



# Class I HLA Allele Predicted Restricted Antigenic Coverages for Spike and Nucleocapsid Proteins Are Associated With Deaths Related to COVID-19

Marco Antônio M. Pretti<sup>1,2</sup>, Rômulo G. Galvani<sup>1,3,4</sup>, Gustavo Fioravanti Vieira<sup>5,6</sup>, Adriana Bonomo<sup>4,7</sup>, Martín H. Bonamino<sup>2,7</sup> and Mariana Boroni<sup>1\*</sup>

<sup>1</sup> Laboratory of Bioinformatics and Computational Biology, Division of Experimental and Translational Research, Brazilian National Cancer Institute (INCA), Rio de Janeiro, Brazil, <sup>2</sup> Program of Immunology and Tumor Biology, Division of Experimental and Translational Research, Brazilian National Cancer Institute (INCA), Rio de Janeiro, Brazil, <sup>3</sup> Faculdade de Biomedicina, Universidade Veiga de Almeida, Rio de Janeiro, Brazil, <sup>4</sup> Laboratory for Thymus Research (LPT), Oswaldo Cruz Institute, Oswaldo Cruz Foundation (FIOCRUZ), Rio de Janeiro, Brazil, <sup>5</sup> Laboratory of Immunoinformatics (NBLI), Department of Genetics, Institute of Biosciences, Federal University of Rio Grande do Sul, Porto Alegre, Brazil, <sup>6</sup> Laboratory of In Silico Human Health, Post Graduate Program in Health and Human Development, La Salle University, Canoas, Brazil, <sup>7</sup> Vice-Presidency of Research and Biological Collections (VPPCB), Oswaldo Cruz Foundation (FIOCRUZ), Rio de Janeiro, Brazil

## OPEN ACCESS

### Edited by:

Linda F. Van Dyk,  
University of Colorado Denver,  
United States

### Reviewed by:

Martin Maiers,  
National Marrow Donor Program,  
United States  
John Sidney,  
La Jolla Institute for Immunology,  
United States

### \*Correspondence:

Mariana Boroni  
mariana.boroni@inca.gov.br

### Specialty section:

This article was submitted to  
Viral Immunology,  
a section of the journal  
Frontiers in Immunology

**Received:** 26 May 2020

**Accepted:** 16 November 2020

**Published:** 16 December 2020

### Citation:

Pretti MAM, Galvani RG, Vieira GF, Bonomo A, Bonamino MH and Boroni M (2020) Class I HLA Allele Predicted Restricted Antigenic Coverages for Spike and Nucleocapsid Proteins Are Associated With Deaths Related to COVID-19. *Front. Immunol.* 11:565730. doi: 10.3389/fimmu.2020.565730

The world is dealing with one of the worst pandemics ever. SARS-CoV-2 is the etiological agent of COVID-19 that has already spread to more than 200 countries. However, infectivity, severity, and mortality rates do not affect all countries equally. Here we consider 140 HLA alleles and extensively investigate the landscape of 3,723 potential HLA-I A and B restricted SARS-CoV-2-derived antigens and how 37 countries in the world are predicted to respond to those peptides considering their HLA-I distribution frequencies. The clustering of HLA-A and HLA-B allele frequencies partially separates most countries with the lowest number of deaths per million inhabitants from the other countries. We further correlated the patterns of *in silico* predicted population coverage and epidemiological data. The number of deaths per million inhabitants correlates to the predicted antigen coverage of S and N derived peptides and its module is influenced if a given set of frequent or rare HLA alleles are analyzed in a given population. Moreover, we highlighted a potential risk group carrying HLAs associated with an elevated number of deaths per million inhabitants. In addition, we identified three potential antigens bearing at least one amino acid of the four-length insertion that differentiates SARS-CoV-2 from previous coronavirus strains. We believe these data can contribute to the search for peptides with the potential to be used in vaccine strategies considering the role of herd immunity to hamper the spread of the disease. Importantly, to the best of our knowledge, this work is the first to use a populational approach in association with COVID-19 outcome.

**Keywords:** COVID-19, SARS-CoV-2, Betacoronavirus, ligandome, population coverage analysis, herd immunity

## INTRODUCTION

Since December 2019, the world has been facing one of the worst pandemics of all times caused by a novel *Betacoronavirus* (Severe Acute Respiratory Syndrome Coronavirus 2—SARS-CoV-2). Until this day, more than 20 million cases and 732 thousand deaths were confirmed. However, the infection does not strike all nations equally when considering multiple epidemiological parameters (1). The viral genome of this etiological agent consists of structural protein genes such as Spike (S), Envelope (E), Membrane (M), and Nucleocapsid (N) and non-structural protein genes (e.g., ORF1ab, ORF3a, ORF6) (2). Their functions have been extensively reviewed elsewhere (2, 3). Once in the human body, SARS-CoV-2 infect cells using the surface molecules ACE2 and TMPRSS2 highly expressed in the lungs and respiratory tract. The infection is recognized by macrophages and alveolar epithelial cells that initiate a pro-inflammatory response, which may trigger an acute respiratory distress syndrome among other symptoms (4–7). The SARS-CoV-2 viral proteins have been shown to elicit both cellular immunity and humoral response (8). One of the most promising viral proteins for vaccine development is the S protein due to its accessibility to antibodies and pivotal role in the infection (9). Nevertheless, antibody response appears first against the N protein, the most abundant protein in coronavirus, and a few days later for S protein (8, 9). The ability to trigger an adaptive immune response greatly relies on the ability to present antigens through the class I and II Histocompatibility Leukocyte Antigen (HLA) molecules. Peptides derived from different SARS-CoV-2 proteins are recognized by CD4 and CD8 T cells from COVID-19 convalescent patients (10). The allelic distribution of the largely polymorphic HLA genes across different countries may affect the appropriate antigen presentation of SARS-CoV-2-derived peptides among distinct populations (11). This genetic variability could be associated with specific profiles of disease susceptibility and severity (1). Some studies predicting HLA-restricted peptides derived from SARS-CoV-2 are already available (12, 13). Therefore, these characteristics must be taken into account when considering epitope-based vaccines since target epitopes must not only be able to bind to the HLA molecule but also prove to be immunogenic to promote a functional response (14). Populational coverage studies considering these features may help narrow the number of candidate peptides to be considered for vaccination strategies (15). However, epitope coverage studies are not sufficient to predict if the immune response would be able to resolve the infection and provide immunity to individuals and/or if this immunization occurs at the expense of exacerbated immune responses. In the context of the SARS-CoV-2 pandemic, the susceptibility to severe cases and deaths may be related to the over or under presentation of SARS-CoV-2-derived peptides by HLA molecules as well as to individual characteristics driving inflammation (3). This is not necessarily restricted to HLA allele distribution among populations since the same peptide can be presented by more than one HLA allele. With that in mind, we evaluated how different populations potentially present HLA class I restricted SARS-CoV-2 peptides and their association with

the local epidemic progression features across different countries. This strategy led us to a set of peptides derived from virus S and N proteins, which showed distinct correlations with disease outcomes. While S may have a protective role as they are potentially presented in HLA alleles with high frequency in countries with less number of deaths per million inhabitants, N has a direct correlation with the unfavorable endpoint. Importantly, the patterns observed are dependent on frequent HLA alleles, and opposite correlations are found when evaluating only rare alleles. These results may prove critical in guiding efforts towards the development of vaccine candidates, as well as determine high-risk groups based on HLA alleles.

## MATERIAL AND METHODS

### Epidemiological Data

The epidemiological data used in this analysis (number of total cases, total deaths, total recovered, serious critical cases, total cases per 1 million of inhabitants, and total deaths per 1 million inhabitants) were retrieved on the Worldometer database ([www.worldometers.info/coronavirus](http://www.worldometers.info/coronavirus), accessed at 01:18PM, May 17, 2020). This database collects data from official reports, directly from Government's communication channels or indirectly, through local media sources when deemed reliable. The 50 countries with the highest death per million inhabitants and more than 1,000 total cases were evaluated. The epidemiological data used can be found in the **Supplementary File 1**.

### HLA-I Gathering

Human Leukocyte Antigen (HLA) I alleles for A and B genes with at least two fields and its frequencies in the selected populations were retrieved from the Allele Frequency Net Database (16) between 23/04/2020 and 07/05/2020. Larger datasets that were not part of anthropological studies nor minority ethnic groups, and with at least 50 individuals were preferred (**Supplementary File 2**). The alleles were ordered from the most frequent to the least frequent and the AF were summed to obtain a cumulative AF of 0.75 or 0.90 separately for HLA-A and B. The values used were the closest to the threshold (equal, above, or below) considering the difference between the cumulative AF and the threshold itself. Countries that haven't reached these thresholds with a difference inferior to 0.04 for AF = 0.75 and 0.02 for AF = 0.9 were excluded. The minor alleles (AF = 0.15) are the alleles exclusive from the dataset generated when using 0.90 as a threshold. AF with resolution greater than two fields were combined, e.g. for BRA: B\*07:02:01 and B\*07:02:03 became B\*07:02. The HLA alleles for four ethnic backgrounds (African American, Caucasian, Hispanic, and Asian) were obtained for the USA population using the cumulative threshold of 0.75 and analyzed as previously described.

### Accessing Experimental Binding Affinities

We downloaded MHC Ligand assays at the Immune Epitope Database (IEDB) on 18/05/2020 that matched the parameters "Linear Epitope," Organism "Virus," positive assays only, MHC Ligand Assays, MHC Restriction: "MHC Class I," HLA-A and HLA-B, Host: "Humans." Assays with NA values

for “Quantitative.measurement” and entries with “>”, “≤” or “<” in “Measurement.Inequality” were excluded. Also, we retrieved T cell assays using the same parameters except for the positive assays only. The IEDB classifications for ligand assays take into account the information provided in deposited papers: peptides categorized into Pos-L/-I/-H have been grouped according to an IC50 threshold. T cell assays are classified into Positive or Negative taking into account in vitro assays such as ELISPOT and chromium release.

## Binding Predictions

The translated proteome from SARS-CoV-2 genome version BetaCoV\_Wuhan\_IPBCAMS-WH-01\_2019 was downloaded on 23/03/2020 from the NCBI repository (<https://www.ncbi.nlm.nih.gov/nucore/MT019529>). Binding predictions were performed using netMHCpan4 (17) with peptide lengths from 8 to 11 for the entire viral genome. We used blastp v2.9.0 to eliminate peptides that matched protein sequences from the Ensembl human genome GRCh38 with 100% of identity and same query length. Next, we performed predictions with netCTLpan to also remove peptides with a TAP score smaller than 0 and a proteasome cleavage score smaller than 0.5 to assure that our analysis only includes peptides with a high probability of being processed by proteasome and TAP. Only strong binder peptides (%Rank <0.5) were kept for further analysis. We used an online tool (Stothard, 2000) to shuffle ligand sequences prior to binding prediction as a control.

## Population Coverage

We used the tool described by Bui and collaborators (14) available at IEDB (<http://tools.iedb.org/population/>) to calculate the population coverages. The inputs used were the same AF for the binding predictions (**Table/Supp file**) and the Strong Binder peptides for each HLA-I allele. The analysis was performed using the “Class I separate” option. Coverages were calculated for the entire SARS-CoV-2, and separately for S, N, M, E, and ORF1ab viral proteins. Area Under the Curve (AUC) was calculated prior to Spearman correlations.

## Statistical Analysis

Significant results for correlations were considered when p-value < 0.05. A bootstrap with 1,000 replicates was performed to access the confidence interval for Spearman correlations with epidemiological data. We used the complete method and 1-Pearson correlation coefficient as the distance to cluster rows and columns of the heatmaps. Cluster stability for heatmaps was calculated using Jaccard distancing through bootstrapping using 10,000 replacements and cluster method “disthclustCBI.” Spearman correlations were used unless stated otherwise.

## Data Analysis

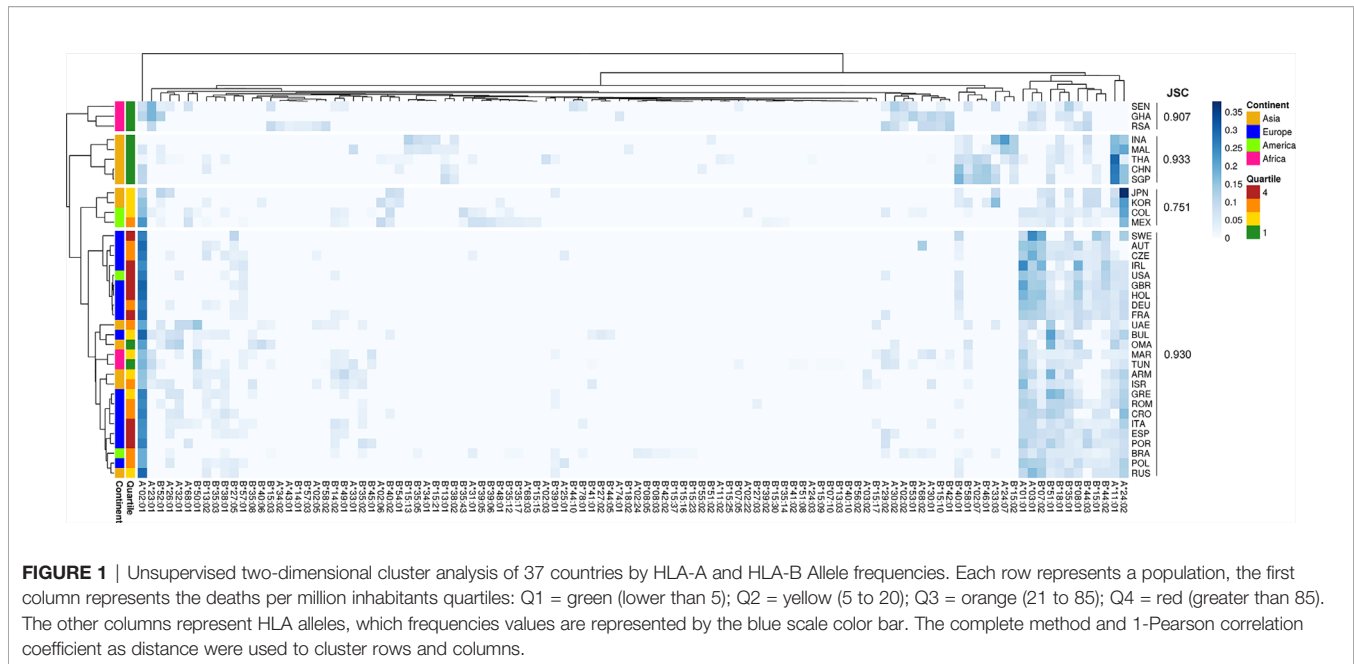
The analysis was conducted using the R environment version 4.0 and the following packages ggpubr v0.2.1, ggplot2 v3.2.1, data.table v1.12.2, gridExtra v2.3, readxl v1.3.1, dplyr v0.8.3, tidytext v0.2.3, pheatmap v1.0.12, RColorBrewer v1.1, cowplot v1.0, GenomicRanges v1.31.1, IRanges v2.18.1, stats v4.0.2, boot v1.3, ggfortify v0.10, and ggbio v1.3.2.

## RESULTS

### HLA-I Alleles Partially Discriminate Different Outcomes Among Countries

To assess the landscape of peptides potentially presented by HLA-I alleles occurring in the global population and how this diversity is associated with the disease outcome, we selected a list of countries hit by the pandemic with known HLAs genetic frequencies. Only countries with more than 1,000 total cases of Coronavirus disease (COVID-19) on 17/05/2020 according to the Worldometer database (18) register were considered, reaching a total of 50 different countries. Two field resolution HLA-A and HLA-B alleles for the aforementioned populations available in the Allele Frequency Net Database (19) were accessed. For each population, a different number of alleles was retrieved, considering the allele in decreasing order of their frequencies to reach a cumulative allelic frequency (AF) as close to 0.75 as possible. Populations whose HLA AF information did not reach the threshold or having less than 50 individuals in the HLA cohort were removed, totalizing 37 countries in our study (**Supplementary File 1, Figure S1**) and 106 HLA alleles. The median cumulative AF varied among populations with a median of 0.752 for HLA-A (minimum of 0.721 for AUT and maximum of 0.781 for GRE) and 0.739 for HLA-B (minimum of 0.714 for SWE and maximum of 0.750 for MEX). The number of HLA-I alleles ranged from 4 to 14 for HLA-A (median = 7) and from 6 to 26 for HLA-B (median = 13, **Table S1**).

Different epidemiological data associated to COVID-19 were retrieved from the Worldometer database (see *Methods, Supplementary File 1*). We then investigated whether different HLA-I alleles among populations could be associated with these epidemiologic features. We chose to use death per million inhabitants as the main epidemiological metric as it is the one that varies least depending on the testing capacity or policy adopted by the country studied (20). To evaluate if the distribution of HLA allele frequencies among populations could explain death per million inhabitants, we performed an unsupervised hierarchical clustering. We used HLA-A and HLA-B frequencies for each particular country, that were divided into quartiles according to the epidemiological parameter (**Figure 1**). We observed, as expected, that geographic location strongly influences the clustering. Cluster 1 is composed exclusively by African countries (pink, RSA, GHA, and SEN), corresponding to 60% of all African countries analyzed, with a Jaccard similarity coefficient (JSC) of 0.91. Cluster 2, also homogeneous regarding the geographic composition, is composed only of Asian countries (yellow, INA, MAL, THA, CHN, and SGP) corresponding to 41% of all Asian countries evaluated in this study, with a JSC of 0.93 (**Figure 1**). However, we observed that all countries comprising clusters 1 and 2 belong to the first quartile (Q1, green, less number of deaths per million). Therefore, although the HLA frequency clustering recaps the country's continental distribution, it also appears to separate the first quartile apart from the remaining groups. The most prevalent HLA-I alleles in Q1 are HLA-A\*24:02, HLA-A\*11:01, HLA-B\*40:01, and HLA-



A\*23:01 while in Q4 (red, more number of deaths per million) are HLA-A\*02:01, HLA-A\*01:01, HLA-A\*03:01, and HLA-B\*07:02. Using a Principal Component Analysis (PCA), the same result was observed: despite the clustering by continents, countries belonging to the first quartile (green) are separated from those of the other quartiles (Figure S2). This result highlights the contribution of HLA alleles to segregate distinct responses to the COVID-19 infection among countries. We reasoned that describing the binding spectra of each allele for peptides related to SARS-CoV-2 could further contribute to a better understanding of the epidemiologic data. Therefore, we next sought to perform a thorough characterization of SARS-CoV-2 derived peptide binding prediction to HLA-I alleles.

### HLA-I Binding of SARS-CoV-2-Derived Peptides Among Different Populations

In order to study the binding of SARS-CoV-2 peptides to different HLA alleles, we predicted the affinities of HLA alleles (33 HLA-A and 73 HLA-B, Supplementary File 2) for each potential SARS-CoV-2 derived peptide using the netMHCpan4 tool (17). Only peptides that met the criteria of scoring high probabilities of being transported by TAP, and processed by the proteasome [predicted both by netCTLpan (21)] were considered as the final dataset (Figure S3A, B). The 38,464 predicted 8–11 mer unique peptides were derived from the complete genome of the SARS-CoV-2. Of those, 18,799 unique peptides passed the TAP transport score, 12,230 the proteasome cutting, and 10,909 both filters (Figure S3C). Peptides were categorized as Strong Binders (SB, %Rank < 0.5), Weak Binders (WB, 0.5 ≤ %Rank < 2), and Non-binder (NB, %Rank ≥ 2) based on the %Rank definition by the developers, which accounts for the probability of a peptide to bind the HLA given a pool of natural ligands. As

expected, their mean predicted affinity to HLA alleles varies accordingly (Figure S4A). We confirmed that the number of SB peptides increases with the number of HLA-I alleles accordingly for a given population by testing the existence of a linear correlation between these two variables ( $R = 0.95$ ,  $p$ -value < 0.01, Pearson correlation test, Figure S3D). Even considering that this correlation could be expected, it is important to emphasize that the number of SB peptides restricted to a given HLA allele varies greatly and shall affect a given population antigen coverage.

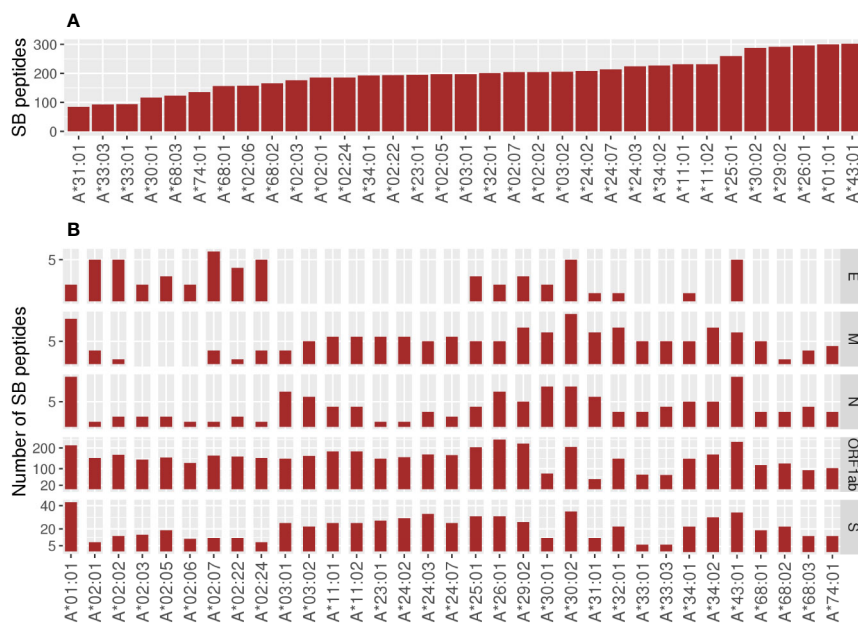
In order to define the best group of peptides to work with, we compared the predicted *in silico* affinities with affinities of viral-derived peptides obtained by experimental assays retrieved at the Immune Epitope Database (IEDB). A total of 49,207 MHC ligand assays (38,558 for HLA-A and 10,649 for HLA-B) were retrieved. Of note, half of the HLA alleles assessed in the *in vitro* binding assays selected from the IEDB are shared by the HLAs used for predictions in our analysis. They were previously classified by the IEDB into Positive-High (Pos-H, 17,353), Positive-Intermediate (Pos-I, 14,927), or Positive-Low (Pos-L, 16,927) according to the results of *in vitro* assays readout. We also predicted the affinity of 18,522 unique peptide/HLA pairs retrieved from the IEDB whose capacities to elicit a TCR response were tested through 34,872 T cell assays classified into Positive (Pos, 2,705) or Negative (Neg, 15,817). In summary, the predicted affinities of peptides classified into SB peptides are equivalent to the vast majority of peptides that were proven to bind the HLA experimentally (Pos-H and Pos-I, Figures S4A, B), potentially eliciting an immune response. A significant difference in the predicted affinities was also observed when comparing Pos and Neg peptides from T cell assays (Wilcoxon,  $p < 0.0001$ , Figure S4C). Overall predicted



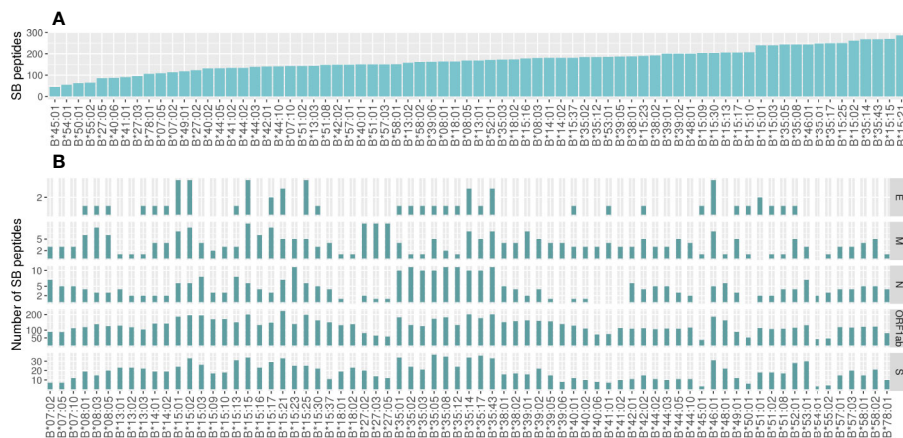
affinities differed for HLA-A and B molecules (**Figure S4D**) but no difference in binding affinities was observed for IEDB ligand assays (**Figure S4E**). Therefore, we decided to work only with the SB peptides in the downstream analyses, totaling 3,723 peptides that passed both TAP and proteasome filtering (**Figure S3C**).

Considering all SB peptides predicted to bind to the 106 analyzed alleles, none of them matched with 100% identity the sequences found in the human proteome using BLAST search. Different HLA alleles can bind to different peptide repertoires, which is defined by the amino acids that compose the cleft. Alleles that bind to a large repertoire of peptides are called generalists and those that bind to a smaller repertoire of peptides can be called specialists (22). The HLA-A alleles with the capacity to present the higher number of SB peptides were HLA-A\*43:01 (302 unique peptides), HLA-A\*01:01 (300 unique peptides), and HLA-A\*26:01 (296 unique peptides). In contrast, HLA-A\*31:01 (84 unique peptides) was predicted to present the lowest number of SB peptides (**Figure 2A**). The number of SB peptides potentially presented by each allele varies greatly depending on the virus open reading frames (ORF) considered, and it is proportional to ORF sizes (**Figure 2B, Table S2**). Some SB peptides were predicted to be presented by multiple HLA alleles and, as so, evidenciate the discrepancy between the number of peptide:HLA pairs and the number of unique SB peptides derived from specific viral ORFs (**Table S2**). For instance, HLA-A\*01:01 is predicted to present the highest number of SB peptides (43 unique peptides) derived from protein

S (1,274 amino acids—aa) and a moderate number of SB peptides (10 unique peptides) derived from the protein N (420 aa, **Figure 2B**). Two of the top five most promiscuous (generalist) HLA-A regarding SARS-CoV-2 are country-specific (HLA-A\*43:01 from RSA and HLA-A\*01:03 from UAE). The most generalist HLA-A (HLA-A\*01:01) is found in low frequency in countries that comprise the first quartile (median AF = 0.000, mean = 0.0252) and high AF in countries from the fourth quartile (median AF = 0.130, mean = 0.148). Regarding the more SARS-CoV-2 fastidious (specialists) HLA-A, HLA-A\*33:03 is present in countries that comprise the first quartile (median AF = 0.021, mean = 0.0468), but is absent in countries belonging to the fourth quartile. The globally common HLA-A\*02:01 is predicted to bind a low number of S- and N-derived SB peptides when compared to the other alleles. On the other hand, it is one of the alleles with the higher number of SB peptides from the ORF1ab region, thus being a specialist allele for this region. In contrast, the HLA-A\*11:01, A\*23:01, and A\*24:02 alleles (more prevalent in the countries belonging to the first quartile) seem to act as specialized alleles for the S regions, having the potential to present a greater number of peptides from this region than from the N region (**Figure 2B**). As for HLA-A, the predicted binding spectrum for HLA-B varies among alleles in which HLA-B\*15:21, HLA-B\*15:15, and HLA-B\*35:43 possess the higher number of SB peptides (286, 269, and 268 unique peptides, respectively), and HLA-B\*45:01 (45 unique peptides) the lowest (**Figure 3A**). The globally common HLA-B\*07:02 and HLA-



**FIGURE 2** | Number of unique Strong Binder peptides for the 33 analyzed HLA-A alleles. **(A)** Global view of the predicted binding spectra for HLA-A considering the entire SARS-CoV-2 genome or **(B)** depicted by ORF1ab, Envelope (E), Membrane (M), Nucleocapsid (N), or Spike (S) proteins. Strong Binder peptides were selected using a %Rank cut-off of 0.5%, which is based on the likelihood of this peptide being presented when compared to a pool of natural ligands. HLA binding predictions performed by netMHCpan were crossed with TAP and proteasome predictions performed by netCTLpan in order to get a more reliable set of peptides. Importantly, the same peptide can be predicted to be presented by more than one HLA.



**FIGURE 3** | Number of unique Strong Binder peptides for the 73 analyzed HLA-B alleles. **(A)** Global view of the predicted binding spectra for HLA-A considering the entire SARS-CoV-2 genome or **(B)** depicted by ORF1ab, Envelope (E), Membrane (M), Nucleocapsid (N), or Spike (S) proteins. Strong Binder peptides were selected using a %Rank cut-off of 0.5%, which is based on the likelihood of this peptide being presented when compared to a pool of natural ligands. HLA binding predictions performed by netMHCpan were crossed with TAP and proteasome predictions performed by netCTLpan in order to get a more reliable set of peptides. Importantly, the same peptide can be predicted to be presented by more than one HLA.

B\*44:02 are predicted to bind low numbers of SB peptides from protein S and N. In contrast, the HLA-B\*15:02 (only found in countries belonging to the first quartile) is predicted to bind high numbers of SB proteins from the S region while having the potential to present low numbers of peptides from N protein (**Figure 3B**).

### Distinct Patterns of Antigen Coverage Among Populations Are Observed Within Different ORFs

Predicted HLA-restricted epitopes capable of binding with substantial affinity to the HLA alleles provides a means of addressing population coverage related to HLA polymorphism and different HLA binding specificities. We, therefore, calculated the percentage of individuals from each population predicted to respond to class I SARS-CoV-2 derived SB epitope set on the basis of HLA genotypic frequencies and HLA binding data, using the population coverage calculation tool available through the IEDB. This allowed us to obtain the number of Epitopes/HLA combinations that cover each population at a given rate (**Table S3, Supplementary File 3**).

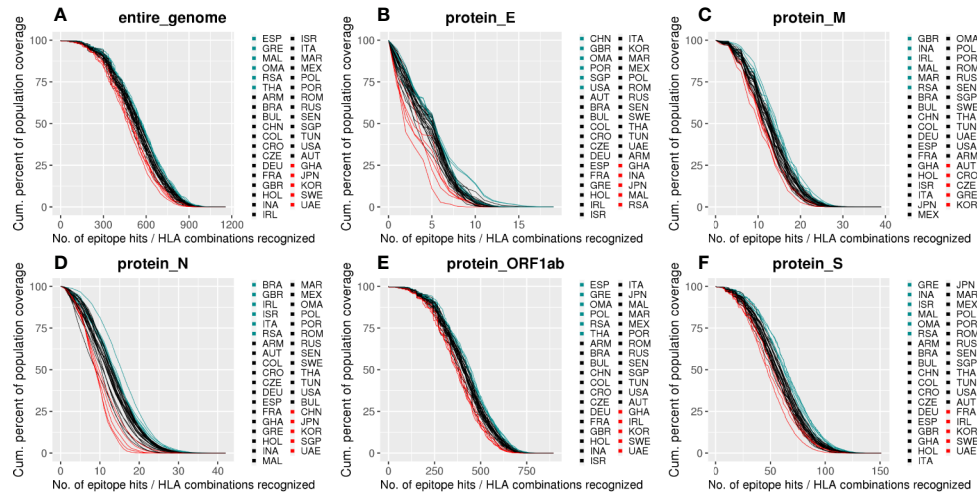
When considering the minimum number of epitopes/HLA combinations predicted to cover 90% of each population, we observed an elevated cumulative coverage throughout the countries, spanning from 466.29 in KOR to 565.58 in OMA, with an average of  $527.61 \pm 24.26$  (**Supplementary File 3**). Despite that, significant differences were spotted when evaluating the Area Under the Curves (AUC) for distinct populations. The AUC reflects the population coverage by accounting the percentage of individuals capable of presenting a set of epitopes considering the epitopes/HLA combinations for a given population. By evaluating the estimated antigen coverage we observed distinct patterns across proteins and divergent

coverage when comparing distinct populations (**Figure 4, Supplementary File 3**). Some populations were better covered for specific portions of the virus, e.g., BRA and ITA showed a high AUC for the protein N (AUC = 1,438.20 and AUC = 1,408.04, respectively, **Figure 4D**) while others showed higher coverages for both N and S proteins, e.g., RSA (AUC = 1,614.47 for N, AUC = 6,248.83 for S) and ISR (AUC = 1,431.51 for N, AUC = 6,416.80 for S) (**Figures 4D, F**). In contrast, IRL showed a high coverage for N protein and a lower for S protein (AUC = 1,428.40 and AUC = 5,392.90, respectively, **Figure 4F**). Importantly, the minimum and maximum AUC values vary according to each protein as they have a different number of Epitope/HLA combinations (**Figure S5**).

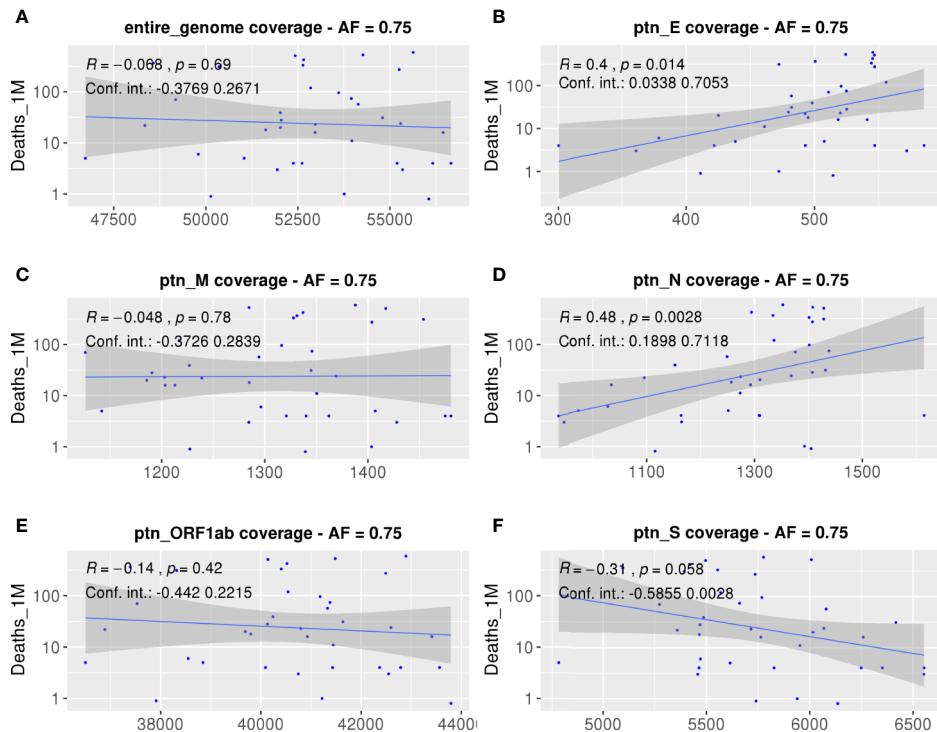
The antigenic coverage within populations can vary since a population is composed of individuals with different ethnic backgrounds. To investigate antigenic coverage differences regardless of geographical localization, but rather associated with distinct ethnic groups in a given population, we performed a population coverage analysis using the HLA alleles described for the four major ethnic backgrounds of the USA: Caucasian, African American, Asian, and Hispanic (**Supplementary File 4**). The estimated coverage varies among ethnic backgrounds depending on the virus protein analyzed (**Figure S6**). African American background has less coverage for E, S, M, and ORF1ab proteins, and high coverage for N protein.

### Epidemiological Parameters Are Associated With Antigen Coverage for Specific ORFs

Aiming to understand the relationship between estimated population antigen coverage and COVID-19 manifestations in different populations, we explored associations between the AUC



**FIGURE 4** | Population coverage for SB peptides derived from SARS-CoV-2 according to the HLA-I set of this study. **(A)** Cumulative population coverage considering the number of possible SARS-CoV-2 derived epitope-HLA allele combinations (%). **(B–F)** Same as in **(A)** but for SB peptides derived from specific viral proteins. Each line represents a population and they were colored according to their AUC: the top six higher AUCs are in light blue, the top six lower are in red, and the remaining curves are black. The abbreviations are alphabetically ordered in the legends, not sorted by AUC. E, Envelope; M, Membrane; N, Nucleocapsid; S, Spike protein.



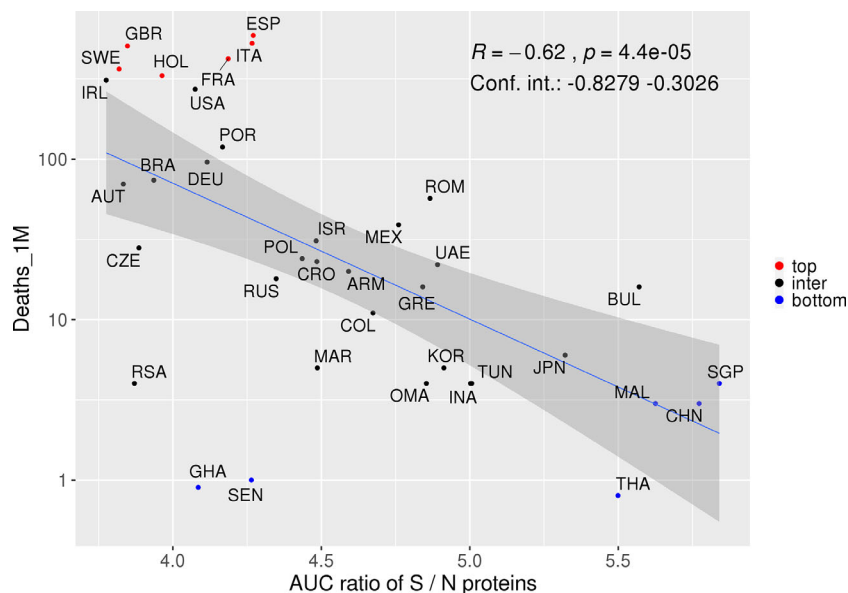
**FIGURE 5** | Correlations between SARS-CoV-2 derived SB peptides coverage of each population and COVID-19 outcome. Spearman correlations between AUC calculated from the population coverage considering the entire SARS-CoV-2 genome **(A)**, Envelope **(B)**, Membrane **(C)**, Nucleocapsid **(D)**, ORF1ab **(E)**, or Spike protein **(F)** derived peptides and deaths per million inhabitants. For this analysis, a threshold of 0.75 for the cumulative AF was considered. The confidence interval shown was calculated using bootstrap with 1000 replacements.

of each country within different ORFs and epidemiological data. As explained before, we have chosen the number of deaths per million inhabitants as the main metric (**Figure 5**), since we believe it may be a more consistent measure, because they are less likely to vary with testing capacity or political strategy. We observed a strong positive correlation between the calculated coverage of each population and the number of deaths per million inhabitants considering potential epitopes derived from the E ( $p = 0.014$ ,  $R = 0.4$ ,  $CI = 0.0423$  to  $0.6995$ ) and N proteins ( $p = 0.0028$ ,  $R = 0.48$ ,  $CI = 0.1518$  to  $0.7036$ ). Consistently, other significant correlations were also found when considering different epidemiological parameters data not shown. The absence of correlation between the epidemiological data and the entire SARS-CoV-2 genome can serve as a random control and reinforces the observed significant correlations for specific viral proteins. Nevertheless, we tested for correlations between scrambled ligands predicted to bind the selected HLAs and epidemiological data. Significant correlations were found for protein N, ORF1ab, and the entire genome ( $R = 0.46$ ,  $p$ -value =  $0.0043$ ,  $R = -0.43$ ,  $p$ -value =  $0.0085$ , and  $R = -0.34$ ,  $p$ -value =  $0.037$ , respectively, data not shown). However, no significant correlation for the S/N ratio was observed (data not shown). Although the number of scrambled ligands has the same magnitude of the one obtained from the original genome, the new peptides supposedly cover less than 10% of population' alleles, which is probably not representative enough.

Interestingly, we observed inverse correlations for S (borderline  $p$ -value), and high AF for HLAs predicted to mainly present S derived peptides in countries from the first quartile. On the other hand, HLAs that potentially present a high

number of N derived peptides were enriched in the fourth quartile. We therefore sought to investigate whether distinct patterns of HLA restriction for S and N-derived peptides could better be associated with deaths. In addition, a given population may have an elevated AUC coverage for both S and N proteins, only for S or only for N proteins. Considering the ratio value, we evaluate the spectrum of S and N derived peptides that simultaneously bind to a collection of HLA alleles, determining if the ratio between the coverage for S/N proteins is more associated with the number of deaths rather than the coverage of each protein separately. In fact, a stronger association with disease outcome was found when evaluating the ratio between the population's AUC predicted for S and N proteins together *versus* the number of deaths per million inhabitants ( $R = -0.62$ ,  $p = 4.4e-5$ ,  $CI = -0.89$  to  $-0.3373$ , Spearman correlation, **Figure 6**). Of note, the AUC ratio for S and N estimated coverage considering ethnic groups was 3.84 for African Americans, 4.08 for Caucasians, 4.55 for Hispanics, and 5.15 for Asians. These AUC ratios fall within the limits of previous analysis (3.78 to 5.84, **Figure 6**) and, although they may indicate a higher number of deaths for people of African American background, the lack of stratified epidemiological data hampered further analysis correlating the ethnic group's coverage with deaths. Therefore, more studies are necessary to confirm these results. Overall, these results highlight the opposite roles associated with potential recognition and presentation of peptides derived from these two SARS-CoV-2 proteins and the COVID-19's outcome.

Interested in better exploring this recognition pattern, we ranked HLA alleles based on the frequency of SB peptides derived from S or N proteins. First, we divided the number of



**FIGURE 6** | Proteins S and N potentially associated with opposite responses associated with COVID-19. Spearman correlation between the S/N AUC ratio and deaths per million inhabitants. Colors represent countries with a higher (red) or lower (blue) number of deaths. The confidence interval shown was calculated using bootstrap with 1,000 replacements.



potential binders individually from each protein by their respective lengths (**Table S2**) to account for the difference in protein sizes, since bigger proteins tend to generate greater numbers of peptides. Next, for each allele, the normalized frequency of SB peptides derived from the S protein was divided by the respective normalized frequency of N-derived peptides. The majority of HLA-A and B alleles have a ratio superior to 1 meaning that they tend to bind to more SB peptides derived from S than to N protein (**Figure S7**). The alleles with the highest ratios are HLA-A\*24:02 (9.54), HLA-A\*23:01 (8.87), and HLA-B\*18:01 (6.25). HLA-A\*30:01 (0.49) and HLA-B\*07:02 (0.32) showed the lowest ratios. HLA-A\*24:02 has an elevated AF in JPN (0.3790) and in many other countries belonging to the first and second quartiles of deaths per million (**Figure 1**). On the other hand, low ratio HLA alleles have a higher AF in GHA (0.1107 for HLA-A30:01) and in SWE (0.186 for HLA-B07:02), the latter possessing a higher number of deaths per million than the average. Moreover, HLA-B\*07:02, which has the lowest ratio for HLA-B alleles, has an elevated AF in several countries of the 4th quartile (**Figure S7, Figure 1**). Importantly, no correlation between the number of deaths per million inhabitants and the number of HLA alleles, either HLA-A, B, or A+B was observed (data not shown). This might suggest a role of specific HLA alleles associated with COVID-19 related deaths rather than the global HLA diversity in a given population.

Until this point, we have focused on representing the majority of each population by using a cumulative AF of 0.75 that is estimated to represent about 93,75% of the individuals in a given population [ $PF = 1 - (1 - AF)^2$ ] (23). Interestingly, when elevating the cumulative HLA AF to 0.9, which is estimated to represent 99% of the individuals, and reperforming the analysis with the same parameters as described for 0.75, a different set of events is observed (**Figure S8**). To reach the 0.9 threshold, 43 HLA-A and 97 HLA-B alleles were included (**Table S1**). For this analysis, two countries (TUN and THA) that haven't reached the cutoff of  $\pm 0.02$  deviation for the 0.90 threshold were excluded. The correlations between AUC population coverage and deaths per million inhabitants were lost except for E protein (**Figure S8B**). Despite the absence of significant correlations between deaths per million inhabitants and AUC coverage for S or N protein, the inverse correlation between AUC ratio for S/N and the number of deaths per million inhabitants ( $R = -0.43$ ,  $p = 0.01$ ,  $CI = -0.7587$  to  $-0.0231$ ) is maintained when considering this new threshold (**Figure S8G**).

To understand the influence of the minor alleles included in the new dataset generated when considering the 0.90 threshold, we performed a new population coverage analysis considering only alleles that were added for this last threshold and are absent from the previous analysis using a threshold of 0.75 (**Figure S9A–G**). Surprisingly, a significant positive Spearman correlation between AUC coverage and deaths per million inhabitants was observed for the S protein ( $R = 0.36$ ,  $p = 0.036$ ,  $CI = 0.0327$  to  $0.5965$ ) and a negative correlation ( $R = -0.36$ ,  $p = 0.033$ ,  $CI = -0.6678$  to  $-0.0105$ ) for the E protein (**Figure S9**). Also in contrast to previous findings, a significant positive correlation ( $R = 0.61$ ,  $p = 0.00011$ ,  $CI = 0.3509$  to  $0.7843$ )

was observed between the AUC ratio for S/N and deaths per million inhabitants (**Figure S9G**) despite no evident correlation with N protein (**Figure S9D**).

## Potential Peptides for Vaccine Development

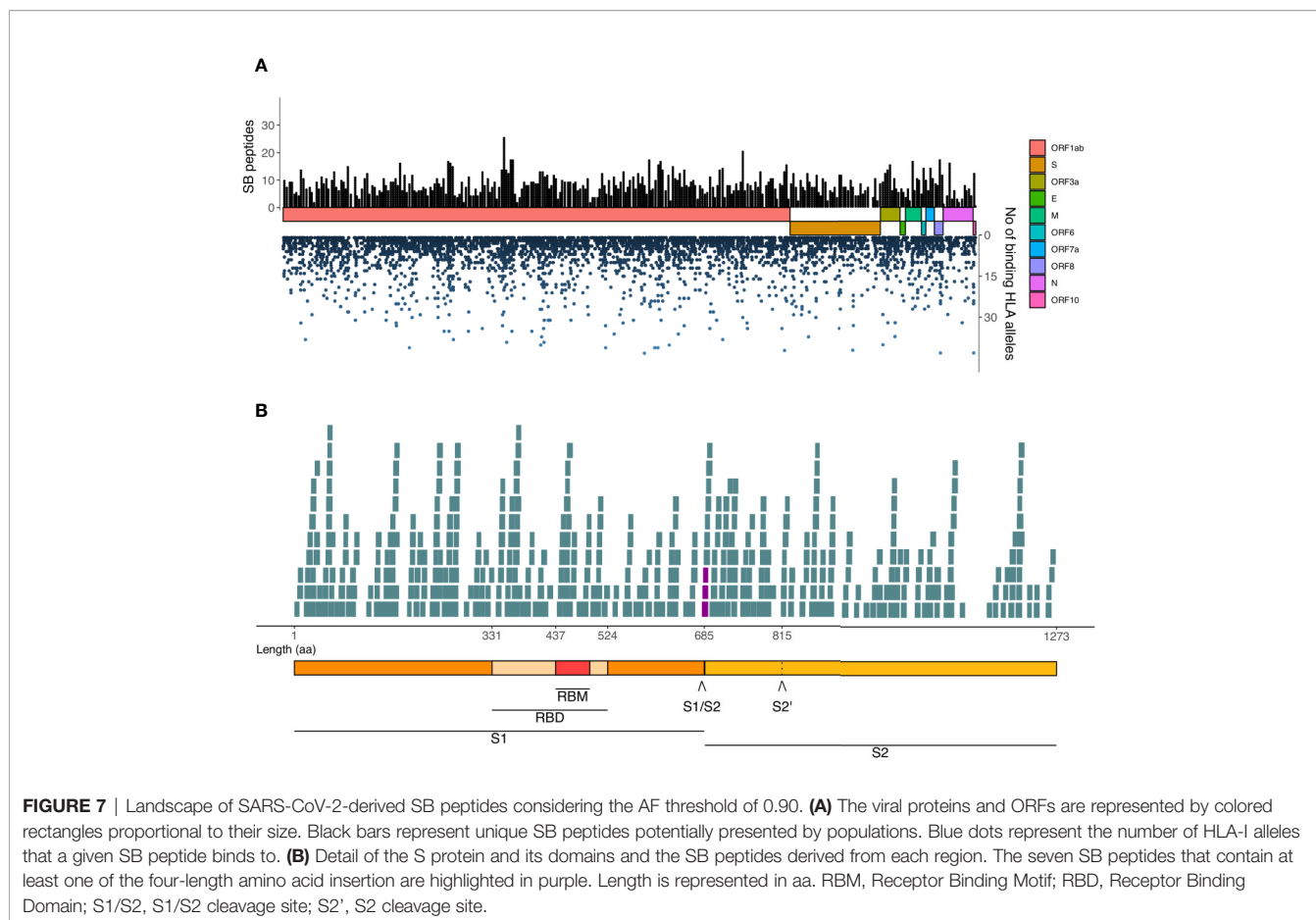
We highlighted the SARS-CoV-2 derived SB peptides generated within each ORF (left Y-axis, **Figure 7A**), as well as the number of HLA-I alleles potentially presenting them (right Y-axis, **Figure 7A**). We observed regions capable of generating an elevated number of unique SB peptides (black bars, top) and the potential of each of these peptides to bind to several HLA-I alleles (blue dots, bottom) which can help to select vaccine candidates. We found 24 unique SB peptides located within the receptor-binding motif and three SB peptides derived from the four amino acid insertion (24) in the S1/S2 cleavage site (highlighted in purple in the **Figure 7B**). These three SB peptides vary in the range of affinities and also the number of HLA binders (**Table 1**). It is also important to note that a 38 amino acid region at the end of S protein (FVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKEYFKN) does not generate any potential HLA-I A or B restricted peptide considering the 140 HLAs analyzed.

Aiming to identify candidate peptides to vaccine development, we listed peptides potentially presented by HLA alleles that possess high allele frequencies in countries belonging to Q1 (**Figure 1**) including but not limited to HLA-A\*24:02, HLA-A\*23:01, HLA-B\*15:02, and HLA-B\*58:01 while avoiding alleles prevalent in Q3 and Q4 (**Figure S10**). It is important to mention that, as stated previously, an unfavorable outcome was observed for countries with high AF for HLA-A\*02:01. Therefore, aiming a vaccine design based on peptides predicted to be presented by this allele may prove ineffective in some populations. A total of 126 candidate peptides were identified considering 85 HLA alleles (**Figure S10, Supplementary File 5**). Six peptides were selected to estimate the global populational coverage considering all deposited populations on IEDB. A coverage of 21.09% is achieved with one peptide, 42.78% with two peptides, and up to 79.99% with six peptides (**Table 2**).

Altogether, these results indicate a distinct pattern of virus recognition and response among populations. HLA allele profiles and predicted antigen coverage for specific ORFs show a strong correlation with mortality by COVID-19. These results highlight the relevance of distinct responses to proteins S and N among different populations to COVID-19 pandemic evolution.

## DISCUSSION

In this work, we investigated the distinct susceptibility associated with HLA diversity profiles among populations facing the SARS-CoV-2. We applied multiple *in silico* approaches to predict the population's response to the SARS-CoV-2 infection and correlated these findings with the number of deaths per million and other epidemiological parameters. We suggest that a positive outcome relies partially on the ability of the majority of a



**TABLE 1** | Strong binder peptides including some of the four amino acid insertion in protein S, which are highlighted in bold.

Peptide	Predicted Aff (nM)	Number of HLA binders	Start (bp)	End (bp)
<b>RA</b> RSVASQSI	129.55 (55.2-203.9)	2	683	692
<b>A</b> RSVASQSI	117.1	1	684	692
<b>RR</b> ARSVASQSI	227.3	1	682	692

population to present peptides derived from the S protein whilst avoiding those derived from the N protein, as it correlates with an unfavorable outcome.

Numerous factors such as political rules implementation, adherence to social distancing measures, health system structures, and population density can affect the spread of SARS-CoV-2 and impact COVID-19 outcome. Besides, underreporting and testing policies vary among countries and also influence the epidemiological reports (i.e.: INA, MEX, SEN, JPN, MAR TUN, BRA, COL, and THA performed tests in less than 1% of the population). Absolute epidemiological parameters such as total number of cases, serious/critical cases, deaths, or recoveries are biased by the number of cases or population size. In this sense, when using the number of deaths per number of cases, the countries' testing policy can directly interfere with the data because countries with an open testing policy tend to have fewer serious cases and deaths than countries which test only

hospitalized or risk group individuals. Therefore, we choose to use deaths per million inhabitants as the main metric since it also has been used as a proxy for disease outcome (25, 26). Other metrics such as excess deaths might be a good representation of deaths related to COVID19 and reducing the testing bias (27), but this information was not available for most of the countries by the time we retrieved data for this work. In addition, other factors like blood type and ethnicity (28, 29), countries BCG vaccination policy (25), or the percentage of obese people (30) could influence mortality, affecting the relationship between deaths per million and Spike or Nucleocapsid proteins AUC coverage. In fact, some countries (e.g.: ITA, ESP, FRA, GBR, USA, and HOL) showed higher deaths per million than expected by the relation between AUC ratio of S/N and deaths per million inhabitants. Considering that, we included as many countries as possible in the analysis to minimize these discrepancies rather than normalize for the particularities of each population.

**TABLE 2** | Global population coverage analysis based on the allele frequencies available at the Immune Epitope Database.

Peptides	Projected population coverage	Average number of epitope hits/HLA combinations recognized by the population	Minimum number of epitope hits/HLA combinations recognized by 90% of the population
TRFQTLAL	21.09%	0.22	0.13
TRFQTLAL	42.78%	0.51	0.17
YFPLQSYGF			
TRFQTLAL	54.12%	0.68	0.22
YFPLQSYGF			
SVLNDILSR			
TRFQTLAL	64.66%	0.9	0.28
YFPLQSYGF			
SVLNDILSR			
ATRFASVYAW			
TRFQTLAL	75.51%	1.18	0.41
YFPLQSYGF			
SVLNDILSR			
ATRFASVYAW			
KEIDRLNEV			
TRFQTLAL	79.99%	1.32	0.5
YFPLQSYGF			
SVLNDILSR			
ATRFASVYAW			
KEIDRLNEV			
ESNKKFLPF			

Our analyses depict the potential population's patterns of SARS-CoV-2 recognition of class I HLA restricted peptides. The HLA alleles repository allowed us to carefully select populations prioritizing blood donor registry rather than anthropology studies to better represent a country. Since the most relevant genes for CTL responses are the HLA-A and B and considering that HLA-C data is not available for several populations, we have not accounted for HLA-C (31). It also allowed us to characterize in detail the potential responses associated with major and minor alleles. Remarkably, HLA-B\*15:03, the allele predicted to present more peptides in previous works (12, 13), ranks 12th out of 76 HLA-B alleles analyzed. As the methodology and metrics used in these previous works differ from ours, different results are expected.

HLA systems have a strong linkage disequilibrium (LD) that can result in conserved multi-locus haplotypes. This happens because the alleles are arranged in adjacent loci in a specific region of the chromosome, where the distance between them, among other factors, can affect the LD (32), and, consequently, their random inheritance in a given population. To correct for LD, we would need to have access to full haplotypes frequencies instead of AF, which is not available or incomplete for many countries impacting the size of the study. Despite that, although our study does not consider the LD of the HLA system, we believe that using high AF thresholds, as we did, minimize the need for the correction for LD. Furthermore, the impact of LD on the calculation of population coverage by IEDB has been considered negligible (14).

HLA binding predictions have become a useful and important method for selecting promising peptides for vaccination. Moreover, by comparing the predicted binding affinities with the affinities verified through assays deposited at the IEDB, we showed that SB peptides, the focus of this study, have comparable

affinities to the peptides experimentally defined as high and intermediate binders. In addition, their binding affinities are in the range of those peptides known to trigger a TCR response as detected by T cell-based assays as opposed to WB peptides.

Previous works have highlighted the importance of population coverage studies to uncover candidate peptides for vaccination. We identified an inverse correlation between S and N coverage and deaths per million inhabitants. This result corroborates a study showing that CD8 T lymphocytes from convalescent patients display more reactive CD8+ T cells for S than for the N protein (10), highlighting its immunogenic role. This same report indicates that several immunogenic HLA-I restricted peptides are derived from S protein, in line with our findings that S peptide coverage has a positive impact on COVID-19 outcome. These results suggest CD8 mediated anti S-derived peptide responses could be one of the mechanisms associated with more favorable COVID-19 outcomes. The observed correlation of the scrambled ligands from protein N and ORF1ab might reflect small biases of the alleles and corresponding predictions, rather than the virus' peptides specifically, corroborating the HLA AF-based clustering profile of countries belonging to the first quartile (with fewer deaths per million). Finally, as the S/N ratio correlation wasn't reproducible when shuffling the ligands, it also suggest a role of a specific HLAs that differentially restrict S- and N-derived peptides to effectively overcome the COVID-19, reinforcing the need of detecting genetic risk groups.

Features of SARS-CoV-2 S proteins are responsible for unique characteristics of this new virus. A recent report indicated differences in the protein S sequence regarding the receptor-binding motif and a four amino acid insertion when comparing with bat and previous strains (24). We found unique SB peptides falling within the receptor-binding motif sequence, including the

four amino acid insertion in the S1/S2 cleavage site. This sequence is considered to be related to the high virulence of SARS-CoV-2 (24) and HLA-I restricted peptides derived from this region could help mounting differential immune responses between SARS-CoV-2 and other coronaviruses. On the other hand, the lack of potential peptides derived from the FVSGNCDVVIGIVNNTVYDPLQ PELDSFKEELDKYFKN region is somehow intriguing and exploiting functional features of this segment and evolutive forces that could have shaped its low immunogenicity can shed light over the relevance of these findings.

The scarceness of N protein derived immunogenic HLA-I restricted peptides reported by Grifoni and colleagues (3) resembles the low proportion of N protein-derived peptides found in our analysis. This reinforces the concept that N derived peptides may not be involved in functional and productive anti-SARS-CoV-2 responses since our correlation analyses indicate that populations with greater N coverage have higher numbers of deaths per million inhabitants. Protein N is highly expressed and has a high immunogenic potential so that antibodies against it are the first found in the serum of infected patients (33). Since the homology between N proteins of SARS-CoV-2 and SARS-CoV is over 90% (15) and the SARS-CoV-2 peak viral load is earlier (34), the immunogenicity of SARS-CoV-2 protein N is believed to be similar to that of SARS-CoV with an even faster response. However, when investigating T lymphocytes from convalescent patients, a lower frequency of targets associated with protein N was observed (10). This aspect indicates that these peptides are either less immunogenic or could lead to the depletion of reacting cells by some T cell hyperactivation mechanism. If this is the case, N derived peptide-based hyperactivation could be related to cytokine storm, a process largely associated with the most severe cases of COVID-19 (4, 35, 36). The number of deaths per million inhabitants showed a significant positive correlation when correlated with N coverage, reinforcing the possibility that N peptides lead to a more aggressive disease. This dysregulation of the immune system has been shown to occur in other viral infections and could be dependent on CD8 T lymphocytes (37).

Assuming that, like rhesus monkeys, individuals previously infected with SARS-CoV-2 are protected from reinfection, herd immunity becomes possible for COVID-19 (38). Herd immunity consists of individuals being protected from infection by being surrounded by immune non-infected individuals, therefore unable to transmit the disease (39). Previous studies have shown that the average number of SARS-CoV-2 transmission caused by a single infectious individual ( $R_0$ ) is from 2 to 6 (40, 41), so assuming a  $R_0$  of 4, the herd immunity would be achieved with 75% immune individuals. By using a threshold of 0.75 for cumulative AF for HLA-A and B within each population, we estimate that 93.75% of the country's population is covered by our analysis. In this context, vaccine strategies based on S derived peptides would induce protective responses to the infection in the majority of the population that carry more frequent HLAs. The opposite would be expected regarding the minority, carrying rare HLAs, that display a positive correlation between S coverage and the number of deaths. On the other hand, the same rationale

may apply to N and E proteins which correlate positively with the number of deaths for the major HLA alleles but display an inverse tendency when only the minor alleles are accounted for. In summary, an elevated S/N coverage for the most frequent HLA alleles showed protective effects whilst lower S/N coverage ratios were protective for the minor HLA alleles within populations. It is important to mention, though, that reaching herd immunity through vaccination for COVID-19 can be a hard task, probably requiring revaccination, since the immune response to SARS-CoV-2 may be lost rapidly (42) and people might thereafter be susceptible to reinfection with the virus.

A country's population is normally composed of multi-ethnic groups and backgrounds, which could undergo distinct susceptibilities to SARS-CoV-2 infection. In addition, there is an increasing debate on whether some ethnic groups are being disproportionately affected by COVID-19, which is frequently enhanced by distinct access to the health system and social inequities (29, 43, 44). Our attempt to investigate antigenic coverage within ethnic groups of the same country aimed to check differences in their coverage patterns, regardless of geographical localization. Overall, our results suggest that differences within ethnic backgrounds may exist and could be associated with a better or worse susceptibility for these groups. However, more studies are necessary to confirm this pattern, and also to explore these discrepancies in other countries since the ethnic background should differ from country to country.

The results also highlight the need for distinct political strategies to deal with the pandemic, which differs among countries. We were able to identify two HLA alleles that are highly prevalent in countries showing less number of deaths and may have protective effects regarding the immune response to SARS-CoV-2. The vast majority of peptides derived from protein S potentially presented by those two alleles were absent from supertype HLAs such as HLA-A\*01:01, A\*02:01, A\*03:01, B\*15:01, and others. In fact, the most prominent worldwide allele HLA-A\*02:01 potentially presents peptides from the ORF1ab region. A recent study has shown that this region is not immunodominant in CD8-mediated immune response when challenged with SARS-CoV-2 Epitope MegaPool in convalescent donors (10). HLA-A\*24:02 is highly frequent in many populations and the predictions made for this allele could help drive vaccine development.

Finally, we highlighted some important viral regions predicted to be presented by a vast number of HLA alleles which can help to understand the response to viral peptides and peptide vaccine design. More importantly, we suggest that peptide-based vaccination strategies should rely mainly on the S protein due to its negative correlation with the number of deaths per million inhabitants and its role in the infection. On the other hand, we present a cautionary note for using the N protein since it showed a strong positive correlation with the number of deaths per million when considering individuals carrying frequent alleles. Last, we identified potential antigens derived from the four amino acid insertion of SARS-CoV-2 that is absent from previous strains. This protein region could be used for studies based on T cell response, as it would allow differentiation from previous



coronavirus infections and importantly, serve as a guide for SARS-CoV-2 specific vaccine development. To our knowledge, this is the first work to focus on antigen population coverage associating with COVID-19 outcome.

## DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**. Further inquiries can be directed to the corresponding author.

## AUTHOR CONTRIBUTIONS

MB, MP, and RG contributed to the conception of the study. GV, AB, MHB, and MB contributed with the design of the study and important scientific discussions. MP and RG performed all the

analysis and wrote the first draft of the manuscript. MB was responsible for the final approval of the submitted version. All the authors contributed to the article and approved the submitted version.

## ACKNOWLEDGMENTS

MP was supported by CAPES fellowship program. The authors also would like to thank the Bioinformatics Core Facility (INCA-RJ) for their support.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fimmu.2020.565730/full#supplementary-material>

## REFERENCES

- WHO. *COVID-19 situation reports* (2020). Available at: <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports> (Accessed May 20, 2020).
- Wu F, Zhao S, Yu B, Chen Y-M, Wang W, Song Z-G, et al. A new coronavirus associated with human respiratory disease in China. *Nature* (2020) 579:265–9. doi: 10.1038/s41586-020-2008-3
- Grifoni A, Sidney J, Zhang Y, Scheuermann RH, Peters B, Sette A. A Sequence Homology and Bioinformatic Approach Can Predict Candidate Targets for Immune Responses to SARS-CoV-2. *Cell Host Microbe* (2020) 27:671–80.e2. doi: 10.1016/j.chom.2020.03.002
- Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* (2020) 395:497–506. doi: 10.1016/S0140-6736(20)30183-5
- Wang D, Hu B, Hu C, Zhu F, Liu X, Zhang J, et al. Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China. *JAMA* (2020) 323:1061–9. doi: 10.1001/jama.2020.1585
- Hamming I, Timens W, Bulthuis MLC, Lely AT, Navis GJ, van Goor H. Tissue distribution of ACE2 protein, the functional receptor for SARS coronavirus. A first step in understanding SARS pathogenesis. *J Pathol* (2004) 203:631–7. doi: 10.1002/path.1570
- Liao M, Liu Y, Yuan J, Wen Y, Xu G, Zhao J, et al. Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. *Nat Med* (2020) 26:842–4. doi: 10.1038/s41591-020-0901-9
- Tay MZ, Poh CM, Rénia L, MacAry PA, Ng LFP. The trinity of COVID-19: immunity, inflammation and intervention. *Nat Rev Immunol* (2020) 20:363–74. doi: 10.1038/s41577-020-0311-8
- Zhang J, Zeng H, Gu J, Li H, Zheng L, Zou Q. Progress and Prospects on Vaccine Development against SARS-CoV-2. *Vaccines (Basel)* (2020) 8:153. doi: 10.3390/vaccines8020153
- Grifoni A, Weiskopf D, Ramirez SI, Mateus J, Dan JM, Moderbacher CR, et al. Targets of T Cell Responses to SARS-CoV-2 Coronavirus in Humans with COVID-19 Disease and Unexposed Individuals. *Cell* (2020) 181:1489–501.e15. doi: 10.1016/j.cell.2020.05.015
- Shi Y, Wang Y, Shao C, Huang J, Gan J, Huang X, et al. COVID-19 infection: the perspectives on immune responses. *Cell Death Differ* (2020) 27:1451–4. doi: 10.1038/s41418-020-0530-3
- Nguyen A, David JK, Maden SK, Wood MA, Weeder BR, Nellore A, et al. Human leukocyte antigen susceptibility map for severe acute respiratory syndrome coronavirus 2. *J Virol* (2020) 94:e00510–20. doi: 10.1128/JVI.00510-20
- Barquera R, Collen E, Di D, Buhler S, Teixeira J, Llamas B, et al. Binding affinities of 438 HLA proteins to complete proteomes of seven pandemic viruses and distributions of strongest and weakest HLA peptide binders in populations worldwide. *HLA* (2020) 96:277–98. doi: 10.1111/tan.13956
- Bui H-H, Sidney J, Dinh K, Southwood S, Newman MJ, Sette A. Predicting population coverage of T-cell epitope-based diagnostics and vaccines. *BMC Bioinf* (2006) 7:153. doi: 10.1186/1471-2105-7-153
- Ahmed SF, Quadeer AA, McKay MR. Preliminary Identification of Potential Vaccine Targets for the COVID-19 Coronavirus (SARS-CoV-2) Based on SARS-CoV Immunological Studies. *Viruses* (2020) 12:254. doi: 10.3390/v12030254
- González-Galarza FF, Takeshita LYC, Santos EJM, Kempson F, Maia MHT, da Silva ALS, et al. Allele frequency net 2015 update: new features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations. *Nucleic Acids Res* (2015) 43:D784–8. doi: 10.1093/nar/gku1166
- Jurtz V, Paul S, Andreatta M, Marcotili P, Peters B, Nielsen M. NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *J Immunol* (2017) 199:3360–8. doi: 10.4049/jimmunol.1700893
- Worldometer. *Coronavirus Update (Live): Cases and Deaths from COVID-19 Virus Pandemic - Worldometer* (2020). Available at: <https://www.worldometers.info/coronavirus/> (Accessed May 20, 2020).
- Gonzalez-Galarza FF, Christmas S, Middleton D, Jones AR. Allele frequency net: a database and online repository for immune gene frequencies in worldwide populations. *Nucleic Acids Res* (2011) 39:D913–9. doi: 10.1093/nar/gkq1128
- Hilborne LH, Wagner Z, Cabrerros I, Brook RH. Linking Statistics With Testing Policy to Manage COVID-19 in the Community. *Am J Clin Pathol* (2020) 154:142–8. doi: 10.1093/ajcp/aqaa099
- Stranzl T, Larsen MV, Lundegaard C, Nielsen M. NetCTLpan: pan-specific MHC class I pathway epitope predictions. *Immunogenetics* (2010) 62:357–68. doi: 10.1007/s00251-010-0441-4
- Kaufman J. Generalists and specialists: A new view of how MHC class I molecules fight infectious pathogens. *Trends Immunol* (2018) 39:367–79. doi: 10.1016/j.it.2018.01.001
- Gonzalez-Galarza FF, McCabe A, Santos EJM, Jones J, Takeshita L, Ortega-Rivera ND, et al. Allele frequency net database (AFND) 2020 update: gold-standard data classification, open access genotype data and new query tools. *Nucleic Acids Res* (2020) 48:D783–8. doi: 10.1093/nar/gkz1029
- Jaimes JA, André NM, Chappie JS, Millet JK, Whittaker GR. Phylogenetic Analysis and Structural Modeling of SARS-CoV-2 Spike Protein Reveals an Evolutionary Distinct and Proteolytically Sensitive Activation Loop. *J Mol Biol* (2020) 432:3309–25. doi: 10.1016/j.jmb.2020.04.009
- Miller A, Reandelar MJ, Fasciglione K, Roumenova V, Li Y, Otazu GH. Correlation between universal BCG vaccination policy and reduced morbidity

- and mortality for COVID-19: an epidemiological study. *medRxiv* (2020). doi: 10.1101/2020.03.24.20042937
26. Hallal PC. *Worldwide differences in COVID19 - related mortality. Worldwide differences in COVID19 - related mortality - Artigos - Revista Ciência & Saúde Coletiva* (2020). Available at: <http://www.cienciaesaudecoletiva.com.br/artigos/worldwide-differences-in-covid19-related-mortality/17555?id=17555&id=17555> (Accessed May 26, 2020).
  27. The Economist. *Covid-19 data - Tracking covid-19 excess deaths across countries. Covid-19 data - Tracking covid-19 excess deaths across countries | Graphic detail | The Economist* (2020). Available at: <https://www.economist.com/graphic-detail/2020/07/15/tracking-covid-19-excess-deaths-across-countries> (Accessed November 5, 2020).
  28. Zhao J, Yang Y, Huang H, Li D, Gu D, Lu X, et al. Relationship between the ABO Blood Group and the COVID-19 Susceptibility. *Clin Infect Dis* (2020). doi: 10.1093/cid/ciaa1150
  29. Williamson EJ, Walker AJ, Bhaskaran K, Bacon S, Bates C, Morton CE, et al. Factors associated with COVID-19-related death using OpenSAFELY. *Nature* (2020) 584:430–6. doi: 10.1038/s41586-020-2521-4
  30. Kass DA, Duggal P, Cingolani O. Obesity could shift severe COVID-19 disease to younger ages. *Lancet* (2020) 395:1544–5. doi: 10.1016/S0140-6736(20)31024-2
  31. Blais M-E, Dong T, Rowland-Jones S. HLA-C as a mediator of natural killer and T-cell activation: spectator or key player? *Immunology* (2011) 133:1–7. doi: 10.1111/j.1365-2567.2011.03422.x
  32. Slatkin M. Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. *Nat Rev Genet* (2008) 9:477–85. doi: 10.1038/nrg2361
  33. Tan Y-J, Goh P-Y, Fielding BC, Shen S, Chou C-F, Fu J-L, et al. Profiles of antibody responses against severe acute respiratory syndrome coronavirus recombinant proteins and their potential use as diagnostic markers. *Clin Diagn Lab Immunol* (2004) 11:362–71. doi: 10.1128/cdli.11.2.362-371.2004
  34. Zou L, Ruan F, Huang M, Liang L, Huang H, Hong Z, et al. SARS-CoV-2 Viral Load in Upper Respiratory Specimens of Infected Patients. *N Engl J Med* (2020) 382:1177–9. doi: 10.1056/NEJMc2001737
  35. Zhou F, Yu T, Du R, Fan G, Liu Y, Liu Z, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *Lancet* (2020) 395:1054–62. doi: 10.1016/S0140-6736(20)30566-3
  36. Qin C, Zhou L, Hu Z, Zhang S, Yang S, Tao Y, et al. Dysregulation of Immune Response in Patients With Coronavirus 2019 (COVID-19) in Wuhan, China. *Clin Infect Dis* (2020) 71:762–8. doi: 10.1093/cid/ciaa248
  37. Walsh KB, Teijaro JR, Brock LG, Fremgen DM, Collins PL, Rosen H, et al. Animal model of respiratory syncytial virus: CD8+ T cells cause a cytokine storm that is chemically tractable by sphingosine-1-phosphate 1 receptor agonist therapy. *J Virol* (2014) 88:6281–93. doi: 10.1128/JVI.00464-14
  38. Deng W, Bao L, Liu J, Xiao C, Liu J, Xue J, et al. Primary exposure to SARS-CoV-2 protects against reinfection in rhesus macaques. *Science* (2020) 369:818–23. doi: 10.1126/science.abc5343
  39. Randolph HE, Barreiro LB. Herd Immunity: Understanding COVID-19. *Immunity* (2020) 52:737–41. doi: 10.1016/j.immuni.2020.04.012
  40. Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. *N Engl J Med* (2020) 382:1199–207. doi: 10.1056/NEJMoa2001316
  41. Sanche S, Lin YT, Xu C, Romero-Severson E, Hengartner N, Ke R. High contagiousness and rapid spread of severe acute respiratory syndrome coronavirus 2. *Emerg Infect Dis* (2020) 26:1470–7. doi: 10.3201/eid2607.200282
  42. Seow J, Graham C, Merrick B, Acors S, Pickering S, Steel KJA, et al. Longitudinal observation and decline of neutralizing antibody responses in the three months following SARS-CoV-2 infection in humans. *Nat Microbiol* (2020) 5:1598–607. doi: 10.1038/s41564-020-00813-8
  43. Office for National Statistics. *Coronavirus-related deaths by ethnic group, England and Wales methodology - Office for National Statistics* (2020). Available at: <https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/methodologies/coronavirusrelateddeathsbyethnicgroupenglandanddwalesmethodology> (Accessed August 12, 2020).
  44. Price-Haywood EG, Burton J, Fort D, Seoane L. Hospitalization and Mortality among Black Patients and White Patients with Covid-19. *N Engl J Med* (2020) 382:2534–43. doi: 10.1056/NEJMsa2011686

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Pretti, Galvani, Vieira, Bonomo, Bonamino and Boroni. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.