Class I HLA allele restricted antigenic coverage for Spike and N proteins is associated with divergent outcomes for COVID-19

Marco Antônio M. Pretti[1,2], Rômulo G. Galvani[1,3,4,], Gustavo Fioravanti Vieira[5,6], Adriana Bonomo[4,7], Martín H. Bonamino[2,7], Mariana Boroni[1*]

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

*** Correspondence:**
Corresponding Author
mariana.boroni@inca.gov.br

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

Abstract

The world is dealing with 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 investigate the landscape of potential HLA-I A and B restricted SARS-CoV-2-derived antigens and how different populations in the world are predicted to respond to those peptides considering their HLA-I distribution frequencies. 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 strong binder peptides and epidemiological data. The number of deaths per million inhabitants inversely correlated with the antigen coverage of peptides derived from viral protein S, while a direct correlation was observed for those derived from viral protein N, highlighting a potential risk group carrying HLAs associated with the latter. In addition, we identified 7 potential antigens bearing at least one amino acid of the small 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.

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 5.2 million cases and 337 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) (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. 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 response to SARS-CoV-2 viral proteins has 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 different populations (11) and as a result, distinct profiles of disease susceptibility and severity are observed (1). 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 in order to promote a functional response (12). Populational coverage studies considering these features may help narrow the number of candidate peptides to be considered for vaccination strategies (13). However, epitope coverage studies are not sufficient to predict if the immune response would be able to resolve the infection and immunize individuals and/or if this immunization occurs at the expense of exacerbated immune responses. In the context of the SARS-CoV-2 pandemics, 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 once 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 outcome: while S coverage has an inverse relationship with the number of deaths, N has a direct correlation with the unfavorable endpoint. These results may prove important in guiding efforts towards the development of vaccine candidates.

Results

HLA-I alleles partially discriminate different outcomes among countries

This is a provisional file, not the final typeset article

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, only countries with more than 1,000 total cases of Coronavirus disease (COVID-19) on 17/05/2020 registered in the Worldometer database (14) were considered, reaching a total of 50 different countries. Four Digits resolution HLA-A and HLA-B alleles for the aforementioned populations available in the Allele Frequency Net Database (15) were accessed. For each population, a different number of alleles was retrieved, considering always the most frequent alleles, to reach a minimum cumulative allelic frequency (AF) of 0.75. Populations whose HLA allele frequencies information did not reach this minimum cutoff were removed, totalizing 38 populations of the initial 50 in our study (Supplementary File 1, Figure S1). The median cumulative AF varied among populations with a median of 0.7705 for HLA-A (minimum of 0.752 for DEU and maximum of 0.822 for AUT) and 0.7622 for HLA-B (minimum of 0.75 for MEX and maximum of 0.798 for NZL) and the number of HLA-I alleles ranged from 4 to 14 for HLA-A (median=7) and from 7 to 27 for HLA-B (median=13.5).

To investigate whether different HLA-I alleles among populations could be associated with epidemiologic features (retrieved from the Worldometer database, see methods), we sorted countries in ascending order regarding these data (Supplementary File 1). The USA was the country with the highest number of cases, followed by RUS, ESP, GBR, BRA, and ITA while the countries with the lowest number of cases were SEN, CZE, TUN, KOR, NZL, and BUL (Figure S2A). However, when we evaluate cases per million inhabitants, IRL was the leading country in the rank, followed by SGP, USA, ITA, SWE, and POR while BRA, THA, CHN, INA, TUN, and JPN were the countries with the lowest ratios (Figure S2B). The USA also led the rank in number of COVID-19 deaths, followed by GBR, ITA, FRA, BRA, and DEU while the countries with the lowest number of deaths were NZL, OMA, SGP, SEN, GHA, and TUN (Figure S2C). When analyzing the number of deaths per million inhabitants, ESP was the country with the highest ratio followed by ITA, GBR, FRA, SWE, and HOL while THA, GHA, SEN, CHN, MAL, and INA were the countries with the lowest ratios (Figure S2D). The USA also had the highest number of serious cases followed by BRA, FRA, GBR, DEU, and ESP while the countries with the lowest number of serious cases were MAR, UAE, TUN, GHA, SEN, and ARM (Figure S2E). On the other hand, the USA also led in the number of recovered, followed by ESP, ITA, BRA, CHN, and RUS while BUL, TUN, SGP, SEN, ROM, and GRE were the countries with the lowest numbers of recovered (Figure S2F).

To evaluate if the distribution of HLA allele frequencies among populations could explain the epidemiologic data, we performed an unsupervised hierarchical clustering using HLA-A and HLA-B frequencies for each distinct population, that were divided into quartiles according to each epidemiological parameter. Clustering of HLA allele frequencies among populations explained better deaths per million inhabitants. Therefore, we chose to use deaths per million inhabitants as the main metric since it also has been used as a proxy for disease outcome (16,17). We observed two branches enclosing 8 of 11 countries that belong to the first quartile (Q1, green, less number of deaths per million), while countries with the worst outcome (Q4, red, more number of deaths per million) tended to cluster together (Figure 1). 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 are HLA-A*02:01, HLA-A*01:01, HLA-A*03:01, and HLA-B*07:02. Despite the clustering of some populations belonging to different quartiles, this result highlights the contribution of HLA alleles to segregate distinct responses to the COVID-19 infection among populations. We reasoned that describing the potential of each allele to present peptides related to SARS-CoV-2 could further contribute to the understanding of the epidemiologic data, so SARS-CoV-2 derived peptide binding prediction to HLA-I alleles was performed.

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 78 HLA-B, Supplementary File 2) for each potential SARS-CoV-2 derived peptide using the netMHCpan tool (18). The 58,470 predicted 8-11 mer peptides were derived from the complete genome of the SARS-CoV-2 deposited on the NCBI (BetaCoV/Wuhan/IPBCAMS-WH-01/2019, see methods). Peptides were categorized as Non-Binders (NB, %Rank ≥ 2; 38,413 peptides), Strong Binders (SB, %Rank < 0.5; 5,785 peptides) and Weak Binders (WB, 0.5 ≤ %Rank < 2; 12,731 peptides) based on the %Rank definition by the developers, which accounts for the probability of a peptide do bind the HLA given a pool of natural ligands. A crescent value in the predicted affinity is observed with the decrease in the qualitative classifications according to the program, lower for SB (median of 180.6 nM), moderated for WB (median of 1,575.9 nM) and higher for NB (median of 36,442.2 nM, Figure S3A). In order to define which group of peptides we should work with, we compared the predicted *in silico* affinities with affinities of viral-derived peptides obtained by experimental assays. The aforementioned data was retrieved from the Immune Epitope Database (IEDB). A total of 49,207 MHC Ligand assays were retrieved following the selection criteria, 38,558 for HLA-A and 10,649 for HLA-B classified by the IEDB according to the results of *in vitro* assays readout into Positive-High (Pos-H, 17,353), Positive-Intermediate (Pos-I, 14,927) or Positive-Low (Pos-L, 16,927). The IEDB classifications take into account the information provided in deposited papers, those categorized into Pos-L/-I/-H have been grouped accordingly with an IC50 threshold. A similar pattern was observed for the experimental affinities, which were lower for Pos-H, median of 10.1 nM, and higher in Pos-L peptides, median of 1,430 (Wilcox, $p < 0.0001$, Figure S3B). Moreover, we also predicted the binding affinities of 18,522 unique peptide/HLA pairs retrieved from the IEDB whose capacity to elicit a TCR response was tested through 34,872 T cell assays. They are qualitatively classified into Positive (Pos, 2,705) or Negative (Neg, 15,817) taking into account *in vitro* assays such as ELISPOT and chromium release. A significant difference in the affinities was observed when comparing Pos (median of 49.1 nM) assays with Neg (median of 188.0 nM) ones (Wilcox, $p < 0.0001$, Figure S3C). It is important to mention that a difference in affinity (Wilcoxon, $p < 0.0001$) was observed for the predicted data (Figure S3D) but no overall difference in binding affinities was observed for IEDB HLA-A and B alleles (Figure S3E). Fifty-seven percent 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. In summary, the predicted affinities of peptides classified into strong, weak or no binder are comparable to the experimental affinities from ligand assays at the IEDB showing that SB peptides have affinities equivalent to the vast majority of peptides proven to bind the HLA experimentally (Pos-H and Pos-I - Figures S3A-B), probably eliciting the immune response. Therefore, we decided to work only with the SB peptides in the downstream analyses.

Of the 111 analyzed alleles, a total of 5,785 unique SB peptides were identified after the removal of 9 SB peptides that matched with 100% identity the sequences found in human proteome using BLAST search. The number of SB peptides potentially presented by each HLA allele is depicted in Figure 2. The alleles presenting the higher number of SB peptides are HLA-A*11:01 (11,385 peptides) and HLA-B*35:01 (11,174 peptides).

We confirmed that the number of SB peptides increases accordingly the number of HLA alleles for a given population by performing a Pearson correlation between these two variables ($R=0.96$, p-value < 0.01, Figure S4). Even considering that the 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.

This is a provisional file, not the final typeset article

Distinct patterns of antigen coverage among populations are observed within different ORFs

Searching for specific virus proteins and regions within the virus ligandome, we analyzed the set of potential HLA-restricted peptides per population and open reading frames (ORF). The number of SB peptides presented by each allele varies greatly depending on the virus ORF considered, and it is proportional to ORF sizes (Figures S5-6, Table S1). For example, HLA-A*24:02 presents the highest number of SB peptides derived from protein S (1,274 amino acids - aa) and a moderate number of SB peptides derived from the protein N (420 aa, Figure S5A).

The use of 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 S2, Supplementary File 3).

We observed an elevated cumulative coverage throughout the populations, spanning from 433.51 in HOL to 511.4 in MAL, with an average of 465.13±17.88, when considering the minimum number of epitopes/HLA combinations that cover 90% of the population. 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 an epitope and the number of epitopes/HLA combinations for a given population. By evaluating the antigen coverage we observed distinct patterns across proteins and divergent coverage when comparing distinct populations within proteins (Figure 3, Supplementary File 3). Some populations were better covered for specific portions of the virus, e.g. BRA (AUC=1,951.36), GBR (AUC=1,824.80), GHA (AUC=1,939.29), IRL (1,949.74), RSA (AUC=2,130.38), and SWE (AUC=1,915.50) show a high AUC for the protein N (Figure 3D). In contrast, three of these populations, e.g. GBR (AUC=7,716.48), IRL (AUC=7,949.25), and SWE (AUC=7,639.13) showed low S coverage (Figure 3F). Importantly, the minimum and maximum AUC values differ according to each protein as they have a different number of Epitope/HLA combinations (Figure S7).

Epidemiological parameters are associated with antigen coverage for specific ORFs

Aiming to understand the relationship between population antigen coverage and SARS-CoV-2 progression in different populations, we explored associations between the AUC of each population within different ORFs and epidemiological data. We performed Pearson correlations between the epidemiological parameters associated with the COVID-19 epidemics and the AUC of the 5 proteins analyzed and the entire SARS-CoV-2 genome. We have tested different epidemiological outcomes (Figures S8-11) and the number of deaths per million inhabitants was the metric most associated with population coverage calculated using class I SARS-CoV-2 derived SB peptides (Figure 4). We observed a strong negative correlation between the calculated coverage of each population (represented by the AUCs) and the number of deaths per million inhabitants considering a set with all potential epitopes derived from the coronavirus genome (p=0.0017, R=-0.49) as well as for ORF1ab (p=0.0031, R=-0.47), S (p=0.0017, R=-0.49) and M (p=0.0025, R=-0.48) proteins. In contrast, a positive correlation was observed for the N protein (p=0.021, R=0.37) (Figure 4). Consistently, other significant correlations were also found when considering different epidemiological parameters that are depicted in Figures S8-11.

In fact, a stronger association with disease outcome was found when evaluating the ratio between population's AUC predicted for S and N proteins together versus the number of deaths per million inhabitants (R=-0.56, p=0.00028, Pearson correlation, Figure 5). This result highlights 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 calculated the frequency of SB peptides derived from S or N proteins by dividing the number of SB peptides by their respective lengths as in Table S1 to account for the difference in protein sizes. A ratio was calculated dividing the SB frequency from protein S by the frequency of SB in protein N. The majority of HLA-A and B alleles have a ratio superior to 1 meaning that they preferentially bind SB peptides derived from S rather than from N protein (Figure S12A-B). The alleles with the highest ratios are HLA-A24:02 (13.16) and HLA-B18:01 (10.5) in contrast to HLA-A30:01 (0.57) and HLA-B-07:02 (0.42), showing the lowest ratios. HLA-A24:02 has an elevated AF in JPN (0.3790) and NZL (0.333) both with a lower number of deaths per million while HLA-B18:01 has the highest AF in GRE (15.80). On the other hand, potentially high-risk 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.

We highlighted the SARS-CoV-2 derived SB peptides generated within each ORF (left Y-axis, Figure 6A) or protein as well as the number of HLA-I alleles capable of presenting them (right Y-axis, Figure 6A). We observed regions capable of generating an elevated number of unique SB peptides (black bars - top) in comparison to others and the ability of each of these peptides to bind to several HLA-I alleles (blue dots - bottom), which can help selecting vaccine candidates. We found 40 unique SB peptides falling within the receptor-binding motif and 7 SB peptides derived from the 4 amino acids insertion (19) in the S1/S2 cleavage site (highlighted in purple in the Figure 6B). These 7 SB peptides vary greatly in the range of affinities and also the number of HLA binders (Table 1). It is also important to note that a 28 amino acid region at the end of S protein (GIVNNTVYDPLQPELDSFKEELDKYFKN) does not generate any potential HLA-I A or B restricted peptide considering the 111 HLAs analyzed.

Altogether, these results indicate a distinct pattern of virus recognition and response among populations. HLA allele profiles and 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.

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 populations' 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 population outcome relies partially on their ability to present peptides derived from the S protein while those derived from the N protein correlate with unfavorable outcomes.

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 (20). A threshold of 0.75 for the

This is a provisional file, not the final typeset article

sum of the allele frequencies for HLA-A and B within each population was established to make binding predictions feasible and to represent the vast majority of individuals.

Numerous factors such as political rules implementation and 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). In addition, other factors could influence mortality, affecting the negative relationship between deaths per million and Spike / N proteins AUC ratios. 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. Several factors could potentially impact these outcomes. As an example, it was proposed that BCG vaccination could be associated with SARS-CoV-2 less severe outcomes. In line with this hypothesis, ITA, ESP, FRA, GBR, USA, and HOL are among the countries that never had or stopped BCG vaccination policy (16). We observed that some countries also had lower numbers of death per million than expected (i.e.: THA, SEN, GHA). These countries had a lower percentage of obese people, which is a risk factor for COVID-19-related death and severity (21). 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.

HLA binding predictions have become a useful and important method for selecting promising peptides for vaccination. Although the binding predictions do not account for peptide cutting and TAP processing, this has been proved to be a robust method since it relies on several HLA-eluted peptides for the algorithm training. Moreover, by comparing the predicted binding affinities with the affinities verified through assays deposited at the IEDB, we showed that SB peptides, 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. The negative correlation between coverage for protein S and deaths per million inhabitants corroborates a study showing that CD8 T lymphocytes from convalescent patients have more targets for the S than for the N protein (3). 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 in 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.

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 (19). We found unique SB peptides falling within the receptor binding motif sequence, including the 4 amino acid insertion in the S1/S2 cleavage site. This sequence is considered to be related to the high virulence of SARS-CoV-2 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 GIVNNTVYDPLQPELDSFKEELDKYFKN 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 (2020) resemble 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 (22). Since the homology between N proteins of SARS-CoV-2 and SARS-CoV is over 90% (13) and the SARS-CoV-2 peak viral load is earlier (23), 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 (3), indicating 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 (4,24,25). The number of deaths per million inhabitants showed a significant positive correlation only when correlated with N coverage, reinforcing the possibility that N peptides lead to a more aggressive disease. This deregulation of the immune system has been shown to occur in other viral infections and could be dependent on CD8 T lymphocytes (26).

Finally, we highlighted some important viral regions predicted to be presented by a vast number of HLA alleles, which can help understanding 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, these strategies must avoid using the N protein since it was the only region showing a positive correlation with the number of deaths per million. Last, we identified potential antigens derived from the 4 amino acid insertion of SARS-CoV-2 that is absent from previous strains and could serve as a guide for SARS-CoV-2 specific vaccine development.

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, 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 of inhabitants and more than 1 thousand total cases were evaluated.

HLA-I gathering

Human Leukocyte Antigen (HLA) I alleles for A and B genes with at least 4 digits and its frequencies in the selected populations were retrieved from the Allele Frequency Net Database (27) between 23/04/2020 and 07/05/2020. Larger datasets that were not part of anthropological studies or minority ethnic groups were preferred (Table S1). The allele frequency of the most frequent HLA alleles was added up to reach a minimum coverage of 0.75 for each population. Allele frequencies with resolution greater than 4 digits were combined, e.g. for BRA: B*07:02:01 and B*07:02:03 became B*07:02.

This is a provisional file, not the final typeset article

Accessing experimental binding affinities

We searched at the Immune Epitope Database (IEDB) on 18/05/2020 for MHC Ligand assays 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.

Binding prediction

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/nuccore/MT019529). Binding predictions were performed using netMHCpan4 (18) with peptide lengths from 8 to 11 for the entire viral genome. Only strong binder peptides (%Rank <0.5) were kept for further analysis. We used blastp v2.9.0 to eliminate peptides that matched protein sequences from the Ensembl human genome GRCh38 with 100% of identity.

Population Coverage

We used the tool described by Bui and collaborators (12) available at IEDB (http://tools.iedb.org/population/) to calculate the population coverage. The inputs used were the same allele frequencies 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. Coverage 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 Pearson correlations.

Data analysis

The analysis was conducted using the R environment version 3.6 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 and ggbio v1.32.

Author Contributions

MB, MAPMP, and RGG contributed to the conception of the study. GFV, AB, MHB and MB contributed with the design of the study and important scientific discussions. MAPMP and RGG performed all the analysis and wrote the first draft of the manuscript. All the authors contributed to manuscript revision, read, and approved the submitted version. MB was responsible for the final approval of the submitted version.

Conflict of Interest Statement

There is no conflict of interest.

Acknowledgments

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

Funding

There is no fund associated with this study.

References

1. **WHO. COVID-19 situation reports. (2020) Available at: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports [Accessed May 20, 2020]**

2. **Wu F, Zhao S, Yu B, Chen Y-M, Wang W, Song Z-G, Hu Y, Tao Z-W, Tian J-H, Pei Y-Y, et al. A new coronavirus associated with human respiratory disease in China. Nature (2020) 579:265–269. doi:10.1038/s41586-020-2008-3**

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) 27:671-680.e2. doi:10.1016/j.chom.2020.03.002**

4. **Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, Zhang L, Fan G, Xu J, Gu X, 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**

5. **Wang D, Hu B, Hu C, Zhu F, Liu X, Zhang J, Wang B, Xiang H, Cheng Z, Xiong Y, et al. Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China. JAMA (2020) doi:10.1001/jama.2020.1585**

6. **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) 203:631–637. doi:10.1002/path.1570**

7. **Liao M, Liu Y, Yuan J, Wen Y, Xu G, Zhao J, Chen L, Li J, Wang X, Wang F, et al. The landscape of lung bronchoalveolar immune cells in COVID-19 revealed by single-cell RNA sequencing. medRxiv (2020) doi:10.1101/2020.02.23.20026690**

8. **Tay MZ, Poh CM, Rénia L, MacAry PA, Ng LFP. The trinity of COVID-19: immunity, inflammation and intervention. Nat Rev Immunol (2020) doi:10.1038/s41577-020-0311-8**

9. **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: doi:10.3390/vaccines8020153**

10. **Grifoni A, Weiskopf D, Ramirez SI, Mateus J, Dan JM, Moderbacher CR, Rawlings SA, Sutherland A, Premkumar L, Jadi RS, et al. Targets of T cell responses to SARS-CoV-2 coronavirus in humans with COVID-19 disease and unexposed individuals. Cell (2020) doi:10.1016/j.cell.2020.05.015**

11. **Shi Y, Wang Y, Shao C, Huang J, Gan J, Huang X, Bucci E, Piacentini M, Ippolito G, Melino G. COVID-19 infection: the perspectives on immune responses. Cell Death Differ (2020) 27:1451–1454. doi:10.1038/s41418-020-0530-3**

12. **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 Bioinformatics (2006) 7:153. doi:10.1186/1471-2105-7-153**

13. **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: doi:10.3390/v12030254**

This is a provisional file, not the final typeset article

14. 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]

15. 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

16. 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

17. 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&id=17555 [Accessed May 26, 2020]

18. Jurtz V, Paul S, Andreatta M, Marcatili 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) 199:3360–3368. doi:10.4049/jimmunol.1700893

19. Jaimes JA, André NM, Millet JK, Whittaker GR. Structural modeling of 2019-novel coronavirus (nCoV) spike protein reveals a proteolytically-sensitive activation loop as a distinguishing feature compared to SARS-CoV and related SARS-like coronaviruses. BioRxiv (2020) doi:10.1101/2020.02.10.942185

20. 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

21. Kass DA, Duggal P, Cingolani O. Obesity could shift severe COVID-19 disease to younger ages. Lancet (2020) 395:1544–1545. doi:10.1016/S0140-6736(20)31024-2

22. Tan Y-J, Goh P-Y, Fielding BC, Shen S, Chou C-F, Fu J-L, Leong HN, Leo YS, Ooi EE, Ling AE, 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) 11:362–371. doi:10.1128/cdli.11.2.362-371.2004

23. Zou L, Ruan F, Huang M, Liang L, Huang H, Hong Z, Yu J, Kang M, Song Y, Xia J, et al. SARS-CoV-2 Viral Load in Upper Respiratory Specimens of Infected Patients. N Engl J Med (2020) 382:1177–1179. doi:10.1056/NEJMc2001737

24. Zhou F, Yu T, Du R, Fan G, Liu Y, Liu Z, Xiang J, Wang Y, Song B, Gu X, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. Lancet (2020) 395:1054–1062. doi:10.1016/S0140-6736(20)30566-3

25. Qin C, Zhou L, Hu Z, Zhang S, Yang S, Tao Y, Xie C, Ma K, Shang K, Wang W, et al. Dysregulation of immune response in patients with COVID-19 in Wuhan, China. Clin Infect Dis (2020) doi:10.1093/cid/ciaa248

26. Walsh KB, Teijaro JR, Brock LG, Fremgen DM, Collins PL, Rosen H, Oldstone MBA. 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) 88:6281–6293. doi:10.1128/JVI.00464-14

27. González-Galarza FF, Takeshita LYC, Santos EJM, Kempson F, Maia MHT, da Silva ALS, Teles e Silva AL, Ghattaoraya GS, Alfirevic A, Jones AR, 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

Figure legends

Figure 1. **Unsupervised two-dimensional cluster analysis of 38 countries by HLA-A and HLA-B Allele frequencies.** Each row represents a population, the first column represents death per million inhabitants quartiles: Q1 = green (lower than 5); Q2 = yellow (5 to 20); Q3 = orange (21 to 85); Q4 = red (greater than 85). The other columns represent HLA alleles, which frequencies values are represented by the blue scale color bar.

Figure 2. Number of Strong Binder peptides for the 111 analyzed HLA-I alleles. **(A)** 33 HLA-A and **(B)** 78 HLA-B were considered for the analysis. 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.

Figure 3. 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 as a function of the fraction of each population (%). **(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 AUC are 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 4. Correlations between SARS-CoV-2 derived SB peptides coverage of each population and COVID-19 outcome. Pearson 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.

Figure 5. Proteins S and N potentially trigger opposite responses associated with COVID-19. Pearson correlation between the S/N AUC ratio and deaths per million inhabitants. Colors represent the countries inside the confidence interval with a higher (red) or lower (blue) number of deaths. Names are shown for those that are discussed in the text.

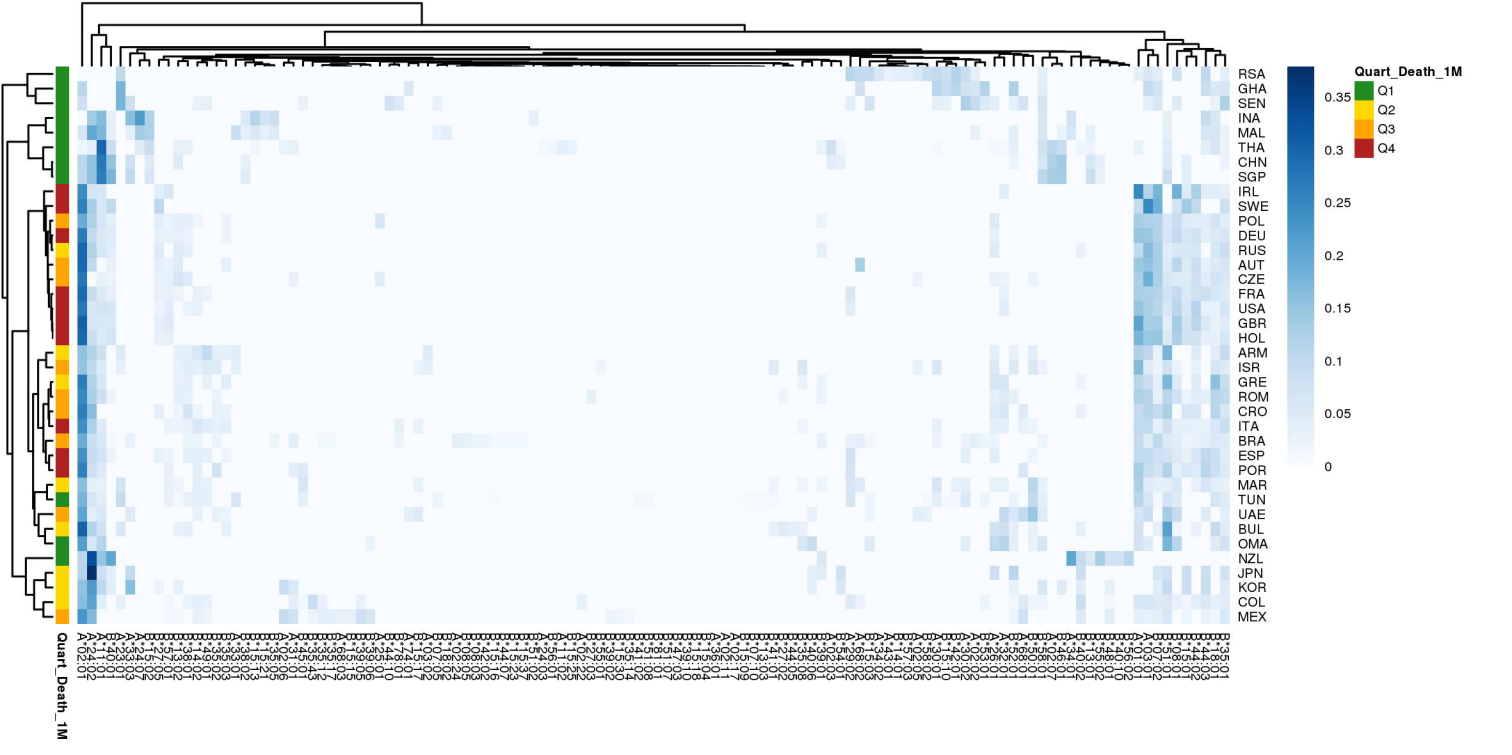
Figure 6. Landscape of SARS-CoV-2-derived SB peptides. **(A)** The viral proteins and ORFs are represented by colored rectangles proportional to their size. Black bars represent unique SB peptides potentially presented by populations. Blue dots represent the number of HLA-I alleles that a given SB peptide binds to. **(B)** Detail of the S protein and its domains and the SB peptides derived from each region. The seven SB peptides that contain at least one of the 4 amino acid insertion are highlighted in purple. Length is represented in aa. RBM, Receptor Binding Motif; RBD, Receptor Binding Domain; S1/S2, S1/S2 cleavage site; S2', S2 cleavage site.

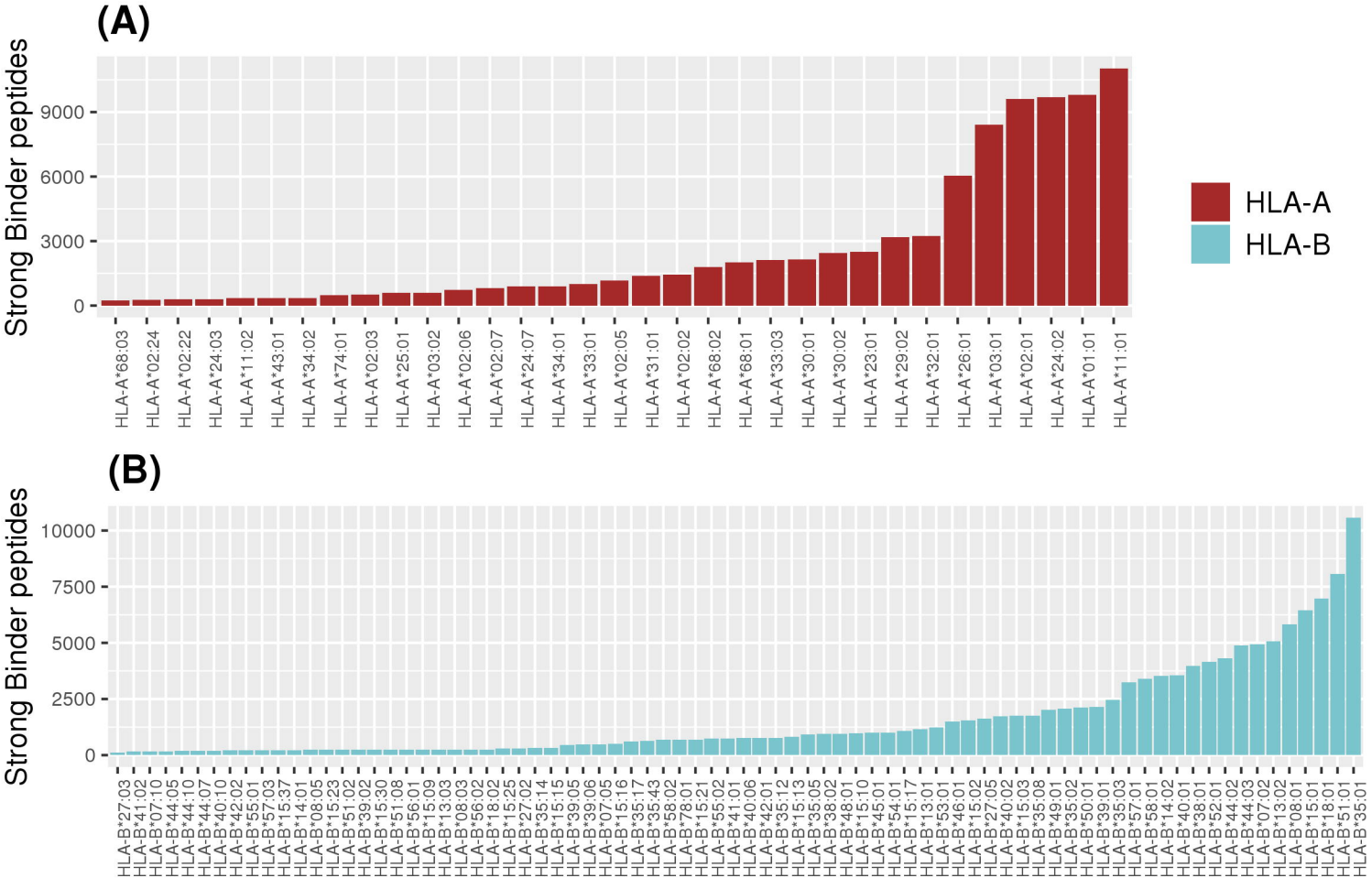
Table 1. Strong binder peptides including some of the 4 amino acid insertion in protein S, which are highlighted in bold.

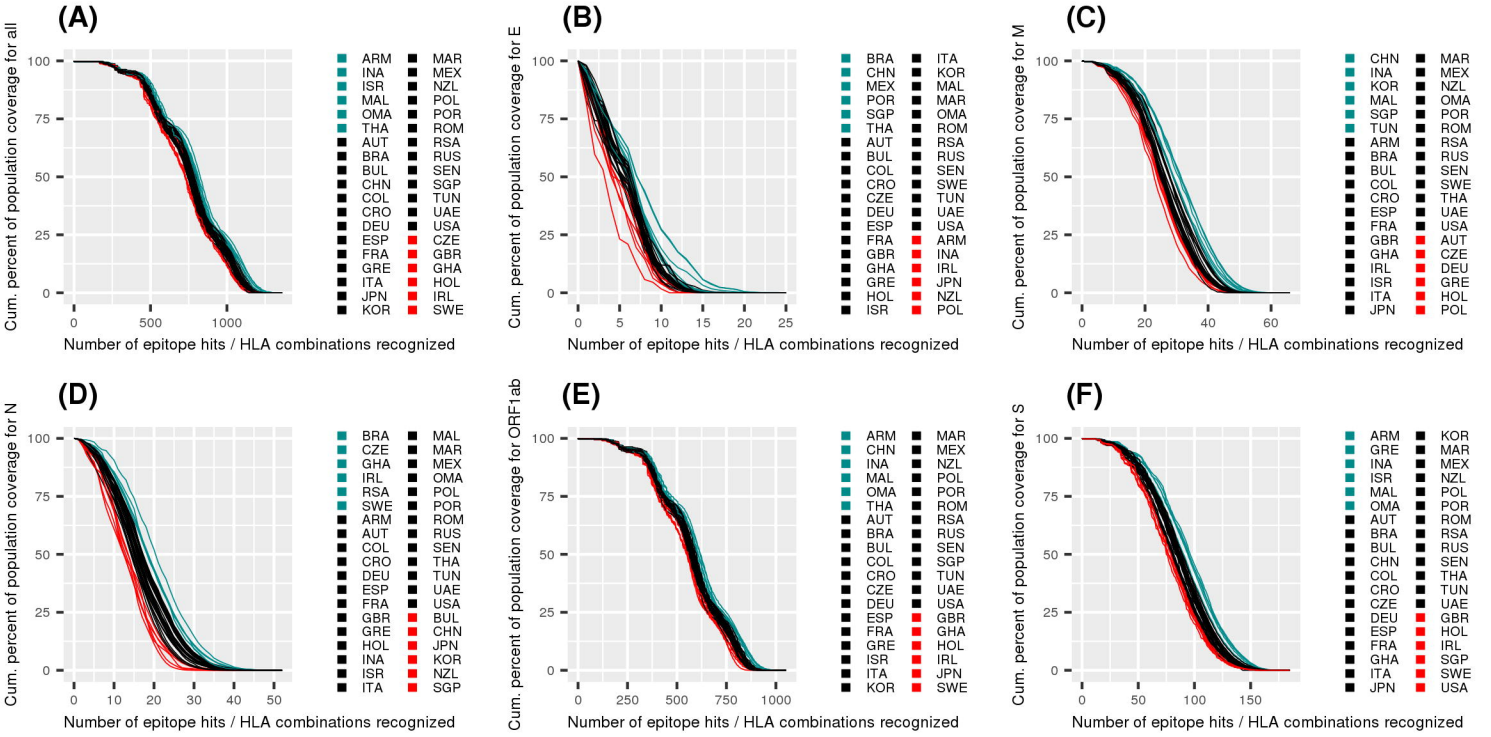
This is a provisional file, not the final typeset article

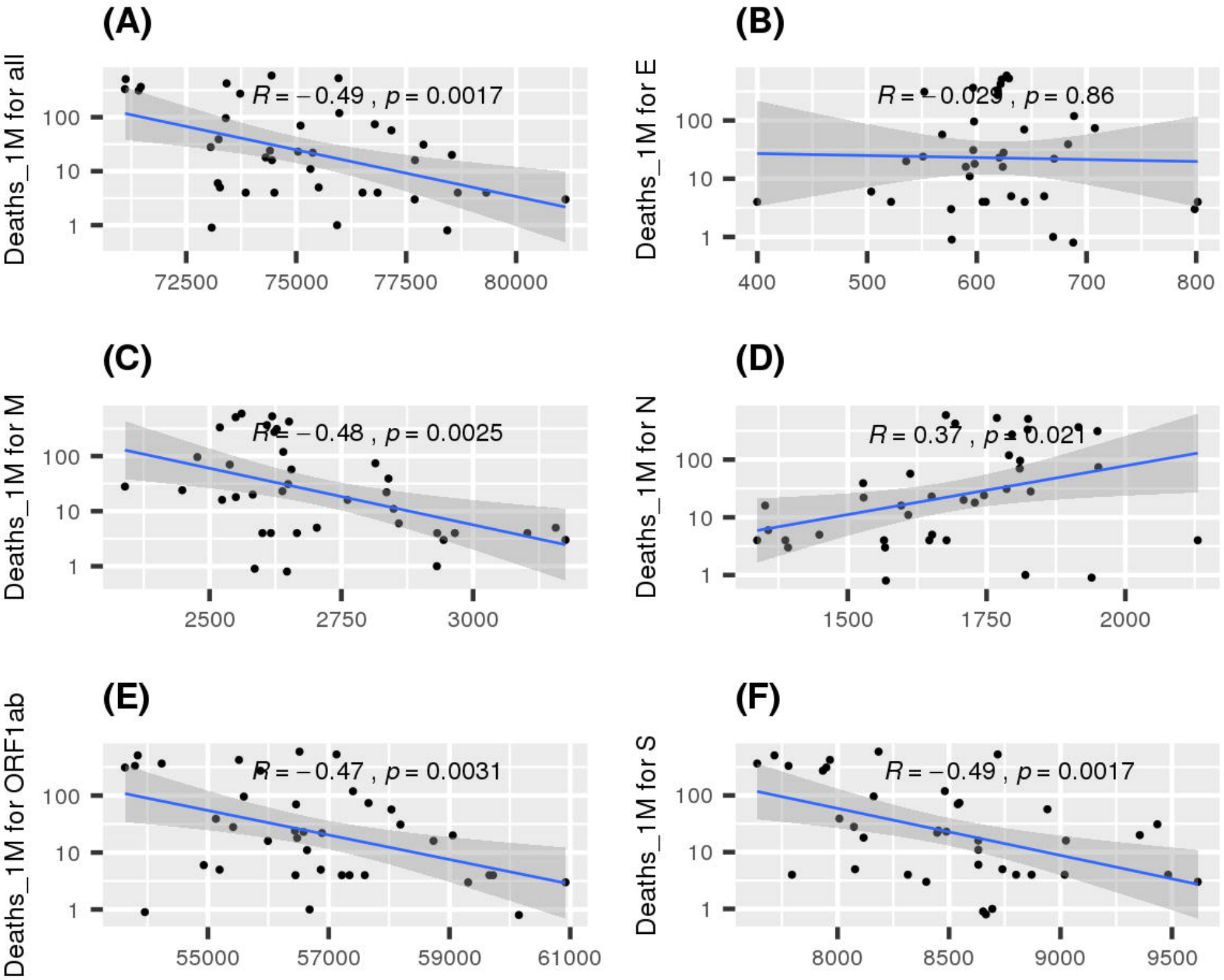
Peptide	**Predicted Aff(nM)**	**Number of HLA binders**	**Start (bp)**	**End (bp)**
NS**PRRA**RSVA	352.6 (159.0-735.6)	3	679	688
QTNS**PRRA**R	54.65 (45.4-63.9)	2	677	685
RARSVASQS	26.4	1	682	691
RARSVASQSI	55.2	1	683	692
S**PRRA**RSV	256.73 (79.6-636.6)	4	680	687
S**PRRA**RSVA	335.11 (6.6-1504.7)	10	680	688
S**PRRA**RSVAS	323.63 (75.1-810.6)	3	680	689

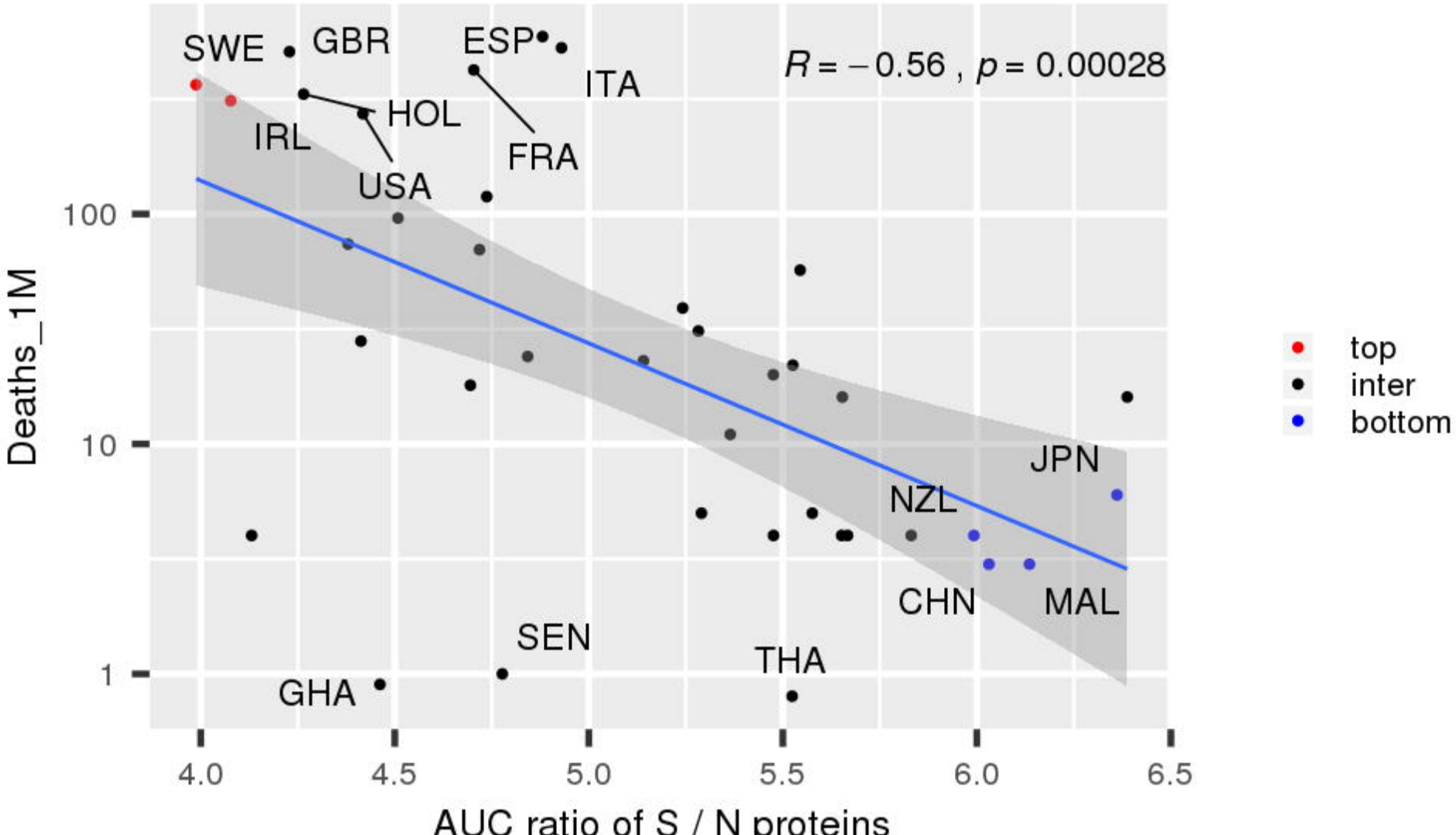
505

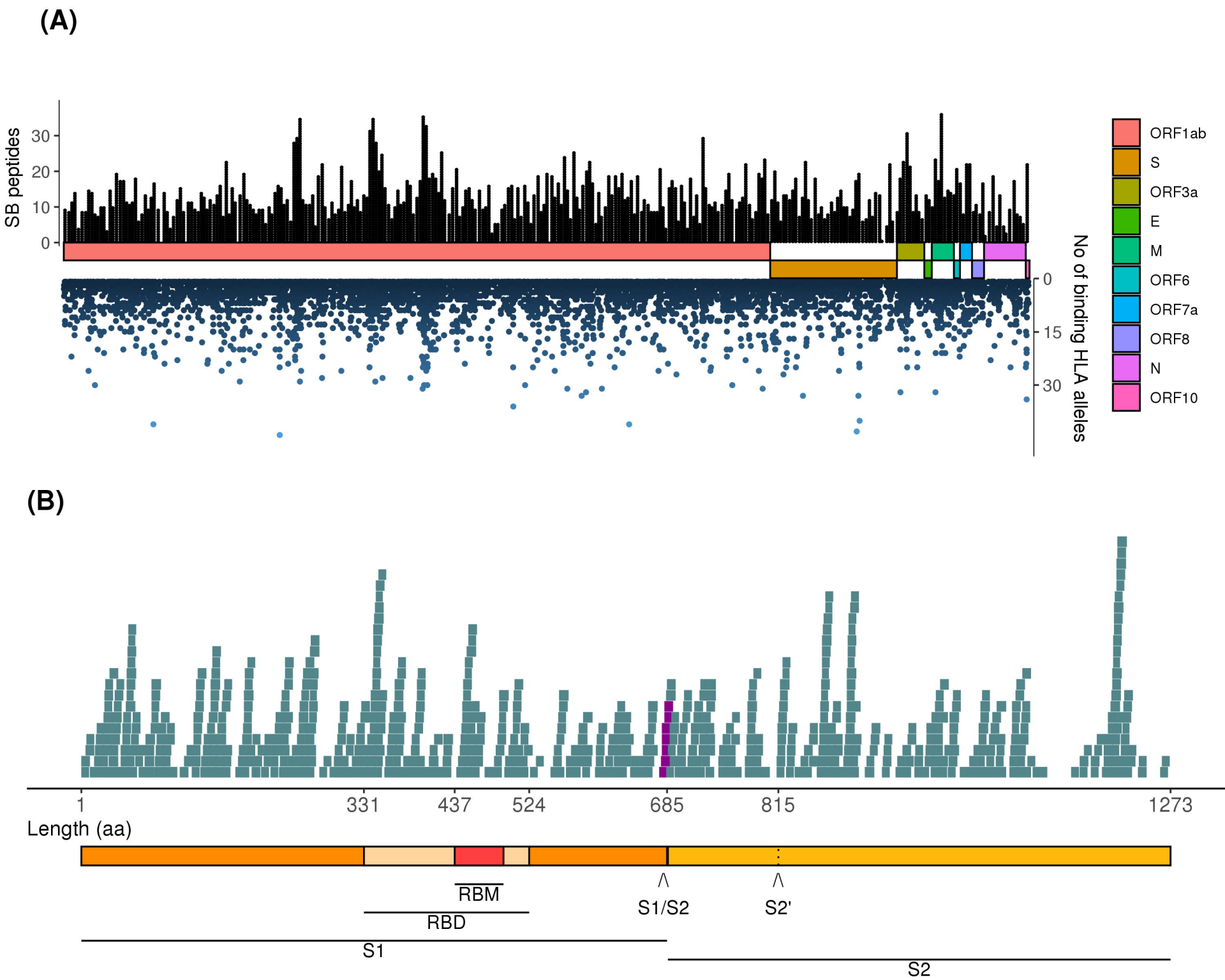

Quart_Death_1M
Q1
Q2
Q3
Q4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
RSA
GHA
SEN
INA
MAL
THA
CHN
SGP
IRL
SWE
POL
DEU
RUS
AUT
CZE
FRA
USA
GBR
HOL
ARM
ISR
GRE
ROM
CRO
ITA
BRA
ESP
POR
MAR
TUN
UAE
BUL
OMA
NZL
JPN
KOR
COL
MEX
Quart_Death_1M


(A)
Strong Binder peptides
9000
6000
3000
0
HLA-A*68:03
HLA-A*02:24
HLA-A*02:22
HLA-A*24:03
HLA-A*11:02
HLA-A*43:01
HLA-A*34:02
HLA-A*74:01
HLA-A*02:03
HLA-A*25:01
HLA-A*03:02
HLA-A*02:06
HLA-A*02:07
HLA-A*24:07
HLA-A*34:01
HLA-A*33:01
HLA-A*02:05
HLA-A*31:01
HLA-A*02:02
HLA-A*68:02
HLA-A*68:01
HLA-A*33:03
HLA-A*30:01
HLA-A*30:02
HLA-A*23:01
HLA-A*29:02
HLA-A*32:01
HLA-A*26:01
HLA-A*03:01
HLA-A*02:01
HLA-A*24:02
HLA-A*01:01
HLA-A*11:01
HLA-A
HLA-B
(B)
Strong Binder peptides
10000
7500
5000
2500
0
HLA-B*27:03
HLA-B*41:02
HLA-B*07:10
HLA-B*44:05
HLA-B*44:10
HLA-B*44:07
HLA-B*40:10
HLA-B*42:02
HLA-B*55:01
HLA-B*57:03
HLA-B*15:37
HLA-B*14:01
HLA-B*08:05
HLA-B*15:23
HLA-B*51:02
HLA-B*39:02
HLA-B*15:30
HLA-B*51:08
HLA-B*56:01
HLA-B*15:09
HLA-B*13:03
HLA-B*08:03
HLA-B*56:02
HLA-B*18:02
HLA-B*15:25
HLA-B*27:02
HLA-B*35:14
HLA-B*15:15
HLA-B*39:05
HLA-B*39:06
HLA-B*07:05
HLA-B*15:16
HLA-B*35:17
HLA-B*35:43
HLA-B*58:02
HLA-B*78:01
HLA-B*15:21
HLA-B*55:02
HLA-B*41:01
HLA-B*40:06
HLA-B*42:01
HLA-B*35:12
HLA-B*15:13
HLA-B*35:05
HLA-B*38:02
HLA-B*48:01
HLA-B*15:10
HLA-B*45:01
HLA-B*54:01
HLA-B*15:17
HLA-B*13:01
HLA-B*53:01
HLA-B*46:01
HLA-B*15:02
HLA-B*27:05
HLA-B*40:02
HLA-B*15:03
HLA-B*35:08
HLA-B*49:01
HLA-B*35:02
HLA-B*50:01
HLA-B*39:01
HLA-B*35:03
HLA-B*57:01
HLA-B*58:01
HLA-B*14:02
HLA-B*40:01
HLA-B*38:01
HLA-B*52:01
HLA-B*44:02
HLA-B*44:03
HLA-B*07:02
HLA-B*13:02
HLA-B*08:01
HLA-B*15:01
HLA-B*18:01
HLA-B*51:01
HLA-B*35:01


(A)
Cum. percent of population coverage for all
Number of epitope hits / HLA combinations recognized
ARM MAR
INA MEX
ISR NZL
MAL POL
OMA POR
THA ROM
AUT RSA
BRA RUS
BUL SEN
CHN SGP
COL TUN
CRO UAE
DEU USA
ESP CZE
FRA GBR
GRE GHA
ITA HOL
JPN IRL
KOR SWE
(B)
Cum. percent of population coverage for E
Number of epitope hits / HLA combinations recognized
BRA ITA
CHN KOR
MEX MAL
POR MAR
SGP OMA
THA ROM
AUT RSA
BUL RUS
COL SEN
CRO SWE
CZE TUN
DEU UAE
ESP USA
FRA ARM
GBR INA
GHA IRL
GRE JPN
HOL NZL
ISR POL
(C)
Cum. percent of population coverage for M
Number of epitope hits / HLA combinations recognized
CHN MAR
INA MEX
KOR NZL
MAL OMA
SGP POR
TUN ROM
ARM RSA
BRA RUS
BUL SEN
COL SWE
CRO THA
ESP UAE
FRA USA
GBR AUT
GHA CZE
IRL DEU
ISR GRE
ITA HOL
JPN POL
(D)
Cum. percent of population coverage for N
Number of epitope hits / HLA combinations recognized
BRA MAL
CZE MAR
GHA MEX
IRL OMA
RSA POL
SWE POR
ARM ROM
AUT RUS
COL SEN
CRO THA
DEU TUN
ESP UAE
FRA USA
GBR BUL
GRE CHN
HOL JPN
INA KOR
ISR NZL
ITA SGP
(E)
Cum. percent of population coverage for ORF1ab
Number of epitope hits / HLA combinations recognized
ARM MAR
CHN MEX
INA NZL
MAL POL
OMA POR
THA ROM
AUT RSA
BRA RUS
BUL SEN
COL SGP
CRO TUN
CZE UAE
DEU USA
ESP GBR
FRA GHA
GRE HOL
ISR IRL
ITA JPN
KOR SWE
(F)
Cum. percent of population coverage for S
Number of epitope hits / HLA combinations recognized
ARM KOR
GRE MAR
INA MEX
ISR NZL
MAL POL
OMA POR
AUT ROM
BRA RSA
BUL RUS
CHN SEN
COL THA
CRO TUN
CZE UAE
DEU GBR
ESP HOL
FRA IRL
GHA SGP
ITA SWE
JPN USA


(A)
Deaths_1M for all
R = −0.49 , p = 0.0017
100
10
1
72500
75000
77500
80000
(B)
Deaths_1M for E
R = −0.029 , p = 0.86
100
10
1
400
500
600
700
800
(C)
Deaths_1M for M
R = −0.48 , p = 0.0025
100
10
1
2500
2750
3000
(D)
Deaths_1M for N
R = 0.37 , p = 0.021
100
10
1
1500
1750
2000
(E)
Deaths_1M for ORF1ab
R = −0.47 , p = 0.0031
100
10
1
55000
57000
59000
61000
(F)
Deaths_1M for S
R = −0.49 , p = 0.0017
100
10
1
8000
8500
9000
9500


SWE
GBR
ESP
ITA
HOL
IRL
FRA
USA
R = −0.56 , p = 0.00028
JPN
NZL
CHN
MAL
SEN
THA
GHA
100
10
1
Deaths_1M
4.0
4.5
5.0
5.5
6.0
6.5
AUC ratio of S / N proteins
top
inter
bottom


(A)
SB peptides
30
20
10
0
No of binding HLA alleles
0
15
30
ORF1ab
S
ORF3a
E
M
ORF6
ORF7a
ORF8
N
ORF10
(B)
1
331
437
524
685
815
1273
Length (aa)
RBM
RBD
S1/S2
S2'
S1
S2