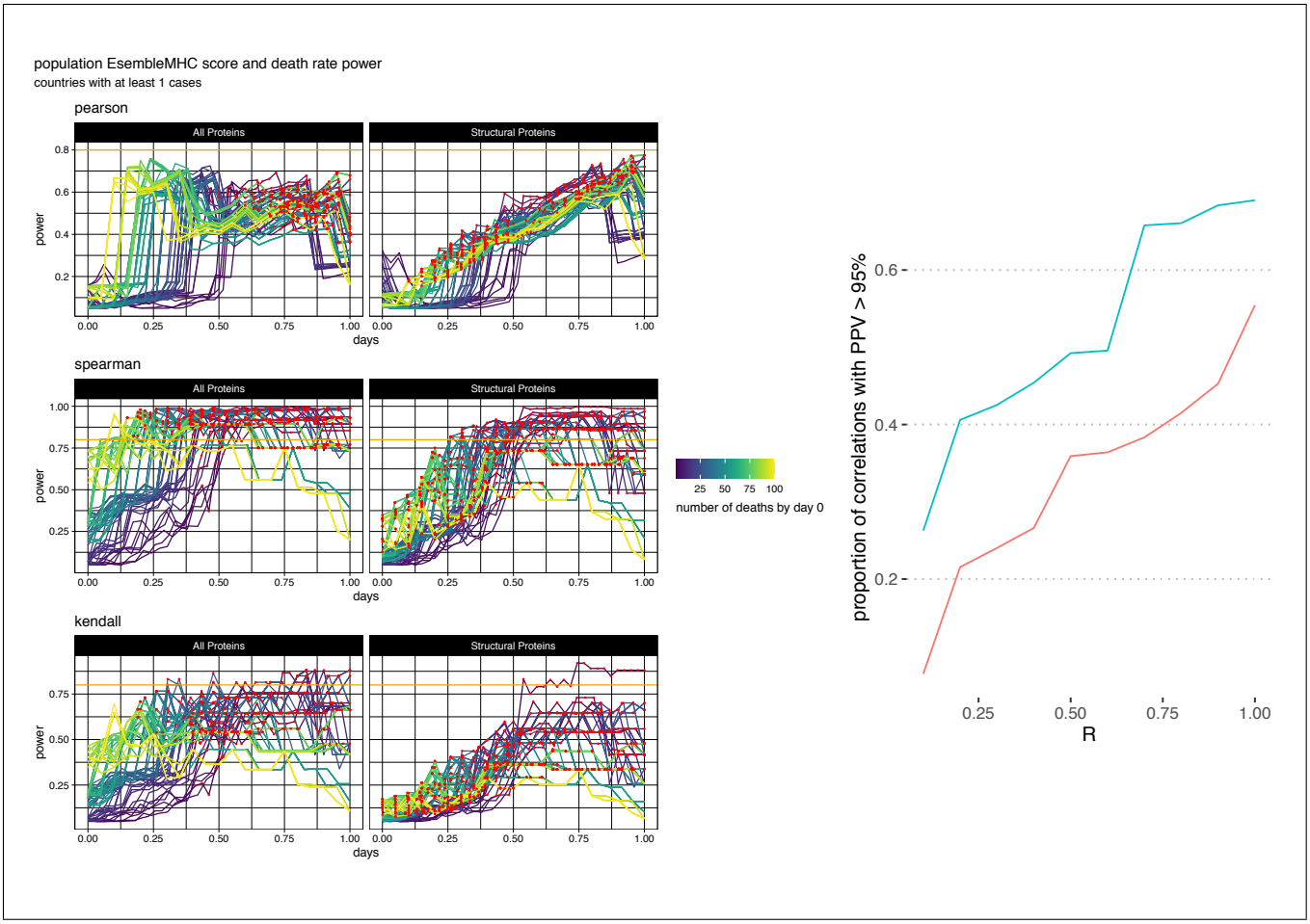

Figure A.7: **Analysis of statistical power for each correlation method.** The statistical power of each reported correlation between EnsembleMHC population score with respect to all SARS-CoV-2 proteins (left column) or structural proteins (right column) and deaths per million were calculated at each day starting from the day a country passed a particular death milestone using Pearson's correlation (top), Spearman's rho (middle), and Kendall's tau (bottom). The days from each start point were normalized, and correlations that were shown to be statistically significant are colored with a red point. The orange line indicates a power threshold of 80%. The line plot on the right shows the proportion of points achieving a PPV value of greater than 95% at different threshold for pre study odd (R) for the spearman correlation carried out in section 3. The blue line represents the proportion of significant correlations for EMP based on structural proteins while the red line represents the same correlations with an EMP score based on all SARS-CoV-2 proteins.

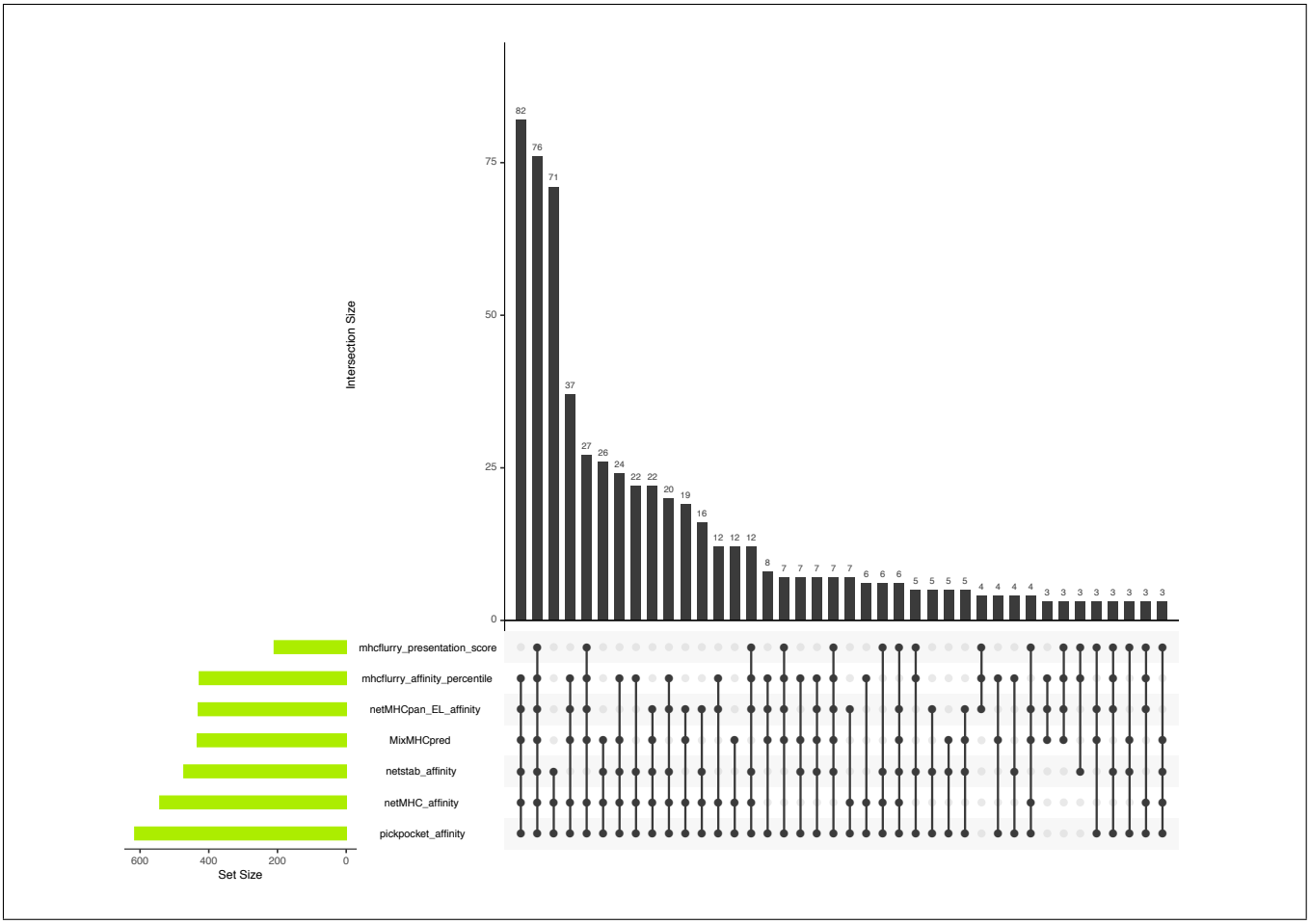

Figure A.8: **Contribution of each algorithm in EnsembleMHC to the total number of predicted peptides.** The UpSet plot shows the contribution of each algorithm to the 658 unique SARS-CoV-2 peptides identified by EnsembleMHC. The top bar plot indicates the number of unique peptides identified by the combination of algorithms shown by the points and segments under each bar. The bars on the left-hand side of the plot indicate the total number of peptides identified by each algorithm.

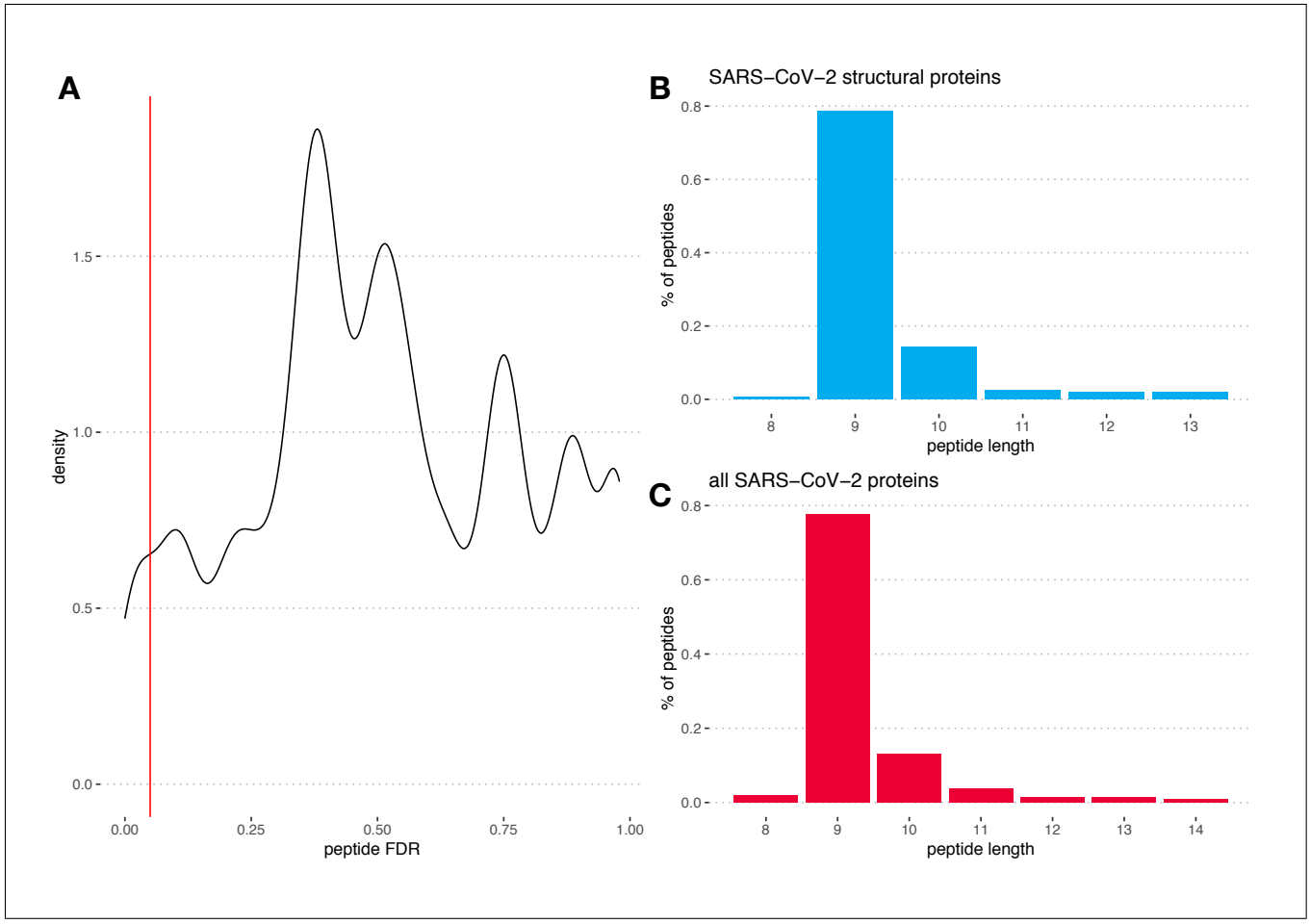

Figure A.9: **EnsembleMHC peptide FDR distribution and length distributions of predicted SARS-CoV-2 MHC-I peptides. (A).** The distribution of peptide FDRs for the 9,712 peptides before the application of the peptide FDR filter. The redline indicates an FDR level of 5%. The length distribution of peptides identified from the entire SARS-CoV-2 proteome **(B)** or SARS-CoV-2 structural protein **(C)**.

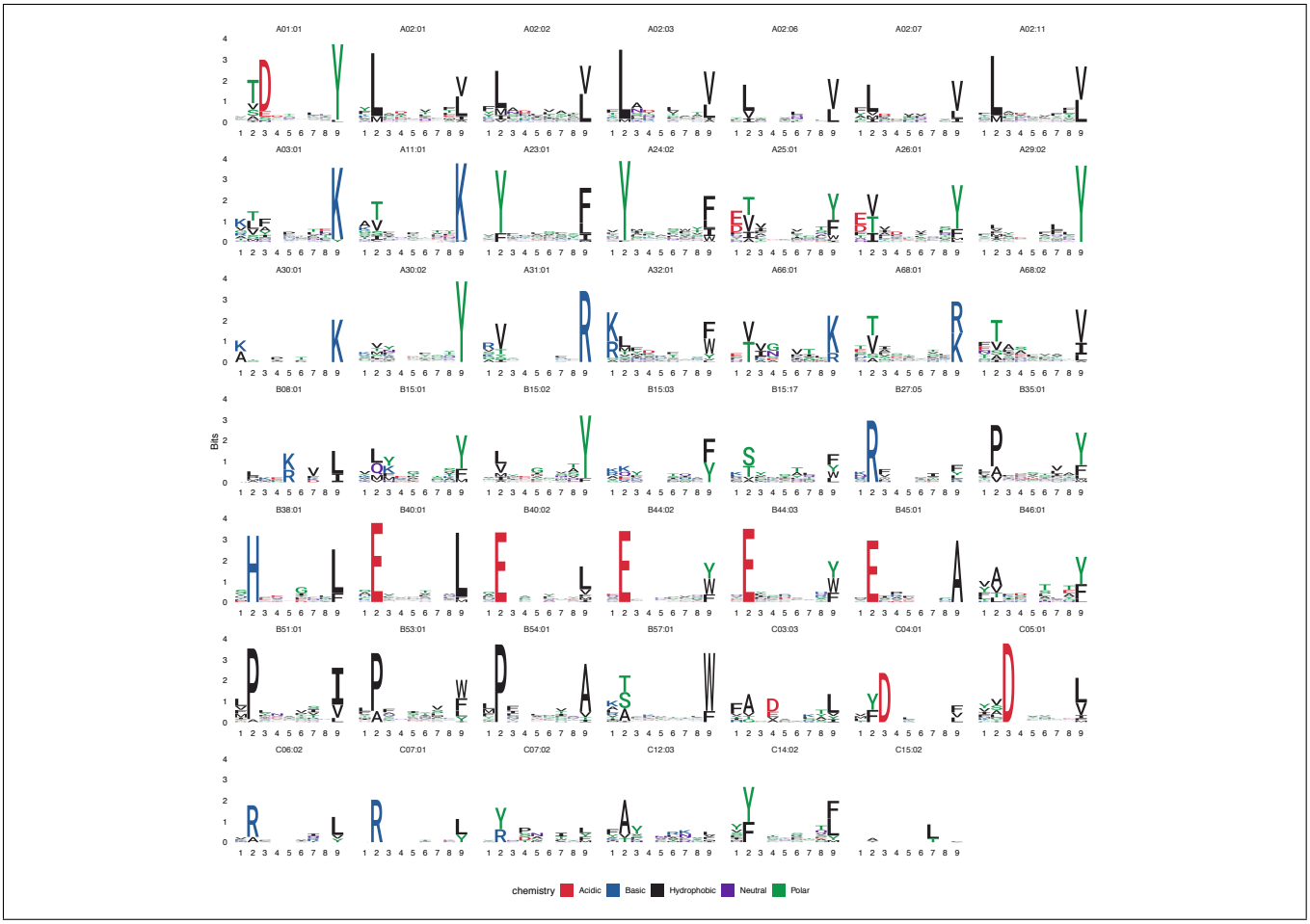

Figure A.10: **Logo plots for the identified peptides from the SARS proteome.** Logo plots were generated for MHC alleles with at least 5 peptides identified by EnsembleMHC prediction of all SARS-Cov-2 proteins.

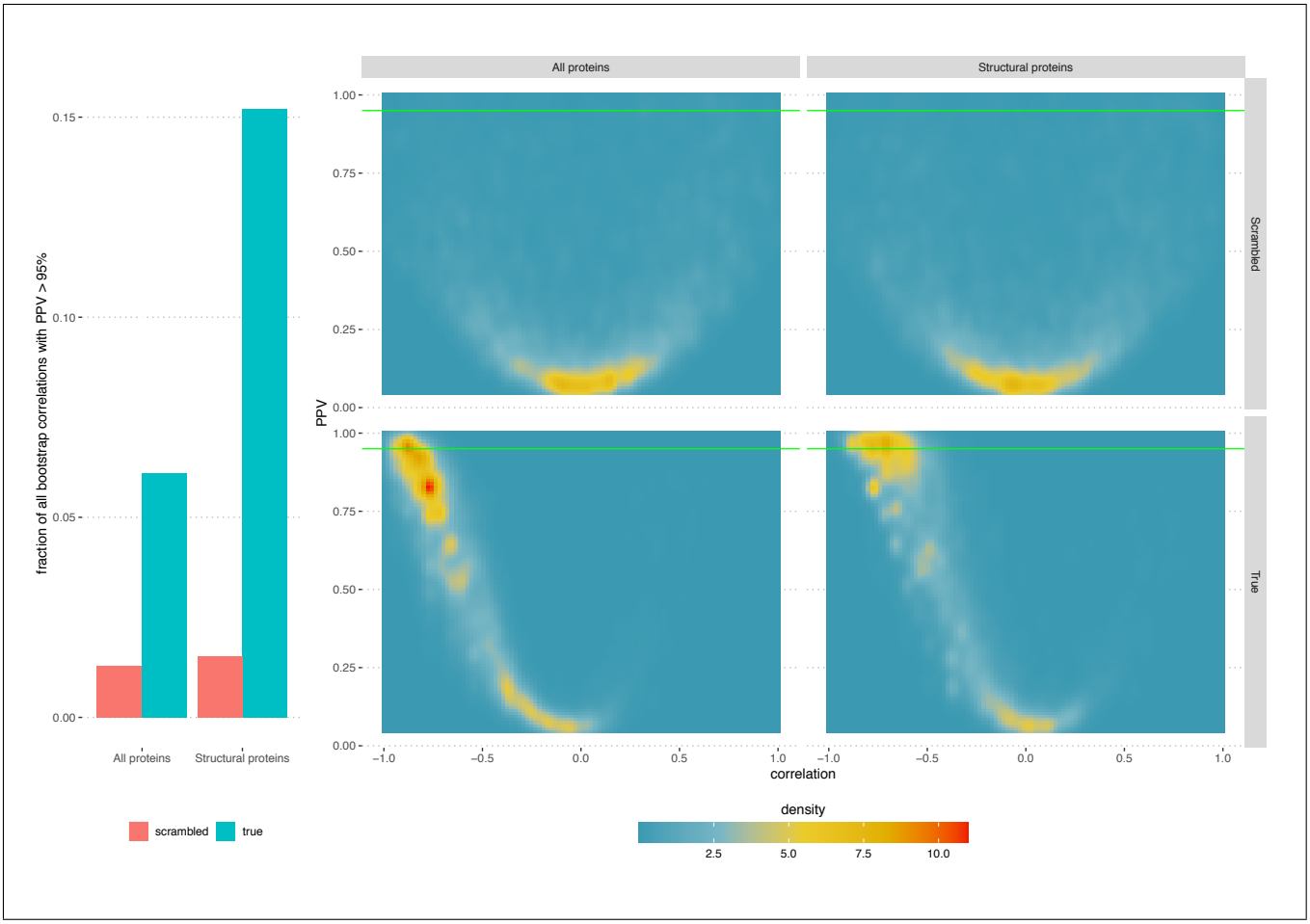

Figure A.11: **Bootstrapping analysis of EnsembleMHC score and deaths per million correlations at the 50 deaths threshold.** The correlations observed in the EnsembleMHC score and deaths per million correlations at the 50 deaths threshold were calculated over 1000 bootstrap iterations (True). In each bootstrap iteration, 50% of the countries were dropped from the analysis. The incidence of spurious statistically significant correlations was simulated by repeating the same bootstrap procedure but with randomized EnsembleMHC scores (scrambled). The green line indicates a PPV value of greater than 95%. The bar plot indicates the proportion of bootstrap iteration for each condition that produced a statistically significant value.

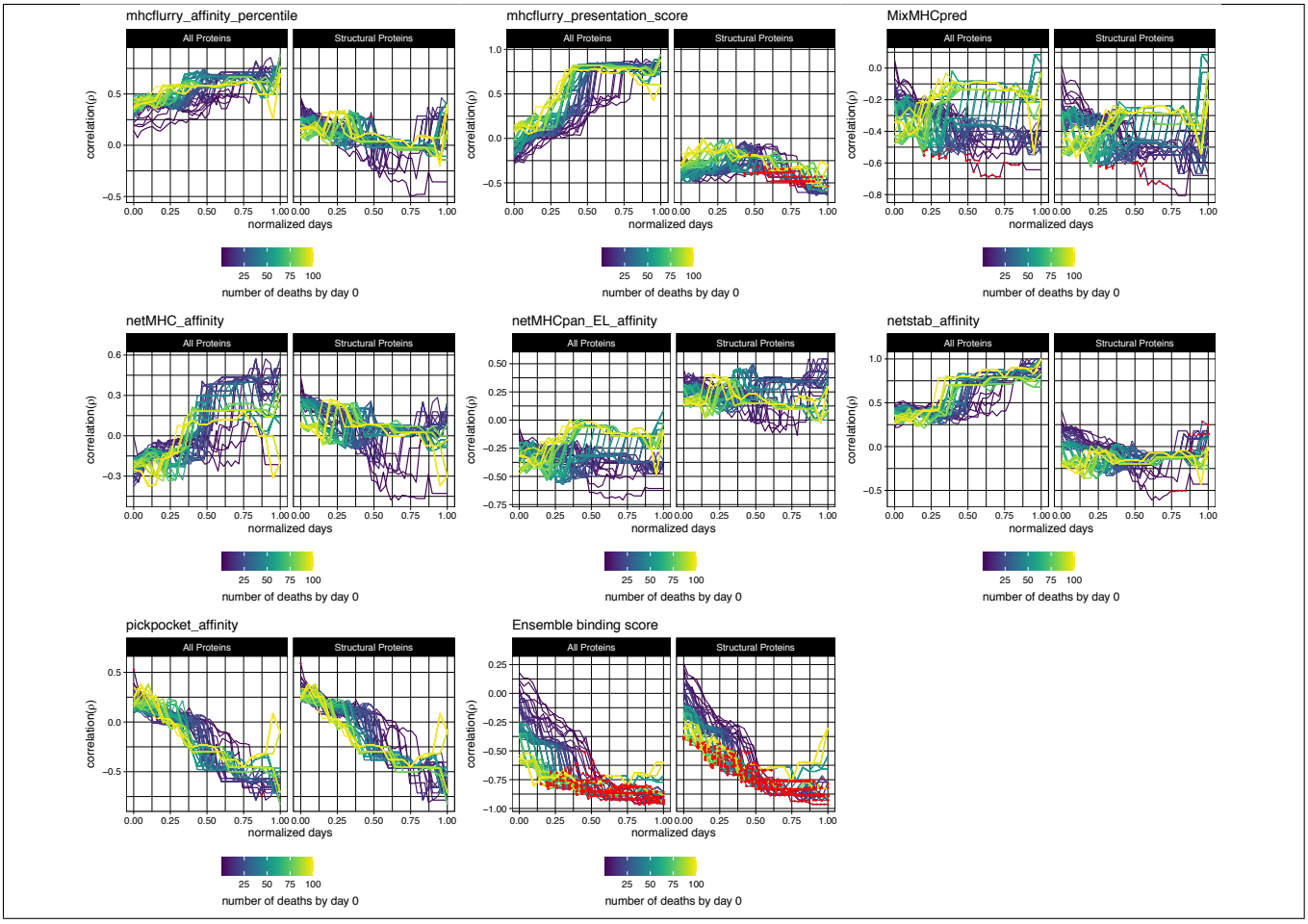

Figure A.12: **Individual algorithms are unable to recreate the correlation reported by EnsembleMHC.** Population presentation scores using only single algorithms were correlated to observed deaths per million. For each algorithm, the algorithm specific population presentation score was calculated from the resulting peptide-allele distribution using global binding affinity cutoffs ($\leq 0.1\%$ for binding percentile scores, top 1% for MHCflurry presentation score, and $\leq 50nm$ for PickPocket). red points indicate a PPV of greater than or equal 95%.