Can the protection be among us? Previous viral contacts and prevalent HLA alleles could be avoiding an even more disseminated COVID-19 pandemic.

Eduardo Cheuiche Antonio[1,3], Mariana Rost Meireles[1,3], Marcelo Alves de Souza Bragatte[1], Gustavo Fioravanti Vieira[1,2,4,*]

1 - Post Graduation Program in Genetics and Molecular Biology - Universidade Federal do Rio Grande do Sul - Brazil

2 - Post Graduation Program in Health and Human Development - Universidade La Salle Canoas - Brazil

3 - These authors contributed equally

4 - Lead contact

*Correspondence: gustavo.vieira@unilasalle.edu.br

Summary

*Background***:** COVID-19 is bringing scenes of sci-fi movies into real life, and it seems to be far from over. Infected individuals exhibit variable severity, with no relation between the number of cases and mortality, suggesting the involvement of the populational genetic constitution and previous cross-reactive immune contacts in the individuals' disease outcome.

Methods: A clustering approach was conducted to investigate the involvement of human MHC alleles with individuals' outcomes. HLA frequencies from affected countries were used to fuel the Hierarchical Clusterization Analysis. The formed groups were compared regarding their death rates. To prospect the T cell targets in SARS-CoV-2, and by consequence, the epitopes that are conferring cross-protection in the current pandemic, we modeled 3D structures of HLA-A*02:01 presenting immunogenic epitopes from SAR-CoV-1, recovered from Immune Epitope Database. These pMHC structures were also compared with models containing the corresponding SARS-CoV-2 epitope, with alphacoronavirus sequences, and with a panel of immunogenic pMHC structures contained in CrossTope.

Findings: The combined use of HLA-B*07, HLA-B*44, HLA-DRB1*03, and HLADRB1*04 allowed the clustering of affected countries presenting similar death rates, based only on their allele frequencies. SARS-CoV HLA-A*02:01 epitopes were structurally investigated. It reveals molecular conservation between SARS-CoV-1 and SARS-CoV-2 peptides, enabling the use of formerly SARS-CoV-1 experimental epitopes to inspect actual targets that are conferring cross-protection. Alpha-CoVs and, impressively, viruses involved in human infections share fingerprints of immunogenicity with SARS-CoV peptides. Interpretation: Wide-scale HLA genotyping in COVID-19

patients shall improve prognosis prediction. Structural identification of previous triggers paves the way for herd immunity examination and wide spectrum vaccine development.

Funding: This work was supported by the National Council for Scientific and Technological Development (CNPq) and National Council for the Improvement of Higher Education (CAPES) for their support.

Keywords

SARS-CoV-2, Covid-19, Cross-reactivity, T cell epitopes, structural analysis, immunogenic fingerprints, genetic background, populational heterogeneity, disease outcome.

Introduction

COVID-19 is bringing scenes of sci-fi movies into real life. Considering more than seven million of already diagnosed cases and the growing number reported in the past few months, it seems to be far from over. The coronaviridae family also includes other respiratory syndrome causative agents in humans, the SARS-CoV, and MERS-CoV viruses[1]. The emergence of this viral pneumonia started in December/19 in the Hubei province, central China[2], rapidly spreading around the globe, receiving the pandemic status at the beginning of March 2020. In the global actual scenario, there are almost 400.000 virus causative deaths, according to COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University[3]- *recovered at June 04, 2020.* Nevertheless, we should consider that there is not a regular correlation between the number of diagnosed individuals and mortality, among the affected countries, as shown in **Table 1.**

The United States, for example, shows the highest number of cases, which may be due to a better diagnosis rate, which influences the death rate calculation. In this context, Russia is the third country with a high number of cases, although its low mortality rate (one to 107 cases). Other countries with the most infected populations also present a low proportion of morbidity in diagnosed individuals as Germany (1:21). These values are quite different from those found in Italy (1:7) and France (1: 6.6), for example. It becomes even more alarming when we realize that each number is a human life.- *calculated data recovered at May 26, 2020.*

What elements could be dictating these differences? Social and cultural differences? The diverse starting point of government measures? Or genetic background presented by the different populations and SARS-CoV-2 strains around the world, causing distinct immune response profiles in critical patients? The present work will discuss this last question.

In the current pandemics, the first clues over the importance of cellular response are arising. For example, a work from Diao et al. (2020) found a negative correlation between CD4+ and CD8+ cell counts and disease severity in 522 patients from Wuhan, with laboratory-confirmed COVID-19. This fact places a pool of especial proteins orchestrating cellular immune mechanisms, as central players. Most of their genes are located on MHC locus. The MHC is the most polymorphic locus in the human genome[4]. This great variability enables that animal species could be able to fight, especially in a populational level, with the comparable mutational potential of pathogens, as viruses. Briefly, it is involved with the processing and presentation of small peptides in cell surfaces, derived from intracellular proteins, allowing the immune system to discriminate self from non-self, thus eliminating pathogens and tumors. At this point, a critical question is raised. We mentioned the MHC locus intentionally, not only MHC genes, considering that other proteins, fundamental to produce true epitopes, belong to this genomic region. Many studies, aiming to prospect tumoral or vaccine targets, focuses its prediction only on the pathogen or cancer sample MHC ligandome[5]. A potential to bind to different alleles do not confer to peptides their full potential to be a T cell epitope. A full T cell synapsis triggering demands additional requirements such as epitope immunodominance and pMHC:TCR physicochemical complementarity. Thus, *in silico* analysis considering additional steps on the antigen processing pathway and comparisons among putative targets and immunogenic epitopes, could present a better performance to prospect actual T cell epitopes, as in the current situation where no previous information is available.

Presenting a central role in this process, the HLAs (peptides presenting molecules in humans) constitute a link between the immune system surveillance and intracellular space status. It is known that different HLA alleles can bind and present a specific viral protein region (peptides) with altered efficiencies, which could provoke both susceptibility or resistance to a disease caused by a specific pathogen, depending on the type of MHC that the individual possesses.[6,7]

Thus, some HLA allotypes are unable to present some immunodominant epitopes (those responsible to initiate T-cell responses) from a specific virus, precluding the detection and fighting by the immune system. The opposite also occurs, the existence of HLAs with improved potential to present optimal viral targets, allowing the infection control[8]. The HLA alleles frequency vary among different populations. Some researches suggest that HLA alleles conferring pathogen resistance are most prevalent in areas with endemic diseases.

An important point to highlight in the current situation is that, even if the allele frequencies do not directly correlate with the percentage of the infected individuals on affected countries, what it can be explained by the probably enhanced affinity of SARS-CoV-2 spike protein by the ACE2 receptor[9,10], the same cannot be said about differential outcomes of critical patients, without a deep HLA genotyping and investigation.

One can argue that this has already be done, in the 2003 SARS outbreak, with few of none significant associations. Nevertheless, we should consider that the number of genotyped individuals in that situation was extremely low compared with the current pandemic. Beyond that, many volunteers in those studies were health workers with putative contact with the virus[11]. Now, unfortunately, we are facing a scenario where we could extract more reliable correlations with alleles indicating better/worse prognosis in COVID19 patients. Large populational samples are essential in studies where we are dealing with a vast number of alleles (variables), as indicated above, contributing to minor effects, which could be important at a global level.

Results and Discussion

The HLA frequency relation and its apparent effect in COVID-19 immune responses

The virus's ability to infect and spread in its pandemic should be related to a shared common global characteristic, as mentioned before. So, it is clear that the infection propagates disconnected from populations' habits and origins, but the lethality varies between them. Important information arises from attending for the most frequent/rare MHC alleles in the most affected regions. The frequencies of these different HLA alleles vary widely among the countries with the highest number of cases [12]. It may indicate a variation in the infection responses, which can influence morbidity. A close

examination pointed out that none of the alleles frequency seems to be able to explain the death rate disparities independently.

Aiming to verify the alleles frequencies influences in disorder outcomes, and in the pandemic picture, we incorporated them into a Hierarchical Clustering Analysis (HCA) using the pvclust package at the RStudio platform. The inclusion of specific human leukocyte antigens (HLA) was refined based on previous studies references and its frequencies on the alleles data bank (considerable prevalence) distributed among the ten more predominant countries in terms of COVID-19 cases, plus the country where occurs the first epicenter of pandemics. Starting from 40 alleles we were excluding them by considering an absence of frequency information for some of the investigated countries and the preliminary HCA results. The final alleles selection includes four allotypes: HLA-B*07, HLA-B*44, HLA-DRB1*03, and HLA-DRB1*04. Each of them is a representative member of a distinct supertype (determining differential recognition and responses in the immune system). The HLA alleles DRB1*03 and DRB1*04 were already described as involved with autoimmune reactions elicitation [4]. In 2006, some works detected autoantibodies directed against lung epithelial cells antigens, which could be mediating tissue damage, in some stages of the disease[13]. In this sense, alleles from the DRB1 gene may play a role in presenting targets from the virus sharing features with self-antigens. Furthermore, some studies highlighted the role of B*07 and DRB1 alleles in the development of SARS susceptibility and resistance[14], and described HLA-B*44 as a predictive element in Hepatitis C Virus clearance[15]. Even considering the prevalence and importance of HLA-A*02 supertype, it was not included in HCA analysis considering its low contrast ratio between the countries. The values used in the HCA were the weighted average of serotype frequencies from all populations belonging to each considered country. **Table 1** summarizes the cited data. Additional negative symbols were artificially attributed to DRB1*03 values intending to detach its alleles frequencies in the HCA. This action was a result of a previous visual inspection correlating its subtle prominence in countries presenting a worse prognosis, classifying the allele as a potential determinant factor.

Figure 1 shows the result for the HCA. The formed clusters seem to explain (in a general way) the global pandemic scenario with exciting results that indicate active participation of the selected alleles and its consequent impact on how the COVID-19 affects countries population. Italy and France are the two more closely related countries, forming a ternary group with Spain. They have, respectively, a mortality index for closed cases of 1 to 5.24, and 1 to 3.41, while in Spain it is from 1:7.87.

Taking into account the active cases the death per case rate decreases, being of 1:7.06, 1:6.60, and 1:10.04, in Italy, France, and Spain, respectively (*all values were calculated considering numbers recovered by May, 23)* [3]. This group comprises three relevant nations considered as former significant pandemic epicenters, presenting high death proportions regarding the number of reported cases. Fortunately, the death and cases rates in these countries have been falling, probably due to protective measures improving their prognosis. A related cluster includes the United States, the United Kingdom, and Brazil. Currently, these countries represent a prominent alarming status, possessing the actual high mortality indexes with an elevated contagious level. Its estimations of death per closed cases and death per general cases show quite a difference (from 1:7.49 to 1:15 in Brazil; 1:5.53 to 1:16.82 in the US, indicating they are in its epidemic course. The UK has no available information about the closed cases, only the death per case index (1 to 7.08). The next cluster grouped Germany and Russia. These two countries present excellent survival indexes. The death per closed cases in Germany is similar to death per general cases (1:20.11 and 1:21.89, respectively), it does not occur in Russia (1:32.86 and 1:106.8) due to a large number of active cases. Turkey branch apart from the above-mentioned nations, maintaining good values of the considered indexes (1:36.1 and 1:28.3). China (1:17.9 and 1:17.89) and Iran (1:17.36 and 1:15.14) fell in a separate group, having very similar death rates, despite their ethnic divergences (**Table 1**).

Trying to establish an initial validation of the noticed alleles effects, we include Chile, Portugal, and Peru (low-rate death indexes) and Romania, in four independent analysis (**Table S1 and Figure S1**). Portugal (**Figure S1A**), Peru (**Figure S1B**), and Chile (**Figure S1C)** cluster with other nations (Turkey and China, respectively) presenting low-rate death indexes. Romania grouped with Brazil (**Figure S1D)**, both countries have death per case rate of approximately 1: 15. Probably there are many other factors involved with the COVID-19 determination, but it seems correct and plausible to presume influences of the above-selected MHC alleles, which highlights the importance of generates populations data about MHC alleles and of use the available information.

HLA-A*02:01 as an example of universal protective alleles in human populations

An alternative hypothesis is that the T cell response governing SARS-CoV-2 recovering and clearance is through prevalent alleles in populations, as the HLA-A*02 supertype, which could be a good sign. Initially, it seems to be a contradictory reasoning, but in

light of the mortality rates in other SARS causative viruses (around 10% or above)[16], the current values may indicate a more effective worldwide cellular response. This idea is supported by many studies reporting HLA-A*02:01 as a pivotal allele in viral infections. A consultation on Immune Epitope Database (www.iedb.org) returned 849 references describing positive T cell response against viral peptides presented in the context of HLA-A*02:01 allele (accessed on May 26, 2020). Besides, 20 out of 34 immunogenic epitopes described for coronaviruses are restricted to HLA-A*02:01. In 2003 pandemics, lymphocytes from previously infected individuals were able to eliminate cells presenting SARS-CoV epitopes restricted to HLA-A*02:01 molecules[11,17] up to six years later from recovering[18], demonstrating the importance of this allele on viral clearance and T cell central memory. A conducted study with individuals from China and Hong Kong evidenced that more than half of the SARS recovered subjects were HLA-A*02:01 positive[19]. It is noteworthy, considering that the frequency of this allele is 0.1090 and 0.0620 in China and Hong Kong, respectively. The HLA-A*02:01 frequencies among the top 10 infected populations is high (median 23.27).

In our analysis, most of the recovered HLA-A*02:01 epitopes described in past epidemics seem to present conserved sequences compared to their equivalent in SARS-CoV-2 (**Table 2**), as described in other works [20,21]. Nevertheless, in case of discordant peptide sequences, the simple sequences comparison of SARS-CoV and SARS-CoV-2 may provide us little information about the impact of these alterations on the immunogenic potential on current putative viral targets. The analysis of structural and physicochemical features in the peptide-MHC (pMHC) surfaces that contact the T cell receptors, considers the combined molecular elements involved in T cell activation. Such structural investigation has already demonstrated its potential, explaining differential immunogenicity among epitopes from diverse viral strains or tumoral origins [22,23]. Since usually there is no available crystal for all specific targets complexed in HLA-A*02:01 allele, we construct all customized complexes through our reliable DockTope tool for pMHC modeling (http://tools.iedb.org/docktope/)(Rigo et al., 2015). This structural analysis gives us two alternative scenarios: the TCR interacting surfaces of both pHLA-A*02:01 complexes were similar **(Figure 2A)**, pointing out for preserved immunogenicity in SARS-CoV-2 targets; or the targets presented subtle physicochemical alterations in complexes harboring SARS-CoV-2 peptides, compared to former SARS-CoV. Amazingly, in this case, some of them turned into closely related surfaces presented by previously immunogenic epitopes, from non-related viruses **(Figure 2B)**. In the depicted example, the SARS-CoV-2 peptide SIIAYTMSL demonstrates structural convergence of physicochemical features with the

immunodominant epitope $M1_{58-66}$ GILGFVFTL, from the Influenza virus. Both epitopes share 2/9 amino acids, reinforcing the importance of structural investigation to prospect cross-reactive targets. In our analysis, to make it evident that the comparisons were not result from structural biases, we provide a small sample of unrelated models from our CrossTope database (www.crosstope.com), containing TCR interacting surfaces from HLA-A*02:01 structures presenting immunogenic epitopes **(Figure S2**). These first comparisons uncovered interesting scenarios. Firstly, even those peptides with sequence alterations in SARS-CoV-2, but without molecular modifications in the TCR interacting surfaces, remain good candidates to immunization strategy. The examples where SARS-CoV-2 peptides shown subtle alterations compared to their corresponding SARS-CoV targets could indicate a change in immunogenicity or even a complete loss of it. However, in this situation it resembles highly immunogenic epitopes from other viral organisms. This evidence emphasizes the need to investigate this new face of the immunogenic prism.

Searching for SARS-CoV-2 shared immunogenic fingerprints in alphacoronaviruses and other prevalent viruses in human populations

The observed molecular similarity between pMHCs complexes containing peptides from SARS-CoV-2 and Influenza viruses brings us toward another attractive hypothesis that refers to a universal previous cytotoxic response present in populations from all over the world, triggered by previous infections. The first suspects to investigate were past contact with HCoVs from the alphacoronavirus genus members (229E, OC43, and NL63). Epidemiological studies reported that around 15-30% of the common cold are caused by this group of pathogens[24]. Even considering that they are viruses with a zoonotic origin, we would expect many spillover events throughout the history of the human species, maintaining regular contact with our species[25]. A codon usage analysis, involving BCoV and HCoV-OC43, suggests that an ancestor coronavirus could be present even 200 kyr ago, in early humans[26]. Therefore, we would expect that this group of pathogens has also contributed to shaping our current immune system repertoire. Guided by this supposition, we compared the immunogenic SARS-CoV epitopes with 229E, OC43, and NL63 corresponding protein sequences, looking for shared elements involved in immunogenicity triggering. Such analysis presented a clear example where sequence comparison might be hiding shared patterns not detectable by single amino acid identity alignment. In **Table 2**, the identities ranged around 50%, a value usually not detected by regular sequence methods prospection.

Nevertheless, when we inspect pMHC structural models harboring peptides from SARS-CoV-1, SARS-CoV2, and alphacoronavirus in HLA-A*02:01 alleles, intriguing fingerprints arose. A similar electrostatic distribution and topography, on the TCR interacting surfaces from the pMHCs, can be observed among SARS-1, SARS-2, and coronaviruses members (229E and OC43) (**Figure 3A**). It is important to reinforce that both 229E and OC43 peptides were predicted as strong binders to HLA-A*02:01 (data not shown), strengthening their potential as actual triggers for SARS-CoV cross-reactivity. The peptides from the beta-CoVs are quite similar (**GLMWLSYF**L and **GLMWLSYF**V, for SARS-1 and 2, respectively). However, the peptides from 229E (LV**MW**VM**YF**A) and OC43 (II**MW**IV**YFV**) are not so conspicuously identical, evidencing the importance of the structural investigation. Furthermore, other peptides derived from alpha-CoV viruses presented a lower similarity in physicochemical features to immunogenic SARS-CoV-1 epitopes (data not shown). They are also potential targets to investigate. A work recently deposited in bioRxiv showed that 34% SARS-CoV-2 of seronegative healthy individuals presented S-reactive CD4 T cells. These cells react almost exclusively with the C-term epitopes region, characterized by higher homology with spike protein of human endemic common cold coronaviruses. Nevertheless, none of the putative cross-reactive epitopes are pointed-out, nor the structural basis hypothesized, but it reinforces our propositions[27]. Evidences of many CD4+/CD8+ cross responses against many SARS-CoV proteins in unexposed individuals were extensively described in Griffoni et al. [20], without the specific identification of sequence targets.

Taking into consideration the previous identification of a similar target from Influenza virus ($M1_{58-66}$ GILGFVFTL) with a SARS-CoV epitope, the next step was to scrutinize CrossTope T cell epitope databases ((*http://crosstope.com/*[28])
searching for structural fingerprints common to SARS-CoV-1 epitopes and unrelated viruses. As demonstrated before, given that SARS-CoV-1 and SARS-CoV-2 peptides are structurally related, the comparison with experimentally described epitopes from SARS-CoV-1 seems to be appropriate. The results from these comparisons were extraordinary. When we look for pHLA-A*02:01 structures, not only the previous example of $M1_{58-66}$ IAV epitope matched with SARS epitopes, but 13 out 20 CD8+ coronavirus epitopes recovered from Immune Epitope Database, has counterparts in targets from common circulating viruses, with respect to depicted molecular features. Three examples of SARS-CoV peptides presenting strikingly structural identity with viral epitopes are presented in **Figure 3B**. The matched targets are derived from viruses belonging to three different families (herpesviridae, poxviridae, and flaviviridae).

These would probably not be investigated targets in an approach using regular methods, given that no apparent identity is evidenced by any of these structurally related sequences.The remaining comparisons can be viewed in **Figure S3**. Importantly, when we consider these images correspondences, the comparisons with pMHCs from unrelated viruses were more conspicuous than those from alpha CoVs. It seems a paradox, given the natural expectation (considering its phylogenetics proximity) of a more intimate relation between alpha and beta CoVs.

Concluding Remarks

In COVID-19, the aetiological agent, SARS-CoV-2, is well established. Nevertheless, the same cannot be regarding all elements participating in severe acute respiratory syndrome (SARS) etiology. For this purpose, two aspects of cellular response were approached: the impact of MHC allelic constitution of populations on COVID-19 outcome and the structural analysis of immunogenic elements presented by SARS-CoV and putative SARS-CoV-2 T cell epitopes.

In the first section, a HLA allele frequency investigation was conducted on the most affected nations (considering COVID-19 cases number) to elucidate populations' genetic elements probably involved in lethality indexes. By only using four frequencies of prevalent alleles, we were able to cluster countries presenting similar death rates profiles concerning both closed and active cases. It seems to contribute with the determination of population response facing COVID-19, in addition to the political and medical interventions which are, in every country context, also extremely essential. It is necessary to emphasize that this approach prototype was just a primary prospection of how these components could operate in complex scenarios. May have countless more factors (and alleles, probably in haplotypes) involved with the mortality determination ratio. Nonetheless, it highlights the importance of a worldwide genotyping effort to clarify this landscape.

The epitope structural analysis and their relationship with other CoV and unrelated viral targets unveiled noteworthy observations. These pieces of evidence may open two avenues of investigation. The high degree of molecular conservation between SARS-CoV-1 epitopes and its corresponding sequences in SARS-CoV-2 allows its use in vaccine development to stimulate cross-reactive responses, covering distinct SARS-CoV-2 strains, for example. In this regard, animal reservoirs can be inspected looking

for beta CoVs with the potential to spill out of their natural hosts to humans. Peptides sequences from these putative HCoV pathogens could be structurally compared searching for cross-reactive T cell targets to be used in a virtual future occurrence of a new coronavirus spillover phenomenon. Such preventive strategy could abbreviate steps to develop immunotherapeutic methods, avoiding the emergence of new pandemics. The second line of the investigation resulted in an even more attractive hypothesis. Previous infections with alpha-CoVs and unrelated common viruses can be generating memory T cells against SARS-CoV-2, in a significant portion of the population. This pool of cells in different individuals is providing a universal immunogenic shield against SARS-CoV-2 and, probably, against other potential emergent viruses. This mechanism seems to be constructed by regular cross-reactive contacts. Moreover, the defense appears to be associated with prevalent alleles, which probably present peptides harboring fingerprints of immunogenicity shared by epitopes that regularly infect humans.

Acknowledgments

We thank to National Council for Scientific and Technological Development (CNPq) and National Council for the Improvement of Higher Education (CAPES) for their support.

Authors Contributions

E.C.A performed the structural analysis in cross-reactivity predictions. M.A.S.B recovered epitope and genome sequences of coronaviruses and performed sequence analysis. M.R.M conducted all the experiments regarding HLA frequencies and wrote the manuscript. G.F.V designed the study, coordinated the experiments and wrote the manuscript.

Declaration of Interests

The authors declare no competing interests.

Methods

Alleles search and data evaluation

To start the investigation, we picked the ten most affected countries in the number of COVID-19 cases, according to the dataset [3] *(recovered at May 26, 2020)*, plus China, which was the first pandemic epicenter. The alleles having a frequency equal to or higher than 10% (in at least one of the selected countries) were examined. Some HLAs were rejected due to its lack of information in some countries or considering its similar frequencies across nations, which could interfere in the future employment of the clusterization method.

The data were obtained from a classic freq search for HLA alleles in the The Allele Frequency Net Database (AFND)[12], selecting the locus and allele of all populations and all regions in the specified country (all countries included in the analysis). This provides a list of country populations, including information on their allele frequencies and sample sizes, which were applied to calculate the weighted average of each allele in each country. For example, to generate **Table 1**, five searches were performed (considering only the five alleles we maintain) for each of the eleven countries, totaling 55. The data for each search (for each allele) was compiled to calculate the weighted average, which are the frequency values observed in the table. A better description of the applied methodology can be visualized as a graphical scheme at **Figure S4.**

Hiararchical Clustering

In the second step, we import one dataset containing alleles information (weighted average of B*07, B*44, DRB1*03, and DRB1*04 frequencies in each country) to statistical software R aiming to assess the hierarchical clustering analysis (HCA) in its add-on package Pvclust [29]. HCA permits to classify several objects (countries) into some groups (clusters) according to similarities between them (allele's values). Pvclust calculates probability values (p-values) for each cluster using bootstrap resampling techniques. Two types of p-values are available: approximately unbiased (AU) p-value and bootstrap probability (BP) value. Multiscale bootstrap resampling is used for the calculation of AU p-value, which has superiority in bias over BP value calculated by the ordinary bootstrap resampling. Generally the AU p-value is a better estimator of the reliability of the clusters obtained. The bootstrap analysis was performed with the number of bootstrap replications being B = 10.000.

SARS-CoV Epitopes propections

We prospected in the IEDB database for experimental coronavirus epitopes, finding 20 immunogenic epitopes restricted to HLA-A*0201 allele. Those epitopes were derived of the N (nucleocapsid) protein, Surface (spike) protein, and Membrane glycoprotein. To verify if those epitopes have a counterpart in the SARS-CoV-2 proteome, we used a Needleman-Wunsch Global Align Nucleotide Sequences (BLAST) using the protein sequences of the 2002/2003 SARS virus and the SARS-CoV-2 from Wuhan. To predict binding affinity of discordant and alpha and unrelated viral epitopes we used NetMHCcons tool.

Alphacoronavirus epitope recovering

We selected viruses from the Alphacoronavirus family (HCoV-229E, HCoV-NL63, and HCoV-OC43), and checked if these strains possess epitopes that could generate similar surfaces as those observed in the SARS viruses. To this, we observed the protein sequences of different Alphacoronaviruses and searched for some amino acid identity with immunogenic targets described for SARS-CoV-1, aiming to verify if they present structural conservation throughout different coronaviruses.

Structural analysis

All selected epitopes were modeled in HLA-A*02:01 receptor by Docktope Tool. The 3D models were used as input in the Pymol software to calculate their electrostatic surfaces to verify how much the change in the amino acid impacted the overall charge disposition in those models. We compared the pMHC complexes of both viruses with several other complexes deposited in the Crosstope database, which harbors several previously described epitope sequences of immunogenic targets from the IEDB.

Structural Hierarchical Clustering

All images of SARS-CoV and CrossTope HLA-A*02:01 epitopes were used to perform the comparisons. We utilized ImageJ to extract the Red-Blue-Green (RGB) values corresponding to the electrostatic surface regions that contact T cell receptor. These regions were inferred by contact calculation of ternary crystals comprising p:HLA-A*02:01:TCR. The regions present colors that can be more positive, neutral or negative

in regards to the electric charges and this color information was converted in numeric values of mean, mode and standard deviation to feed the hierarchical clusterization program pvclust, ran by the R Studio to perform this pairwise comparison and group the most similar surfaces together.

Figure Legends

Figure 1. ***Hierarchical Clustering Analysis from the most affected countries in number of COVID-19 cases reported, based on its HLA frequencies.*** The image represents the HCA result for a clusterization concerning the frequencies of the following alleles: HLA-B*07, HLA-B*44, HLA-DRB1*03, and HLA-DRB1*04. Four clusters comprising countries with similar death rates are showed. High-rates countries as Spain, Italy, and France grouped. In the same way, low-rate nations like Germany and Russia fall in the same branch. Its required frequencies were obtained from Alleles Frequency Net Database (http://www.allelefrequencies.net/) and are also compiled in Table 1. Values at branches are AU p-values (left), BP values (right), and cluster labels (bottom). Clusters with AU ≥ 95 are indicated by the rectangles. Death per case ratios of each country are expressed as 1: X (one death in X cases). The calculated values for "X" are depicted in HCA.

Figure 2. ***Structural analysis of SARS-CoV HLA-A*02:01 epitopes.*** A sequence demonstrating the TCR-interacting surface is depicted above. From left to right we can observe 1) the MHCs in the cell membrane; 2) a side view of pMHC with alpha chain in pale blue, B2-microglobulin in pale purple and the peptide in green; 3) a top view of pMHC with the peptide region highlighted in green and 4) the pMHC surface with electrostatic potential distribution and topography, the main elements involved in immune response stimulation. Negative values are represented in red and positive ones in blue. A) SARS-CoV peptide sequences showing high similarity in sequence and structural features. B) SARS-CoV SI**V**AYTMSL/SI**I**AYTMSL sequences presenting subtle differences. An HCA analysis with the IAV **G**I**LGFVFT**L epitope demonstrates that it is even more similar to SIIAYTMSL SARS-CoV-2 peptide sequence.

Figure 3. ***Comparison of electrostatic surface distribution and topography of SARS-CoV peptides with alphacoronaviruses and other prevalent viruses in***

human populations. pMHCs models were compared in terms of topography and electrostatic distribution. In A, two beta (SARS-CoV) and two alpha (229E and OC43) coronaviruses show similar electrostatic distribution and topography despite its sequence divergences. In B, a panel of three SARS-CoV peptides is compared to pMHCs containing viral epitopes from members of other families. An unexpected shared structural similarity arises, disregarding their lack of sequence identity and phylogenetic relationship. Electrostatic calculations are represented as negative (red) and positive (blue) charges. Sequence identity is depicted as red letters in the peptide sequences.

Figure S1. ***Additional Hierarchical Clustering Analysis based on countries HLA frequencies.*** The figure depicts the inclusion of four countries: Portugal (A), Peru (B), Chile (C), and Romania (D) in separate HCAs. It illustrates the results for clusterization concerning the frequencies of the following alleles: HLA-B*07, HLA-B*44, HLA-DRB1*03, and HLA-DRB1*04. Values at branches are AU p-values (left), BP values (right), and cluster labels (bottom). Clusters with AU ≥ 95 are indicated by the rectangles. Death per case ratios of each country are expressed as 1: X (one death in X cases).

Figure S2. ***TCR interacting surfaces from pMHC immunogenic structures.*** Nine pMHC from HLA-A*02:01 bound to experimental epitopes are represented. This panel of images was provided to evidence molecular signs diversity that can be found in these targets.

Figure S3. ***Remaining comparison of SARS-CoV epitopes and immunogenic epitopes from CrossTope Database.*** Shared fingerprints are also observed in the above-exemplified pairs of structures indicating a general structural signature.

Figure S4. ***Resume the method applied in "The HLA frequency relation and its apparent effect in COVID-19 immune responses" section*****.** The weighted average (WA^1) of each allele´s frequency was obtained for the selected countries at *the Allele frequency net database (AFND*). It was used in a hierarchical clustering analysis (HCA) through the pvclust package in the R-Studio software.

References:

1. Wu F, Zhao S, Yu B, et al. A new coronavirus associated with human respiratory disease in China. *Nature* 2020; **579**(7798): 265-9.
2. Zhou P, Yang XL, Wang XG, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* 2020; **579**(7798): 270-3.
3. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect Dis* 2020.
4. Arango MT, Perricone C, Kivity S, et al. HLA-DRB1 the notorious gene in the mosaic of autoimmunity. *Immunol Res* 2017; **65**(1): 82-98.
5. Mösch A, Raffegerst S, Weis M, Schendel DJ, Frishman D. Machine Learning for Cancer Immunotherapies Based on Epitope Recognition by T Cell Receptors. *Frontiers in Genetics* 2019; **10**(1141).
6. Trowsdale J, Knight JC. Major histocompatibility complex genomics and human disease. *Annu Rev Genomics Hum Genet* 2013; **14**: 301-23.
7. Xi YSaY. HLA and Associated Important Diseases. 2014.
8. Dronamraju K. The Compatibility Gene: How Our Bodies Fight Disease, Attract Others, and Define Our Selves by Daniel M. Davis. *The Quarterly Review of Biology* 2015; **90**(3): 343-.
9. Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. The proximal origin of SARS-CoV-2. *Nature Medicine* 2020; **26**(4): 450-2.
10. Hoffmann M, Kleine-Weber H, Schroeder S, et al. SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* 2020.
11. Chen YM, Liang SY, Shih YP, et al. Epidemiological and genetic correlates of severe acute respiratory syndrome coronavirus infection in the hospital with the highest nosocomial infection rate in Taiwan in 2003. *J Clin Microbiol* 2006; **44**(2): 359-65.
12. FF G-G, A M, EJMD S, et al. Allele Frequency Net Database (AFND) 2020 Update: Gold-Standard Data Classification, Open Access Genotype Data and New Query Tools. *Nucleic acids research* 2020; **48**(D1).
13. Lo AW, Tang NL, To KF. How the SARS coronavirus causes disease: host or organism? *J Pathol* 2006; **208**(2): 142-51.
14. Ng MHL, Department of Anatomical and Cellular Pathology HKS, China, Lau K-M, et al. Association of Human-Leukocyte-Antigen Class I (B*0703) and Class II (DRB1*0301) Genotypes with Susceptibility and Resistance to the Development of Severe Acute Respiratory Syndrome. *The Journal of Infectious Diseases* 2020; **190**(3): 515-8.
15. Frias M, Rivero-Juárez A, Rodriguez-Cano D, et al. HLA-B, HLA-C and KIR improve the predictive value of IFNL3 for Hepatitis C spontaneous clearance. *Scientific Reports* 2018; **8**(1): 1-7.
16. Janice Oh HL, Ken-En Gan S, Bertoletti A, Tan YJ. Understanding the T cell immune response in SARS coronavirus infection. *Emerg Microbes Infect* 2012; **1**(9): e23.
17. Tsao YP, Lin JY, Jan JT, et al. HLA-A*0201 T-cell epitopes in severe acute respiratory syndrome (SARS) coronavirus nucleocapsid and spike proteins. *Biochem Biophys Res Commun* 2006; **344**(1): 63-71.
18. Channappanavar R, Fett C, Zhao J, Meyerholz DK, Perlman S. Virus-specific memory CD8 T cells provide substantial protection from lethal severe acute respiratory syndrome coronavirus infection. *J Virol* 2014; **88**(19): 11034-44.
19. Wang Y-D, Sin W-YF, Xu G-B, et al. T-Cell Epitopes in Severe Acute Respiratory Syndrome (SARS) Coronavirus Spike Protein Elicit a Specific T-Cell Immune Response in Patients Who Recover from SARS. 2004.
20. 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**(4): 671-80.e2.

21. Nguyen A, David JK, Maden SK, et al. Human leukocyte antigen susceptibility map for SARS-CoV-2. *Journal of Virology* 2020: JVI.00510-20.
22. Antunes DA, Rigo MM, Freitas MV, et al. Interpreting T-Cell Cross-reactivity through Structure: Implications for TCR-Based Cancer Immunotherapy. *Frontiers in Immunology* 2017; **8**(1210).
23. Mendes MF, Antunes DA, Rigo MM, Sinigaglia M, Vieira GF. Improved structural method for T-cell cross-reactivity prediction. *Mol Immunol* 2015; **67**(2 Pt B): 303-10.
24. Mesel-Lemoine M, Millet J, Vidalain P-O, et al. A Human Coronavirus Responsible for the Common Cold Massively Kills Dendritic Cells but Not Monocytes. 2012.
25. Ye ZW, Yuan S, Yuen KS, Fung SY, Chan CP, Jin DY. Zoonotic origins of human coronaviruses. *Int J Biol Sci* 2020; **16**(10): 1686-97.
26. Brandão PE, Universidade de São Paulo SP, Brazil. Could human coronavirus OC43 have co-evolved with early humans? *Genetics and Molecular Biology* 2018; **41**(3): 692-8.
27. Braun J, Loyal L, Frentsch M, et al. Presence of SARS-CoV-2 reactive T cells in COVID-19 patients and healthy donors. 2020.
28. Sinigaglia M, Antunes DA, Rigo MM, Chies JA, Vieira GF. CrossTope: a curate repository of 3D structures of immunogenic peptide: MHC complexes. *Database (Oxford)* 2013; **2013**: bat002.
29. Suzuki R, Shimodaira H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* 2006; **22**(12): 1540-2.

Table 1. Alleles frequency information and data of the most affected countries considering the number of COVID-19 cases.

Country		United States	Brazil	Russia	Spain	United Kingdom	Italy	France	Germany	Turkey	Iran	China [E]
cases		1.666.828	347.398	335.882	282.370	257.154	229.327	182.219	179.986	155.686	133.521	82.971
death		98.683	22.013	3.388	28.678	36.675	32.735	2.8332	8.366	4.308	7.359	4.634
active cases		1.121.231	182.798	224.558	56.734	N/A	57.752	85.590	11.720	33.776	22.090	79
closed cases [A]		545.597	164.600	111.324	225.636	N/A	171.575	96.629	168.266	121.910	111.431	82.892
population		330.769.370	212.376.810	145.927.122	46.752.654	67.843.268	60472166	65.256.433	212.376.810	84.277.597	83.859.705	1.439.323.776
death:cases [B]		16,82	15,002749	106,79	10,04	7,08	7,06	6,60	21,89	36,10	17,36	17,90
death: closed cases [B]		5,53	7,4773997	32,86	7,87	N/A	5,24	3,41	20,11	28,30	15,14	17,89
death:population		3351,88	9647,80	43071,76	1630,26	1849,85	1847,32	2303,27	25385,70	19563,05	11395,53	310600,73
Alleles Frequency [C]	*HLA-A*02*	0.2016	0.2593	0.2843	0.2439	0.2550	0.2539	0.2710	0.2665	0.2525	0.1713	0.2836
	*HLA-B*07*	0.1144	0.0691	0.0947	0.0900	0.1215	0.0570	0.1035	0.1311	0.0428	0.0405	0.0277
	*HLA-B*44*	0.1292	0.1081	0.0908	0.1634	0.1265	0.0916	0.1552	0.1270	0.1110	0.0413	0.0372
	*DRB1*03*	0.1153	0.0973	0.0808	0.1234	0.1490	0.0995	0.1106	0.1055	0.0932	0.0867	0.0464
	*DRB1*04*	0.1625	0.1254	0.1109	0.1201	0.1911	0.0800	0.1355	0.1316	0.1362	0.1054	0.1206
Population size reported [D]	*HLA-A*02*	2.833.193	2.854.388	11.521	6.518	7.987	165.944	15.429	81.446	512	16.743	72,136
	*HLA-B*07*	12.370	2.854.075	10.162	2.025	7.987	164.791	15.429	11.581	370	16.519	65.320
	*HLA-B*44*	12.370	2.854.075	10.162	2.025	7.987	164.791	15.429	11.581	370	16.519	65.433
	*DRB1*03*	11.988	2.853.876	10.563	10.679	980	171.535	15.397	11.407	528	16.455	40.114
	*DRB1*04*	11.988	2.853.876	10.240	4.274	980	171.588	15.397	11.407	528	16.513	66.380

The data about reported cases and deaths was recovered from COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (Dong et al., 2020)- *on May 26, 2020.*

[A] Closed cases number was calculated excluding the active ones from total cases.

[B] Death per case ratios of each country are expressed as 1: X (one death in X cases) and the X values were specified in the table.

[C] The allele's frequency for each HLA was obtained as a weighted average (considering the sample size) of all the populations reported in *the allele frequency net database* (http://www.allelefrequencies.net/) for each country.

[D] Population size reported is the sum of the sample sizes included in the present study.

[E] China was added due to its crucial role as the first pandemic epicenter of COVID-19.

Table 2. Recovered HLA-A*02:01 epitopes from SARS-CoV, SARS-Cov-2 and other alphacoronaviruses

IEDB - experimental positive epitopes from coronaviruses [A]					Peptides in other coronaviruses [B]			
Epitope ID	Description	Antigen Name	Organism Name	Experimental Assays (Positive/All)	SARS-CoV-2	HCoV-229E	HCoV-NL63	HCoV-OC43
2801	ALNTLVKQL	S protein	SARS-related coronavirus	2/2		**S**LN**H**L**TS**QL	ALN**H**L**TS**QL	ALN**N**L**LQ**QL
2802	ALNTPKDHI	Nucleoprotein	SARS-related coronavirus	2/2		**RVTV**PKDH**P**		**DV**NTP**ADIV**
16156	FIAGLIAIV	Spike glycoprotein precursor	SARS-related coronavirus	2/2				FI**NGIFAKV**
21347	GMSRIGMEV	Nucleoprotein	SARS-related coronavirus	10/10				
27182	ILLNKHIDA	Nucleoprotein	SARS-related coronavirus	1/3				
27241	ILPDPLKPT	Spike glycoprotein precursor	SARS-related coronavirus	2/2	ILPDP**S**KP**S**			
34851	LALLLLDRL	Nucleoprotein	SARS-related coronavirus	4/4				
36724	LITGRLQSL	Spike glycoprotein precursor	SARS-related coronavirus	5/8		LITGRL**AA**L	LITGRL**AA**L	LI**N**GRL**TA**L
37473	LLLDRLNQL	Nucleoprotein	SARS-related coronavirus	11/12				
38881	LQLPQGTTL	Nucleoprotein	SARS-related coronavirus	4/6		**Q**KLP**N**G**V**T**V**		**GTV**LPQG**YY**
44814	NLNESLIDL	S protein	SARS-related coronavirus	3/4		NIN**ST**L**V**DL		**V**LN**HSY**I**N**L
54690	RLNQLESKV	Nucleoprotein	SARS-related coronavirus	9/11	RLNQLESK**M**			
58730	SIVAYTMSL	S protein	SARS-related coronavirus	3/3	SI**I**AYTMSL			
69657	VLNDILSRL	S protein	SARS-related coronavirus	3/4			**ET**ND**VSSM**L	**SLQE**ILSRL
71663	VVFLHVTYV	Spike glycoprotein precursor	SARS-related coronavirus	5/5				**LY**FIH**FN**YV
125100	ILLNKHID	Nucleoprotein	SARS-related coronavirus	1/1				
21041	GLMWLSYFV	Membrane glycoprotein	SARS coronavirus TJF	7 / 8	GLMWLSYF**I**	**LV**MW**VM**YF**A**	**LCL**W**VM**YFV	**II**MW**IV**YFV
64710	TLACFVLAAV	Membrane glycoprotein	SARS coronavirus TJF	5 / 5			**VLALSIFDCF**V	
32069	KLPDDFMGCV	Spike glycoprotein precursor	SARS coronavirus BJ01	2 / 2	KLPDDF**T**GCV			**YSF**D**SYL**GCV
54680	RLNEVAKNL	Spike glycoprotein precursor	SARS coronavirus BJ01	1 / 1		RLN**Y**VA**LQT**	**TLQE**FA**Q**NL	RL**QEAIKV**L

[A] The selection and data of the epitopes were accessed on Immune Epitope Database and Analysis Resource (IEDB - https://www.iedb.org/ - *acessed on May 26, 2020*).

[B] In the comparison among other coronoviruses peptides, the blank cells in the columns mean no amino acid difference among SARS-CoV-1 epitope and its equivalent in the viruses, the occurred discrepancies were **bolded.**

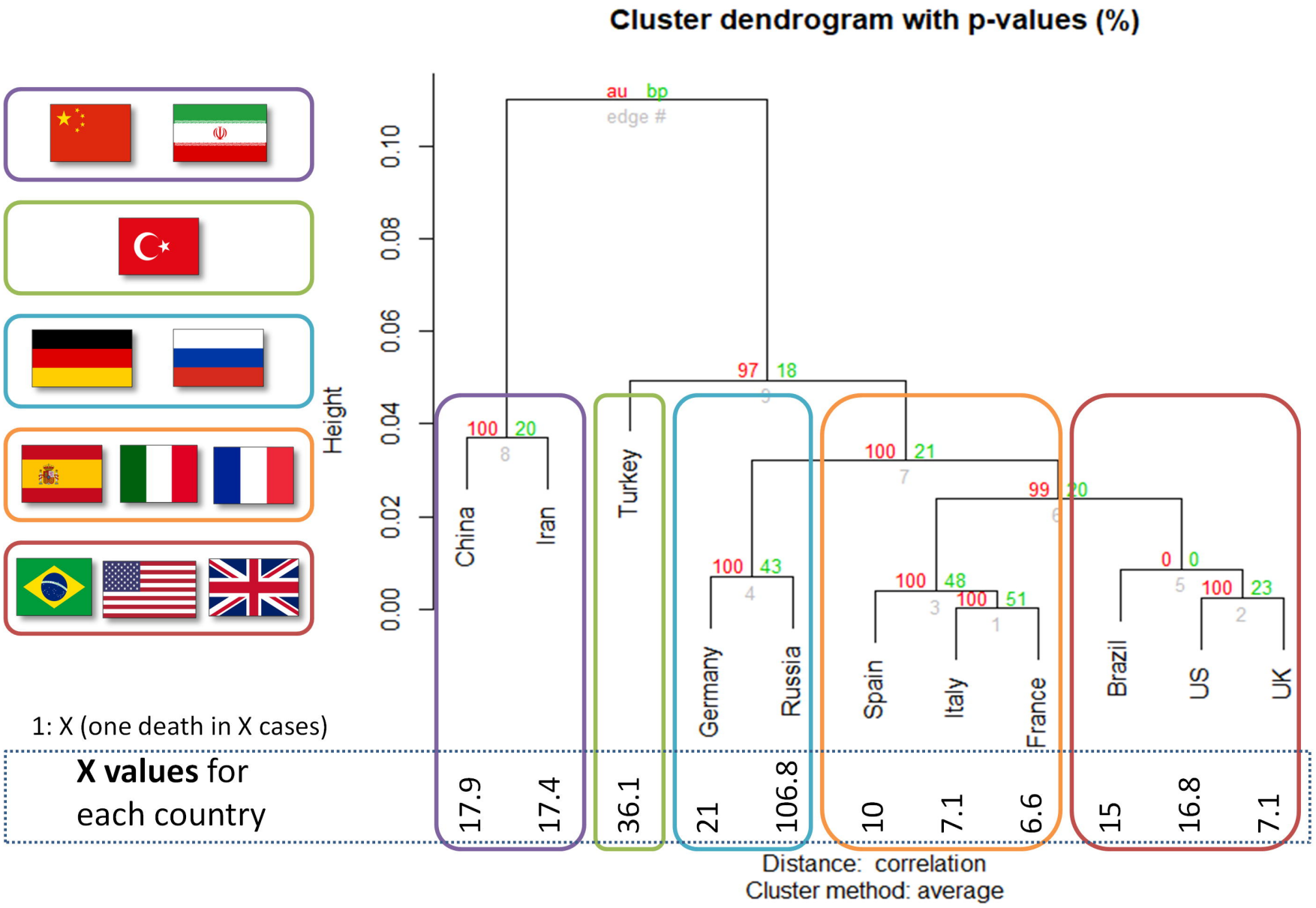
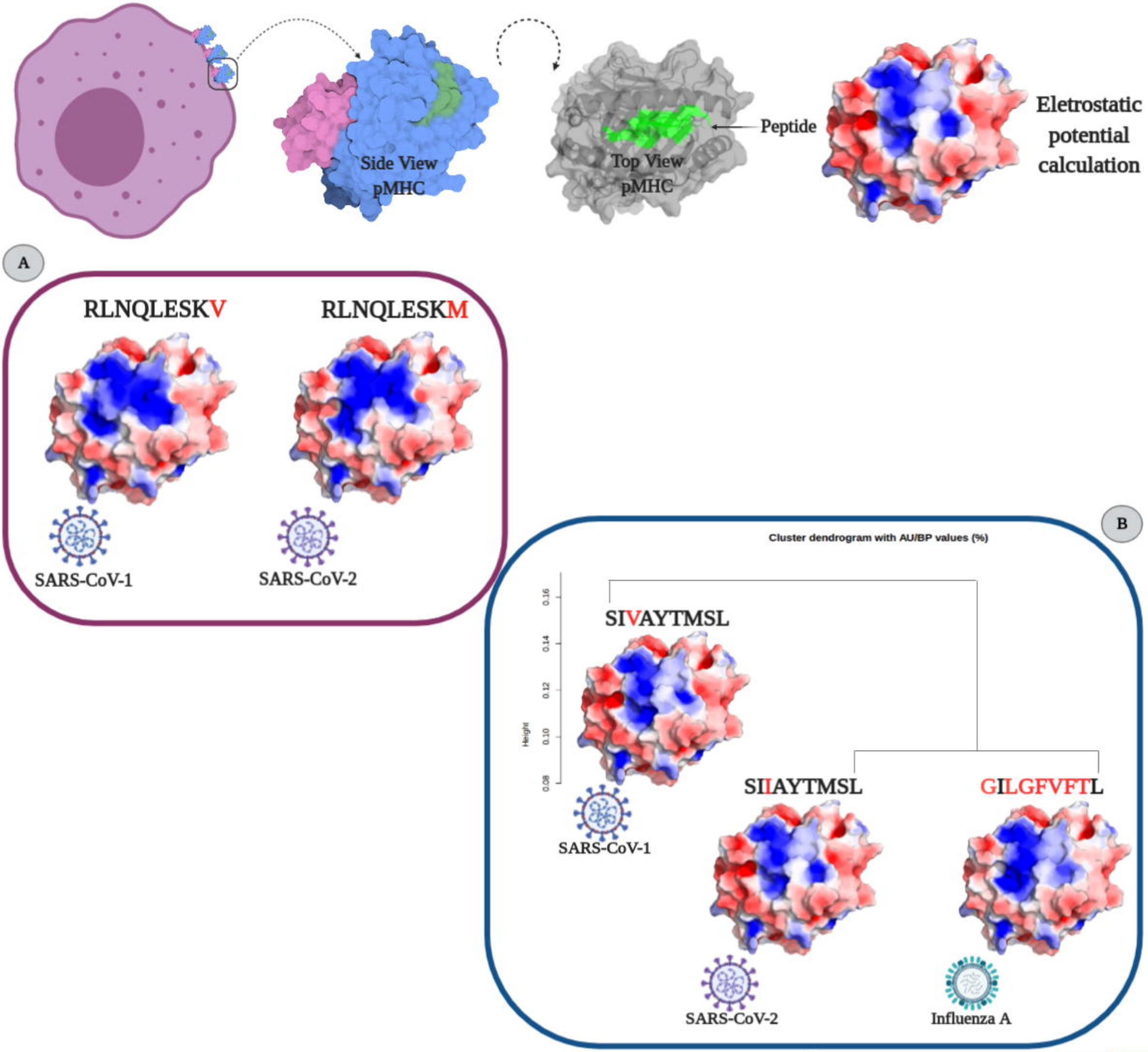

Cluster dendrogram with p-values (%)
au bp
edge #
Height
0.00
0.02
0.04
0.06
0.08
0.10
97 18
9
100 20
8
100 21
7
99 20
6
100 43
4
100 48
3
100 51
1
0 0
5
100 23
2
China
Iran
Turkey
Germany
Russia
Spain
Italy
France
Brazil
US
UK
1: X (one death in X cases)
X values for each country
17.9
17.4
36.1
21
106.8
10
7.1
6.6
15
16.8
7.1
Distance: correlation
Cluster method: average

Figure 1.


Side View
pMHC
Top View
pMHC
Peptide
Eletrostatic
potential
calculation
A
RLNQLESKV
RLNQLESKM
SARS-CoV-1
SARS-CoV-2
B
Cluster dendrogram with AU/BP values (%)
Height
0.16
0.14
0.12
0.10
0.08
SIVAYTMSL
SIIAYTMSL
GILGFVFTL
SARS-CoV-1
SARS-CoV-2
Influenza A

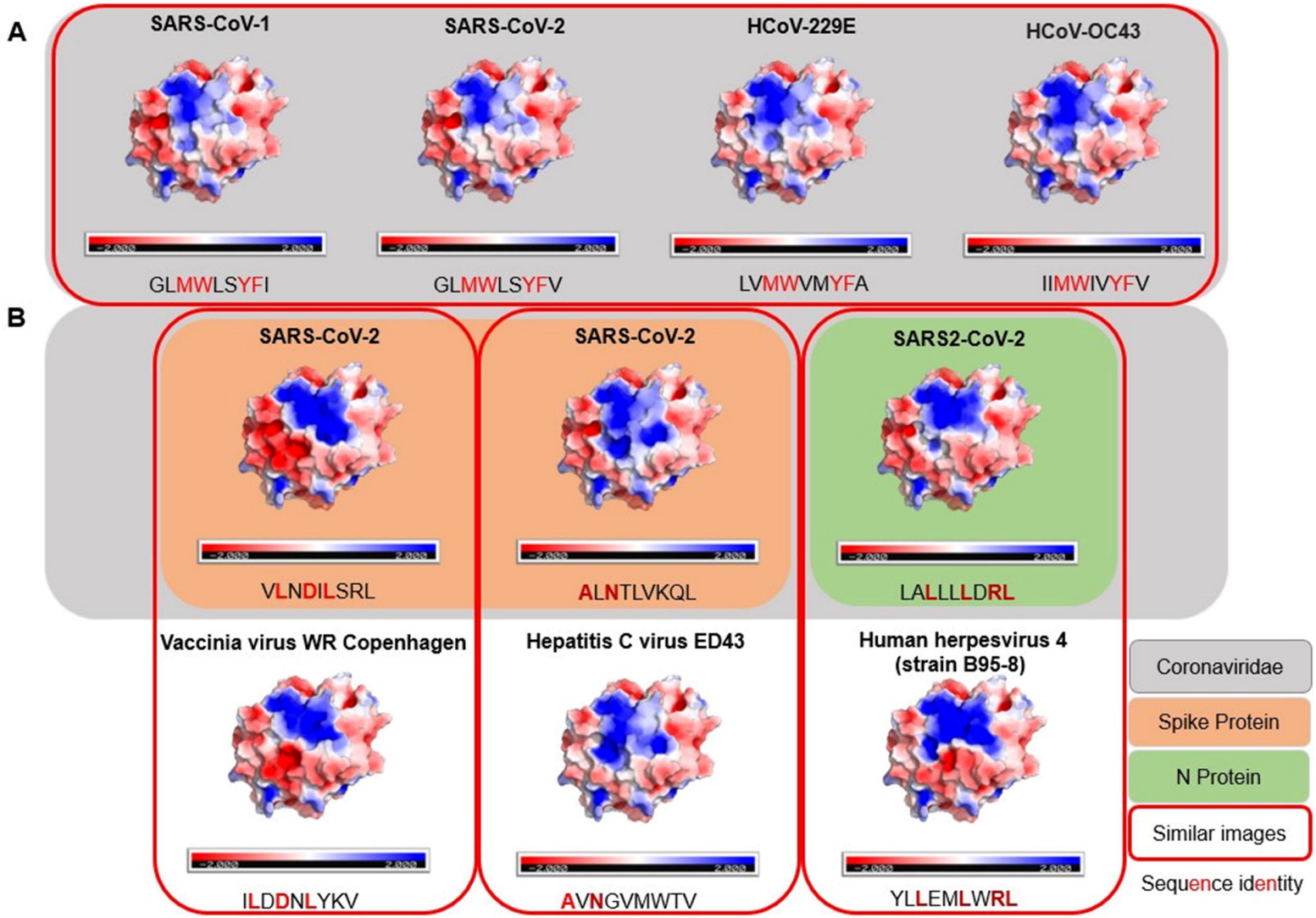

A
SARS-CoV-1
SARS-CoV-2
HCoV-229E
HCoV-OC43
GLMWLSYFI
GLMWLSYFV
LVMWVMYFA
IIMWIVYFV
B
SARS-CoV-2
SARS-CoV-2
SARS2-CoV-2
VLNDILSRL
ALNTLVKQL
LALLLLDRL
Vaccinia virus WR Copenhagen
Hepatitis C virus ED43
Human herpesvirus 4 (strain B95-8)
ILDDNLYKV
AVNGVMWTV
YLLEMLWRL
Coronaviridae
Spike Protein
N Protein
Similar images
Sequence identity

Figure 3.