Cell Reports Medicine, Volume *2*

Supplemental information

Total predicted MHC-I epitope load is inversely associated with population mortality from SARS-CoV-2

Eric A. Wilson, Gabrielle Hirneise, Abhishek Singharoy, and Karen S. Anderson

A Supplemental figures: Total predicted MHC-I epitope load is inversely associated with population mortality from SARS-CoV-2

Eric Wilson, Gabrielle Hirneise, Abhishek Singharoy, Karen S Anderson

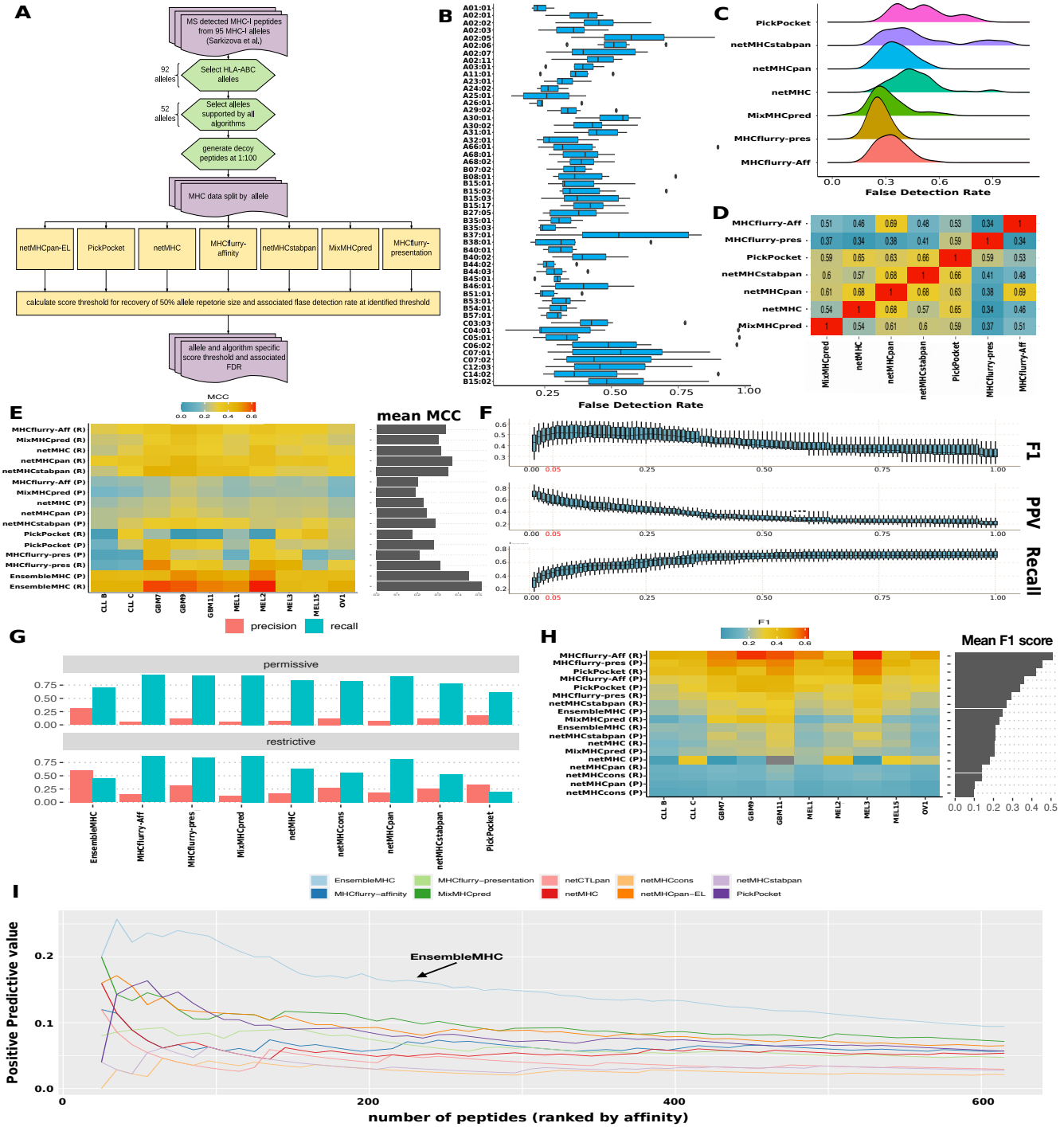

Figure S1: **EnsembleMHC Parameterization overview and viral peptide analysis, Related to figure 1.** (A) EnsembleMHC Parameterization workflow. (B) The EnsembleMHC score algorithm was parameterized using high quality mass spectrometry-detected MHC-I peptides paired with a 100-fold excess of randomly generated decoy peptides. Each bar represents the distribution of algorithm-specific false detection rates (n = 7) at that MHC allele. (C) A density plot of the observed FDRs for each algorithm across all alleles (n = 52). (D) The correlation between individual peptide scores for each algorithm across all alleles was calculated using Pearson correlation. Warmer colors indicate a higher level of correlation while cooler colors indicate lower correlation. (E) Matthew's correlation coefficient was calculated for each algorithm. Warm colors indicate higher MCC while cooler colors indicate lower MCC. The average MCC for each algorithm is represented by the bar plot on the right margin. (F) The effect of different $peptide^{FDR}$ cutoff thresholds on the results reported in **figure 1** was evaluated for a range of 0.01-1. The $peptide^{FDR}$ selected for use in this study is highlighted in red. (G-H) The analysis reported in **figure 1 (A-B)** were repeated with additional comparisons to consensus-based MHC-I prediction algorithms, namely netMHCcons[49] and netCTLpan[48]. (I) The positive predictive value of each algorithm was calculated with respect to ability to identify immunogenic peptides derived from Hepatitis-C genome polyprotein, Dengue virus genome polyprotein, and the HIV-1 POL-GAG protein when selecting n number of top scoring peptides **(STAR METHODS).**

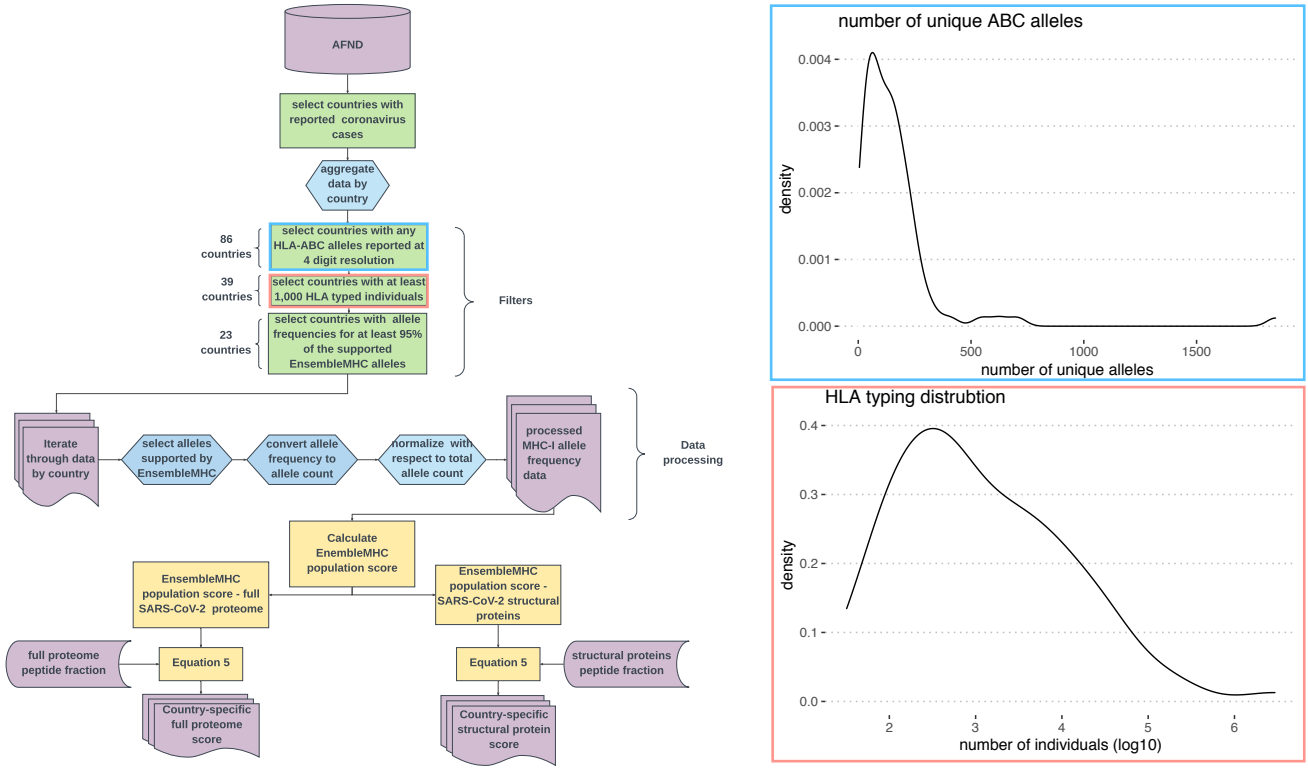

Figure S2: **Data processing and EnsembleMHC population score calculation workflow, Related to figure 3.** The overview of the data processing steps used on the global MHC-I allele frequency data and the calculation of the EnsembleMHC population score with respect to the full SARS-CoV-2 proteome and SARS-CoV-2 structural proteins. **(inset plots)**, The blue inset plot illustrates MHC-typing breadth and depth variation by showing the distribution of the total number of MHC-I alleles reported at 4-digit resolution in 86 countries. The red inset plot shows the distribution of the number of MHC-genotyped individuals in the set of countries with at least 1 reported coronavirus case. **AFND = Allele Frequency Net Database**

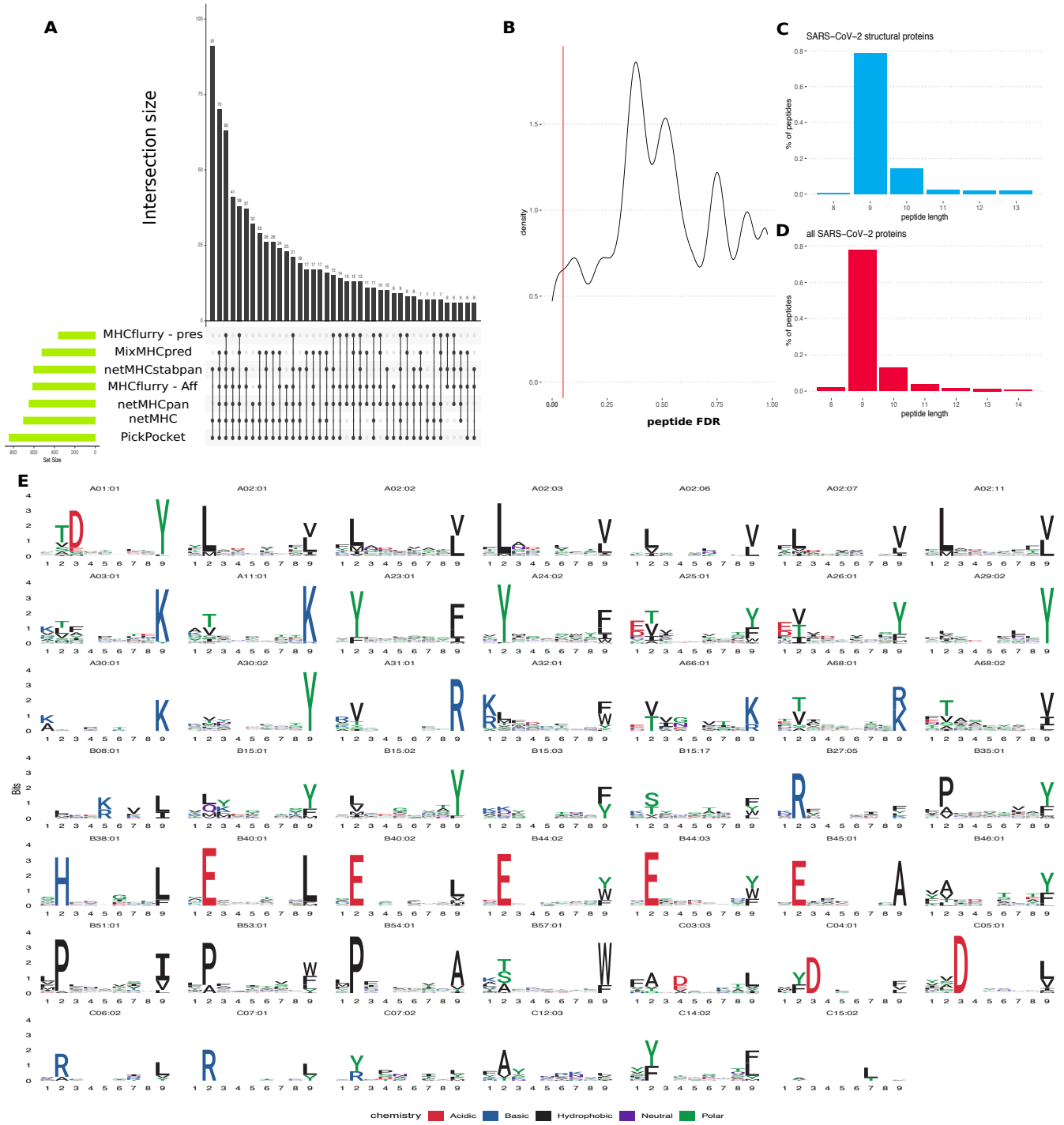

Figure S3: **Characteristics of peptides predicted by EnsembleMHC, Related to figure 2** (A) The UpSet plot shows the contribution of each individual component algorithm to the 658 unique SARS-CoV-2 peptides identified by EnsembleMHC. The top bar plot indicates the number of unique peptides identified by the combination of algorithms shown by the points and segments located under each bar. The bar plot on the left-hand side of the plot indicates the total number of peptides identified by each algorithm. (B) The $peptide^{FDR}$ distribution of the 9,712 SARS-CoV-2 peptides that fell with the score threshold of at least one component algorithm. The red line indicates an $peptide^{FDR}$ level of $\leq 5\%$. (C) The length distribution of the 108 high-confidence peptides identified from SARS-CoV-2 structural proteins. (D) The length distribution of the 658 high-confidence peptides identified from full SARS-CoV-2 proteome. (E) Logo plots were generated for MHC alleles with at least 5 peptides identified by EnsembleMHC. Peptides shorter than 9 amino acids had random amino acid inserted into a non-anchor position while peptides longer than 9 amino acids had a random non-anchor position deleted. Large amino acid character height indicates a high frequency of that amino acid at that position. Amino acids are colored residue type.

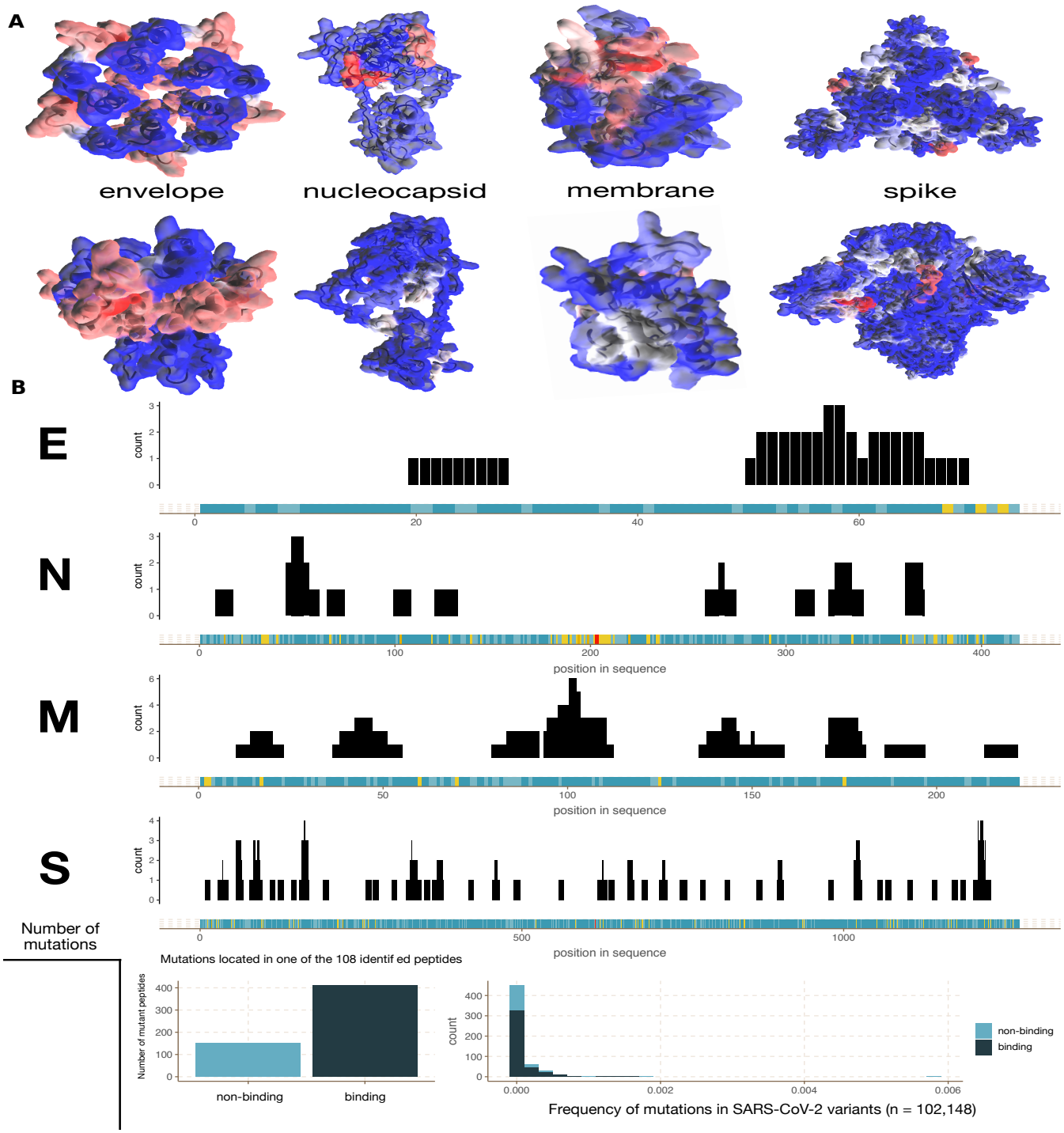

Figure S4: **Molecular origin of predicted SARS-CoV-2 structural protein MHC-I peptides and impact of sequence polymorphism, Related to Figure 2** (A) The predicted SARS-CoV-2 structural protein MHC-I peptides were mapped onto the solved structures for the envelope and spike proteins, and the predicted structures for the nucleocapsid and membrane proteins. Red highlighted regions indicate an enrichment of predicted peptides while blue regions indicate a depletion of predicted peptides. (B) The incidence of protein sequence mutations (colored bar) and the frequency of that position in one of the 108 SARS-CoV-2 structural protein peptides (black bars) were calculated for 102,148 SARS-CoV-2 sequence variants. **Lower left panel,** all potential mutations arising in one of the 108 peptides identified by EnsembleMHC were evaluated for changes in binding affinity ($peptide^{FDR} > 0.05$). **Lower right panel,** The overall frequency of mutations impacting EnsmebleMHC-predicted peptides with light blue indicating deleterious mutations, and dark blue indicating neutral mutations.

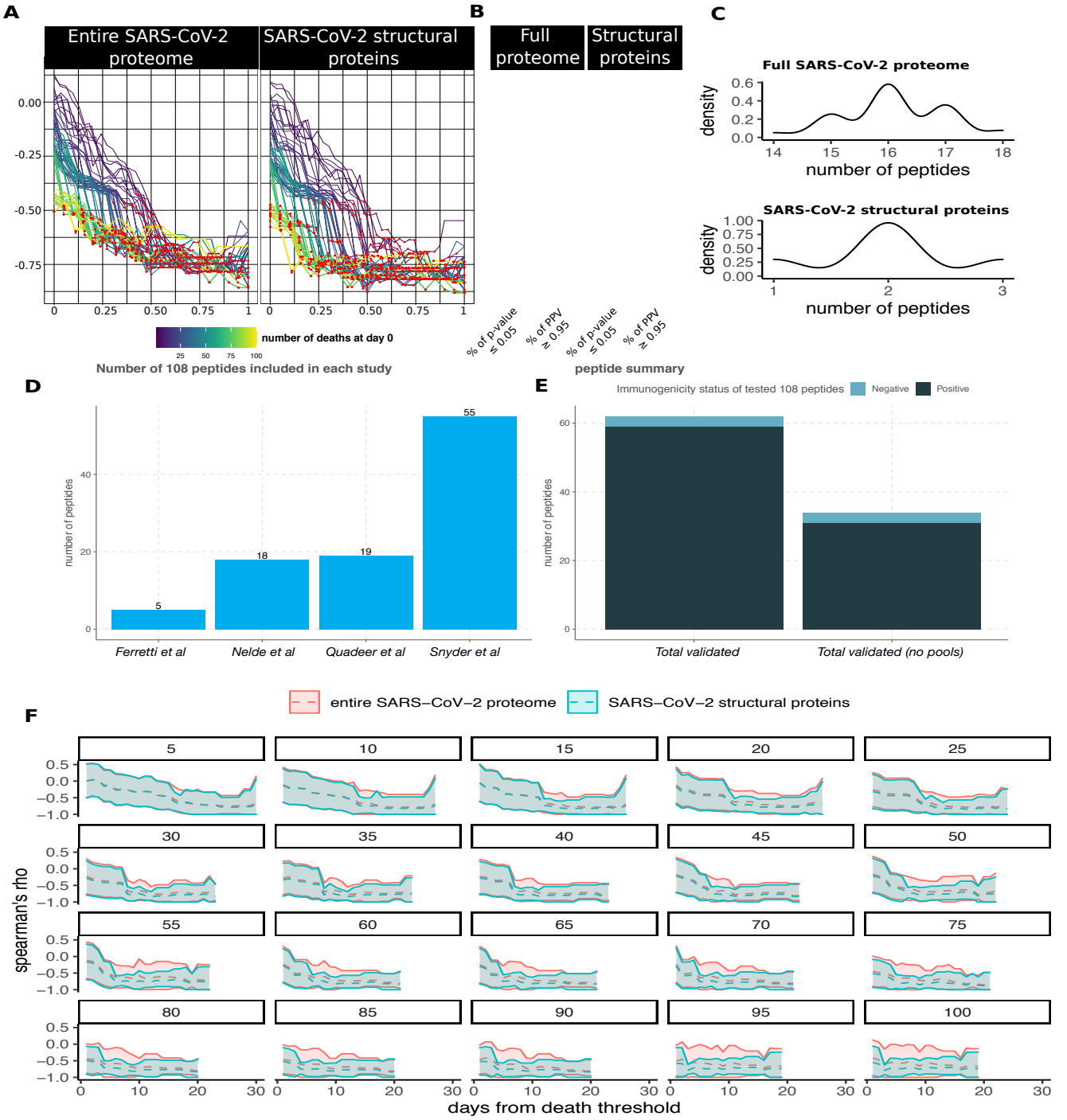

Figure S5: **Comparison of entire SARS-CoV-2 EnsembleMHC population score and structural protein EnsembleMHC population score, Related to Figure 3.** (A) The correlations between EnsembleMHC population score based on the full SARS-CoV-2 proteome **(left)** or only SARS-CoV-2 structural proteins **(right)**. (B) The difference in the proportions of significant p-values and PPV between the full SARS-CoV-2 proteome **(left)** and SARS-CoV-2 structural proteins **(right)** (not corrected for multiple testing). (C) The SARS-CoV-2 peptide-MHC allele distribution resulting from uniform allele sampling. These distribution were used as the partner distributions for the Kolmogorov-smirnov test described in the results. (D) 62 (57%) EnsembleMHC-identified SARS-CoV-2 structural protein peptides were included for testing in 4 different studies. (E) The summary of immunogenicity status of tested EnsembleMHC peptides across all studies. These summaries were split into two groups. *Total validated* indicates the total number of experimentally validated peptides while *total validated (no pools)* indicates the number of experimentally validated peptides excluding those only tested in peptide pools. This distinction was made due to the potential of peptide pools to obscure which tested peptide is truly responsible for the observed immune response. (F) Each individual plot shows the 95% confidence interval (shaded region) for the correlations between EMP scores based on the entire SARS-CoV-2 proteome (red) or SARS-CoV-2 structural proteins (blue) and observed deaths per million for different starting minimum death thresholds (indicated by number above plot).

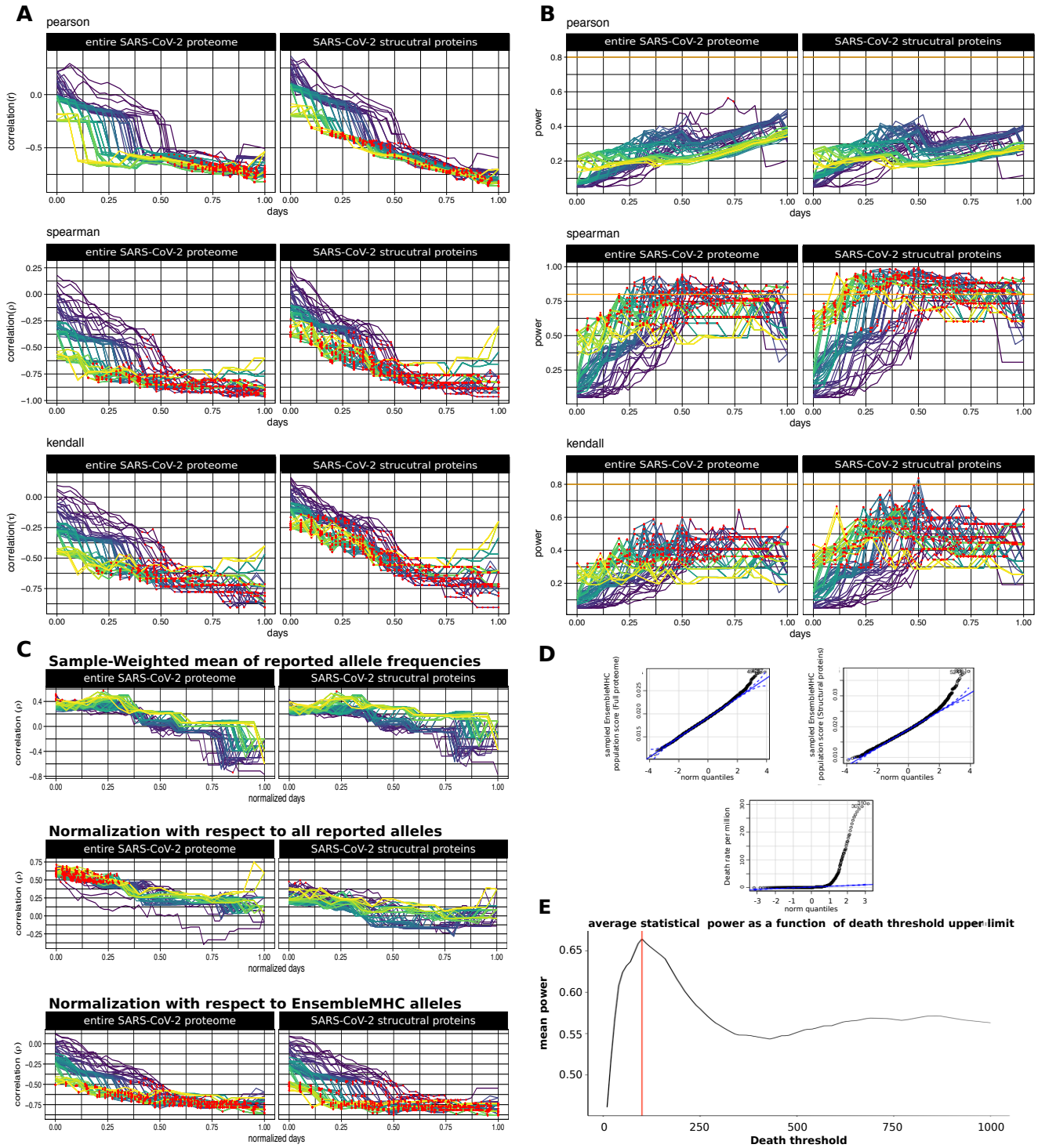

Figure S6: **Justification of statistical tests, Related to STAR METHODS** (A) The correlation between EnsembleMHC population score with respect to all SARS-CoV-2 proteins **(left column)** or SARS-CoV-2 structural proteins **(right columns)** and deaths per million using Pearson's r **(top)**, Spearman's rho **(middle)**, and Kendall's tau **(bottom)**. Correlations that were shown to be statistically significant are colored with a red point. (B) The statistical power of each reported correlation. Correlations that were shown to be statistically significant are colored with a red point. The orange line indicates a power threshold of 80%. (C) The effect of different allele frequency normalization techniques on the reported correlations between SARS-CoV-2 mortality and EMP scores based on the full SARS-CoV-2 proteome (left column) or SARS-CoV-2 structural proteins (right column). **Top panel**, The aggregation of allele frequencies within a particular country by taking the sample-weighted mean of reported frequencies for the 52 selected MHC-I alleles. **Middle panel**, Normalizing allele count with respect to all detected alleles in a given population. **Bottom panel**, Normalizing allele count with respect to only the 52 select alleles. (D) QQ plots were generated from the respective distributions of the full proteome EnsembleMHC population scores, structural protein EnsembleMHC population scores, and deaths per million. To provided more descriptive distributions, EnsembleMHC population scores based on the full SARS-CoV-2 proteome and SARS-CoV-2 structural proteins were calculated for 10,000 simulated countries. Allele frequencies for simulated countries were generated by randomly sampling an observed allele frequency for each of the 52 alleles and re-normalizing to ensure the sum of allele frequencies were equal to one. Points falling outside of the blue lines indicate non-normal data skewing. (E) The mean statistical power of all resulting correlations between EnsembleMHC population scores and observed deaths per million at different minimum reported death thresholds. The red line indicates a minimum death threshold of 100 deaths by day 0, the selected upper limit for analysis.

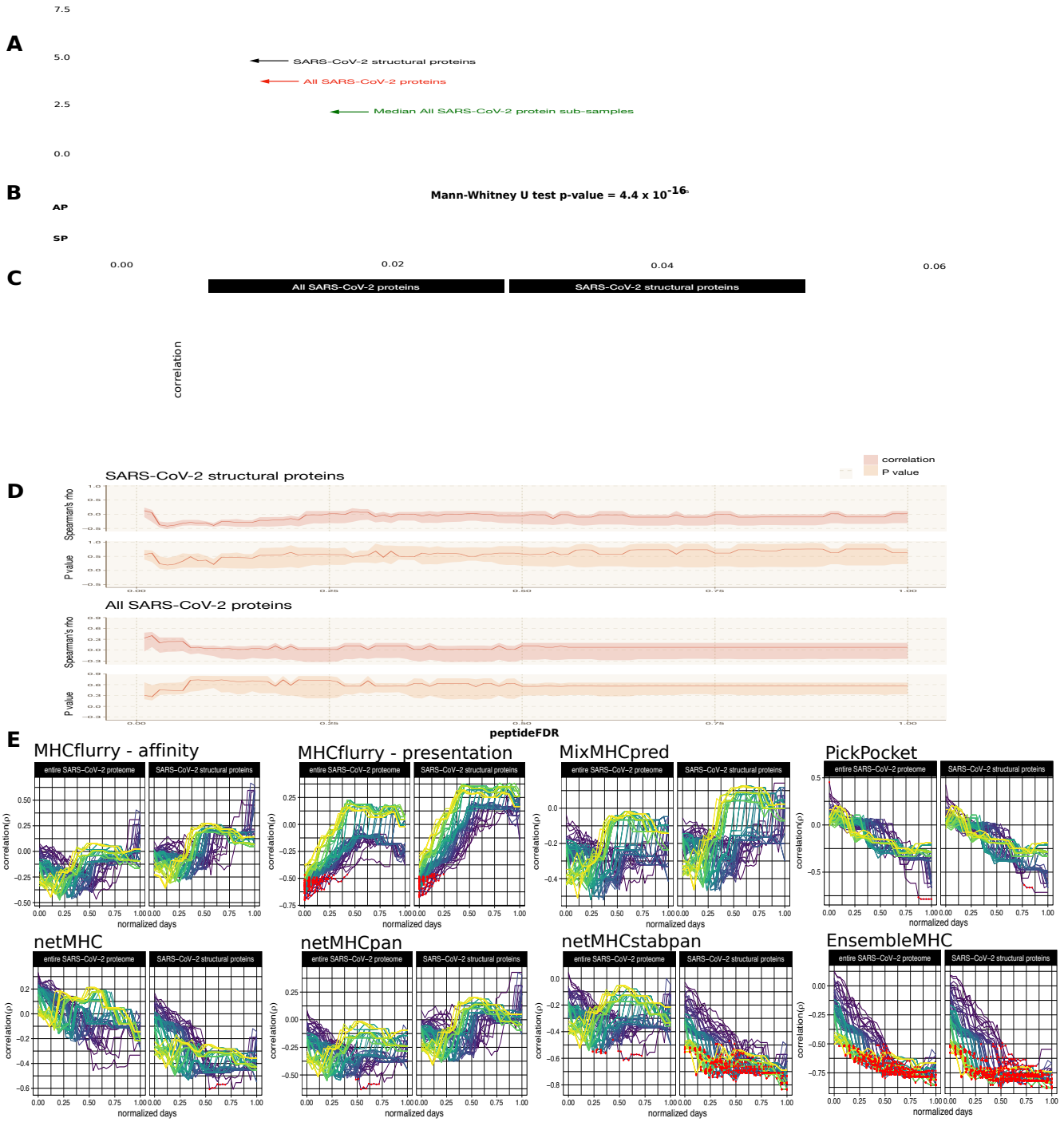

Figure S7: **Robustness of EMP score correlation analysis, Related to Figure 3** (A) 1,000 sub-sampling iterations were performed by randomly selecting 108 peptides from the full SARS-CoV-2 proteome that passed the 5% $peptide^{FDR}$ filter. The correlation between the population EMP score produced by each sub-sampled set of peptides and observed deaths per million were plotted (grey lines). The correlation distribution observed for identified SARS-CoV-2 structural protein peptides (black line), all SARS-CoV-2 proteins (red line), and the median correlation distribution across all subsampling iterations (green line) were plotted for comparison. (B) Kullback-Leibler divergence was calculated for the correlation distribution of each down sample iteration relative to either the correlation distribution of the all peptide group (AP) or the structural peptide group (SP). (C) The MHC-I allele assessment of peptides that passed an individual algorithm binding affinity thresholds were shuffled prior to $peptide^{FDR}$ filtering. The red points indicate correlations with a p-value $\leq 5\%$. (D) The impact of varying $peptide^{FDR}$ cutoff threshold on the shuffled MHC data set. For each $peptide^{FDR}$ cutoff threshold (x-axis), the upper bound of the shaded region indicates the 75^{th} percentile, the lower bound indicates the 25^{th} percentile, and the solid line indicates the median.(E) Population SARS-CoV-2 binding capacities using only single algorithms were correlated to observed deaths per million. For each algorithm, the population SARS-CoV-2 binding capacity was calculated from the resulting viral peptide-MHC allele distribution using restrictive MHC-I binding affinity cutoffs ($\leq 0.5\%$ for binding percentile scores, top 0.5% MHCflurry presentation score, and $\leq 50nm$ for PickPocket). Red points indicate a PPV $\geq 95\%$.

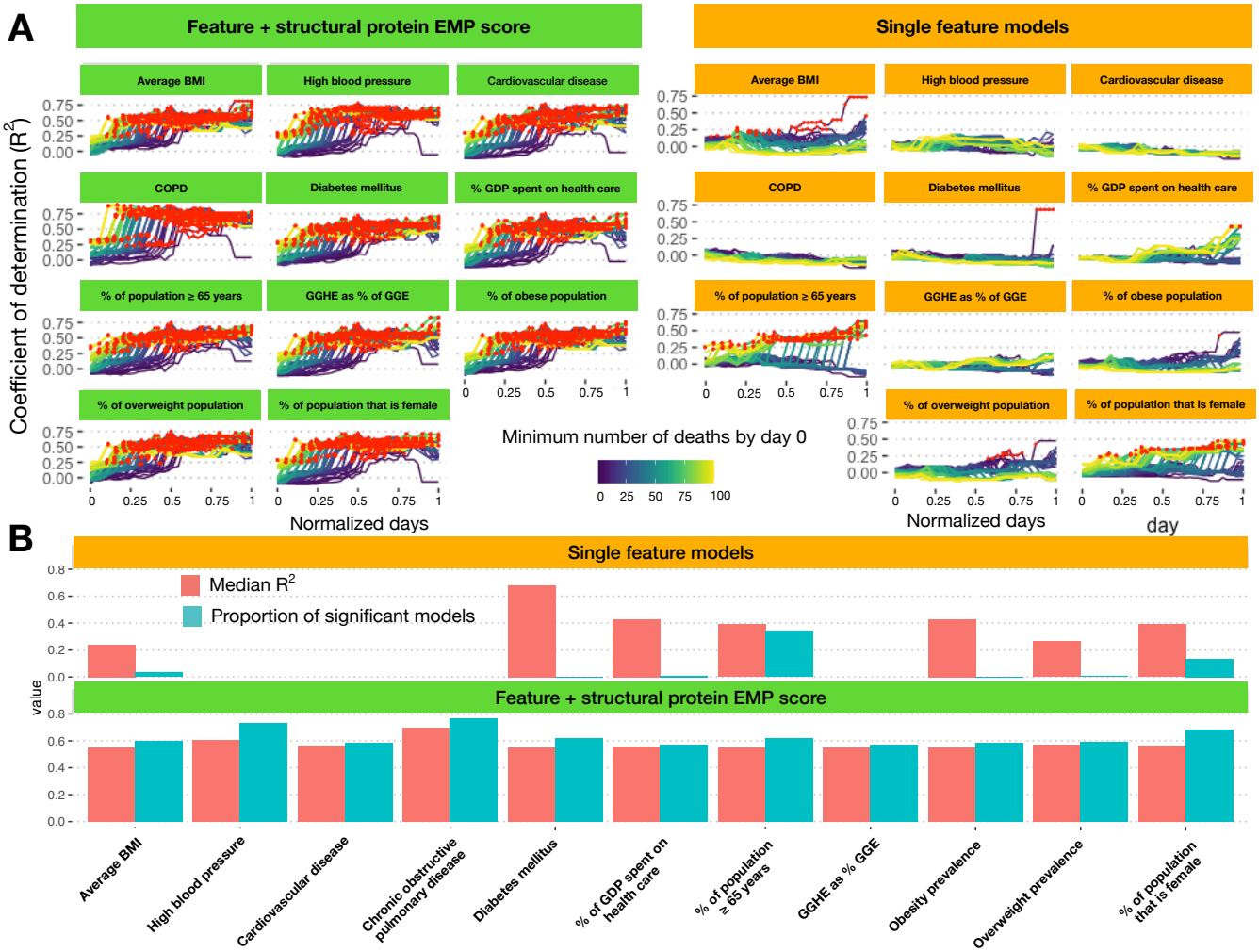

Figure S8: **Addition of structural protein EMP score significantly improves linear model fit to observed deaths per million, Related to Figure 4** (A) Linear models were constructed using either a single risk factor (yellow) or a combination of a risk factor and structural protein EMP scores (green). The x-axis indicates the number of normalized days from when a minimum death threshold was met (line color), and the y-axis indicates the observed adjusted R^2 value. (B), A summary of results obtained from single feature linear models (top panel, yellow) or the combination models (bottom panel, green). The red bars indicate the median R^2 value achieved by that model and the blue bars indicate the proportion of regressions that were found to be significant (F-test ≤ 0.05).

A

Countries
China
Japan
South Korea
US
France
Germany
India
Italy
Russia
UK
Iran
Israel
Croatia
Romania
Netherlands
Mexico
Ireland
Czechia
Morocco

B

Factor	Abbreviation	Description
% of population ≥ 65 years	65	Percentage of the population that is 65 years of age or older (2020).
Average BMI	Avg. BMI	The age-standardized average population body mass index (2016).
Cardiovascular disease	CD	The deaths per million due to cardiovascular disease (2016).
Chronic obstructive pulmonary disease	COPD	The deaths per million due to complications from chronic obstructive pulmonary disease (2016).
Diabetes mellitus	DM	The deaths per million due to complications from diabetes mellitus (2016).
High blood pressure	BP	The age-standardized percentage of the population with a systolic blood pressure ≥ 140 or diastolic blood pressure ≥ 90 (2015).
Obesity prevalence	OBS	The age-standardized percentage of the population with a BMI ≥ 30 (2016).
Overweight prevalence	OVW	The age-standardized percentage of the population with a BMI ≥ 25 (2016).
Structural protein EMP score	SP	The SARS-CoV-2 structural protein presentation score.
% of GDP spent on health care	GDP	Current health expenditure (CHE) as percentage of gross domestic product (2017).
% of total gov. expenditure on health care	GGHE	General government expenditure on health as a percentage of total (2014).
% of population that is female	SEX	The proportion of the total population that is female (2020).

Table S2. Socioeconomic and health-related risk factors, Related to Figure 4. (A) 21 countries were selected for analysis based on the existence of data in the Global Health Observatory data repository and inclusion in the 23 country set used for EMP score analysis. (B) Descriptions and abbreviations for the selected risk factors. Each factor is labeled with the year that the data was collected. In every case, the most recent data was selected for analysis.