scientific reports

Check for updates

OPEN

Considering epitopes conservity in targeting SARS-CoV-2 mutations in variants: a novel immunoinformatics approach to vaccine design

Mohammad Aref Bagherzadeh[1], Mohammad Izadi[1], Kazem Baesi[2], Mirza Ali Mofazzal Jahromi[3,4,5]✉ & Majid Pirestani[6]✉

Severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) has gained mutations at an alarming rate in the past years. Developing mutations can increase the virus's pathogenicity and virulence; reduce the efficacy of vaccines, antibodies neutralization, and even challenge adaptive immunity. So, it is essential to identify conserved epitopes (with fewer mutations) in different variants with appropriate antigenicity to target the variants by an appropriate vaccine design. Yet as, 3369 SARS-CoV-2 genomes were collected from global initiative on sharing avian flu data. Then, mutations in the immunodominant regions (IDRs), immune epitope database (IEDB) epitopes, and also predicted epitopes were calculated. In the following, epitopes conservity score against the total number of events (mutations) and the number of mutated sites in each epitope was weighted by Shannon entropy and then calculated by the Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS). Based on the TOPSIS conservity score and antigenicity score, the epitopes were plotted. The result demonstrates that almost all epitopes and IDRs with various lengths have gained different numbers of mutations in dissimilar sites. Herein, our two-step calculation for conservity recommends only 8 IDRs, 14 IEDB epitopes, and 10 predicted epitopes among all epitopes. The selected ones have higher conservity and higher immunogenicity. This method is an open-source multi-criteria decision-making platform, which provides a scientific approach to selecting epitopes with appropriate conservity and immunogenicity; against ever-changing viruses.

About 2 years have passed since the first case of new coronavirus disease-19 (COVID-19) has been identified in Wuhan, China. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the notorious claimant behind COVID-19, has spread all around the world, soon after confirmation of the first cases in Wuhan, which resulted in a global health problem till now[1–3]. On 11 March 2020, the World Health Organization (WHO) announced this worldwide health problem as a new pandemic. Although the pandemic is quite young, now our knowledge about the etiology, pathogenesis, and treatment of COVID-19 has improved[4–6]. SARS-CoV-2, the new member of the Coronaviridae family (Betacoronavirus), has improved its ability to survive and preserve its generation[5]. Compared to other Coronaviridae members, higher virulence besides significant pathogenicity may justify this unprecedented pandemic[7].

According to the latest report of WHO on 2/13/2022, so far 404,910,528 people in the world have been infected with the SARS-CoV-2, and 5,783,776 have died from COVID-19[1]. COVID-19 clinical presentations range from non-symptomatic infection to severe acute respiratory distress syndrome (ARDS) and even death, but it is best known for its flu-like symptoms (cough, low-grade fever, chills, myalgia, sweating, and fatigue), as

[1]Student Research Committee, Jahrom University of Medical Sciences, Jahrom, Iran. [2]Department of Virology, Pasteur Institute of Iran, Tehran, Iran. [3]Zoonoses Research Center, Jahrom University of Medical Sciences, Jahrom, Iran. [4]Department of Immunology, School of Medicine, Jahrom University of Medical Sciences, Jahrom, Iran. [5]Department of Advanced Medical Sciences & Technologies, School of Medicine, Jahrom University of Medical Sciences, Jahrom, Iran. [6]Parasitology and Entomology Department, Faculty of Medical Sciences, Tarbiat Modares University, Tehran, Iran. ✉email: alimofazzal@yahoo.com; alimofazzal@jums.ac.ir; pirestani@modares.ac.ir

well as pneumonia[8,9]. By passing time, observations in clinical practice declared other organ involvement like the gastrointestinal and central nervous system in COVID-19 that just complicated the issue[10,11]. The impact of COVID-19 on individuals and also public health although is drastic, it is far to be understood. So far, several vaccines with different platforms have been approved by Food and Drug Administration (FDA) or local authorities[12]. Vaccination and adhering to safety protocols (reducing human-to-human transmission through monitoring, social distancing, and hygienic principles) are promising possible options[13,14].

Therefore, the vaccine-mediated immunity (vaccination) of the population seems to be even more important in the current situation of the COVID-19 pandemic, where SARS-CoV-2 spreads rapidly and there are therapeutic challenges in the management of patients suffering from COVID-19. In another hand, the naturally acquired immunity (NAI) could not prevent the spread of the SARS-CoV-2[15]. So, the importance of developing vaccines to target SARS-CoV-2 during this particular circumstance is undeniable[16]. Many noteworthy considerable efforts have been made in this field around the world. Till now, different approaches have been administrated for COVID-19 vaccine design, such as mRNA-based vaccines (Comirnaty (BNT162b2 or Pfizer, BioNTech) and Moderna), adenovirus vaccine (AstraZeneca), recombinant adenovirus vaccine (Sputnik V), non-replicating viral vector (Janssen (JNJ)), inactivated vaccine (CoronaVac), peptide vaccine (EpiVacCorona), nanoparticle vaccine (NVX-CoV2373), and DNA vaccine (plasmid) (ZyCoV-D). Meanwhile, some have international approval for global use, some have limited licenses in some countries, and a large number of vaccines are under the clinical trials and development phase (more information on vaccine tracker-NY)[17]. According to WHO, as of 2/13/2021, 10,095,615,243 vaccine doses have been administrated in the world[1].

As vaccines are being developed in the laboratories and vaccination is under process, the novel coronavirus is cleverly changing and mutating in nature. So far, lots of mutations have been identified in the SARS-CoV-2 genome, which resulted in new specific variants. Potentially developing mutations can result in structural changes in key proteins involved in the pathogenesis and spread of the virus. The frequent and rapid genetic mutations and the consequent changes in SARS-CoV-2 implicate the necessity of a method for naming new variants. As expected, new mutations in the target sequences of vaccines could decrease vaccines efficacy against new variants, besides increasing pathogenicity, transmission, and virulence, a point that has also been shown in studies[18–20]. According to the United States centers for disease control and prevention (CDC), variant of interest (VOI) stands for a variant with specific genetic markers, which potentiate changes to receptor binding, decreased neutralization by antibodies generated against vaccination, and some other clinical issues in diagnosis and treatment. In another hand, variant of high consequence (VOHC) indicates a variant with significantly reduced effectiveness of prevention measures or medical countermeasures relative to previously circulating variants, in literature with acceptable evidence[21].

The CDC illustrated that B.1.617.2 and AY lineages (Delta), B.1.1.529 (Omicron), are categorized as current variant of concern (VOC), which means they are associated with reduced neutralizing antibodies titer in the convalescent and post-vaccination sera besides increased transmissibility and more severe disease[22–24].

It shows regardless of various approaches to vaccine design; the vaccine target needs to be selected precisely. As SARS-CoV-2 evolves spontaneously, two major points should be considered in selected parts' properties; first, the selected part must have an acceptable antigenicity to induce the significant immune response, second, it must be conserved during mutations in order to obtain proper coverage against different present variants. Also, it helps to maintain its efficacy against new variants. But on the other side of the coin, it might be more tremendous. Deployment of data and tools to understand the manner of SARS-CoV-2 mutations and simulation of the immune response against new variants seem to be a reliable way to deal with this issue[25].

On the other hand, different bioinformatics approaches have been utilized for finding potential drugs[26–28], designing vaccines[29,30], and finding the concept behind COVID-19 pathogenesis[31,32]. Many complex biochemical processes that have to be spent in the laboratory with a lot of time and money can be modeled and predicted[33,34]. For example, finding herbal drugs' interaction with SARS-CoV-2 proteins[33,35,36], predicting immunogen and antigen epitopes from SARS-CoV-2 proteins[37,38], and finding new potential pathways in COVID-19 progression and pathogenicity[32,39] and etcetera are all available through different bioinformatics techniques.

In this study, first, we demonstrated how investigated immune-dominant parts of SARS-CoV-2 proteins have mutated significantly. Furthermore, we presented evidence of alterations the spike (S), membrane (M), nucleocapsid (N), and envelope (E) proteins of SARS-CoV-2, which are discussed in peer reviews from the immune epitope database (IEDB)[40–42]. Because of the necessity of finding the appropriate target for vaccines, we wondered how an algorithm could propose an epitope with significant antigenicity besides being adequately conserved.

Here, we predicted B cell, helper T lymphocyte (HTL), and cytotoxic T lymphocyte (CTL) epitopes and then aligned them to their reference sequence to find mutations in each epitope. Then we scored their events (mutations) with the Shannon entropy and Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) method to quantify their conserved manner. In the setting of this algorithm, the final output still included dozens of different types of epitopes from different proteins making the decision hard to select the best epitope. To select ones with superiority, epitopes were illustrated in a plot of antigenicity score–conserved manner, and then epitopes were undergone illustrated two-dimensional comparison, which allows facile and confident selection of epitopes among a large volume of data. In this logical analysis, we tried to clarify the importance of the conserved property of epitopes along with the immunogenicity of the selected parts of SARS-CoV-2 proteins for the vaccine design (study design—sup. Diagram 1).

Herein, we provided a scientific and practical approach selecting epitopes with appropriate conserved and immunogenicity properties (whether in predicted epitopes or real ones from databases or peer reviews), with the hope that this protocol will aid in the development of methods to design vaccines against ever-changing viruses such as SARS-CoV-2, human immunodeficiency virus (HIV) and so on.

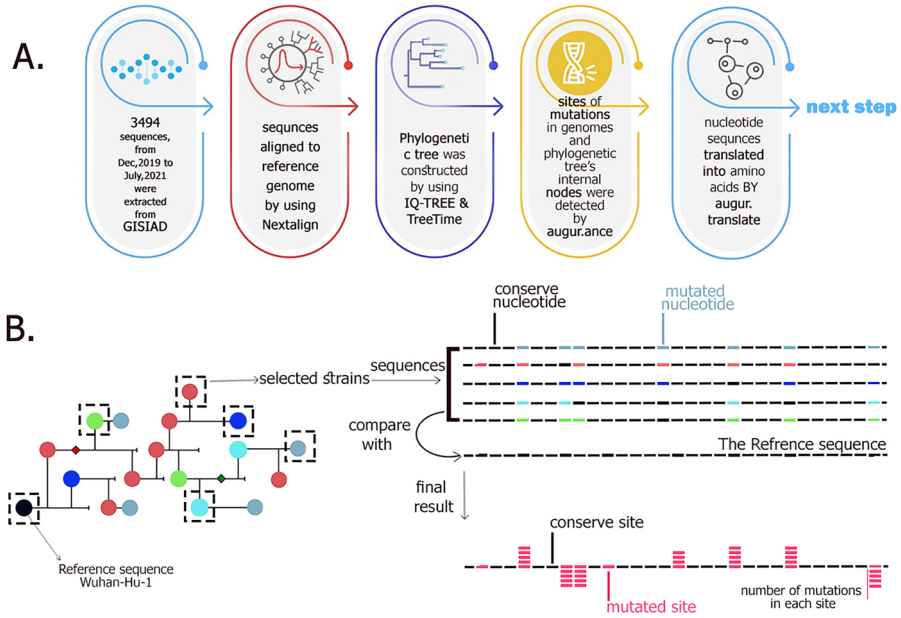

Figure 1. Schematic illustration of 5 steps to find the site of mutations and number of mutations. As illustrated in panel (**A**) number of mutations in each site of the SARS-CoV-2 genome was identified through 5 steps for the next step (finding mutations in different sites of epitopes and IDRs and quantifying their conservity). In panel (**B**), the identification of mutations and sites with comparing with the reference sequence (3rd and 4th steps) is illustrated schematically. In this example we have chosen 5 new variant sequences and compared them with the reference sequence in order to find the site of mutations and number of events in SARS-CoV-2 genome.

Results

Background. Genetic mutations elsewhere in the SARS-CoV-2 genome resulted in sophisticated conditions both in the virus virulence and pathogenicity and also in the host's immune system response by repeated infections with other variants that rechallenge adaptive immunity response[20,43]. Perhaps, the most important issue in dealing with the coronavirus vaccine design is its genetic variation[44–46]. Not only does it affects the vaccine design, but also it challenged the immune system response and treatment[46,47]. This issue, typicality is not a new concern, as we have seen this challenge in other viruses (such as, Influenza virus, Herpes, Zika, other coronaviridea members, and etc.)[48–50]. The most prominent example of these is HIV, where all these problems (in the immune system response against the virus, clinical presentation and management, and also vaccine design) are perspicuous[51,52]. As previously predicted, this problem grew rapidly with the introduction of VOC (B1.1.7, 501Y. V2, etc.), which prompted the WHO to quickly propose a system for naming new variants of the virus. We could be optimistic that by developing a strategy to deal with RNA virus mutations, the main burden against these viruses will be removed.

Finding mutations in SARS-CoV-2 variants. In order to detect mutations in the SARS-CoV-2 genome, 3369 sequences, from Dec 2019 to July 2021 were extracted from the global initiative on sharing avian flu data (GISAID) database (Fig. 1). Then, the sequences were aligned to the reference genome (hCoV-19/Wuhan/Hu-1/2019 (NC_045512.2)) using *Nextalign*. According to literature and the National Center for Biotechnology Information (NCBI), NC_045512.2 was considered as the reference sequence[5,53–55]. After that, sequences that were incomplete or much shorter than the reference sequence [having more than 3000 (non-template nucleotide) Ns and gaps ('–')] were excluded (Sup. Algorithm 1). The phylogenetic tree was constructed by using *IQ-TREE & TreeTime* (phylogenic tree—S1)[56]. Afterward, using the *augur ancestral module*, sites of mutations in genomes and phylogenetic tree's internal nodes were detected, as well as in the following translated into amino acids by the *augur translate module*. In the following, the number of mutations in each site was calculated with Python programming (Figs. 1 and 2). In this way, we could illustrate mutations in each region of the genome for each variant or total variation in each protein (Figs. 2 and S2).

In this study, we mainly focused on surface proteins that are mostly targeted by the immune system and vaccine design. In fact, events (alteration in the amino acid sequence of proteins) of the spike (S), membrane (M), nucleocapsid (N), and envelope (E) proteins have been magnified and illustrated (Fig. 2). As demonstrated in Sup. Table 3, about 30% of the amino acid positions of S, E, and M proteins and about half of the amino acid positions of N protein have been mutated or at least have one event. A brief look at the charts (Fig. 2) reveals some areas with a high number of events (red area) and some areas with fewer mutations (green area). For example, in the case of S protein, in about the first 220 amino acid sites, there is a high density of events besides the great number of mutations in several sites. In another hand, we can see areas with a low number of mutations that seem to be more conserving (green area). However, this type of categorization into green and red areas has been done upon qualitative decision, but it is clear enough to demonstrate the obvious difference in terms of conservity

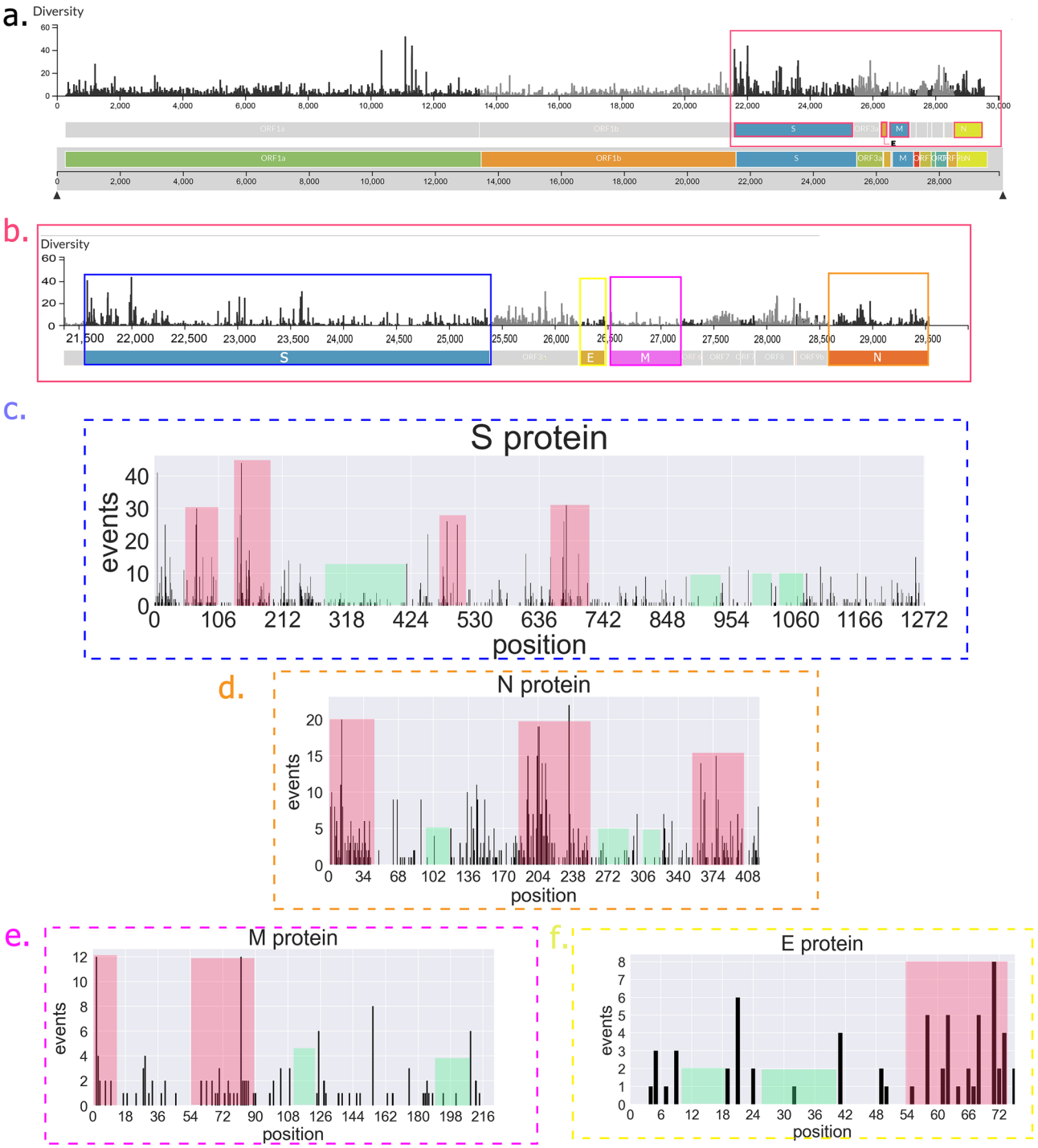

Figure 2. Mutations (events) in SARS-CoV-2 surface proteins (S, N, M, and E proteins) based on changes in their amino acid sequences. All events have been recorded on SARS-CoV-2 proteins illustrated in the first diagram (panel **a**). We focused on 4 main surface protein changes (the red box, panel **b**). In panel c (dash lined blue box), S protein events are illustrated. Qualitatively, we can see some red areas that have a higher number of mutations in their sequences (or high density of mutations), in comparison with green areas with the lower number of mutations in their sequences. All these changes and this qualitative view (red and green areas) are also presented for N, M, and E proteins in panels d, e, and f respectively.

among sites of each protein. In the following, we tried to calculate this qualitative difference into quantitative form for better decision-making. Before addressing the next step, here we highlighted the concerns with these differences (green and red areas).

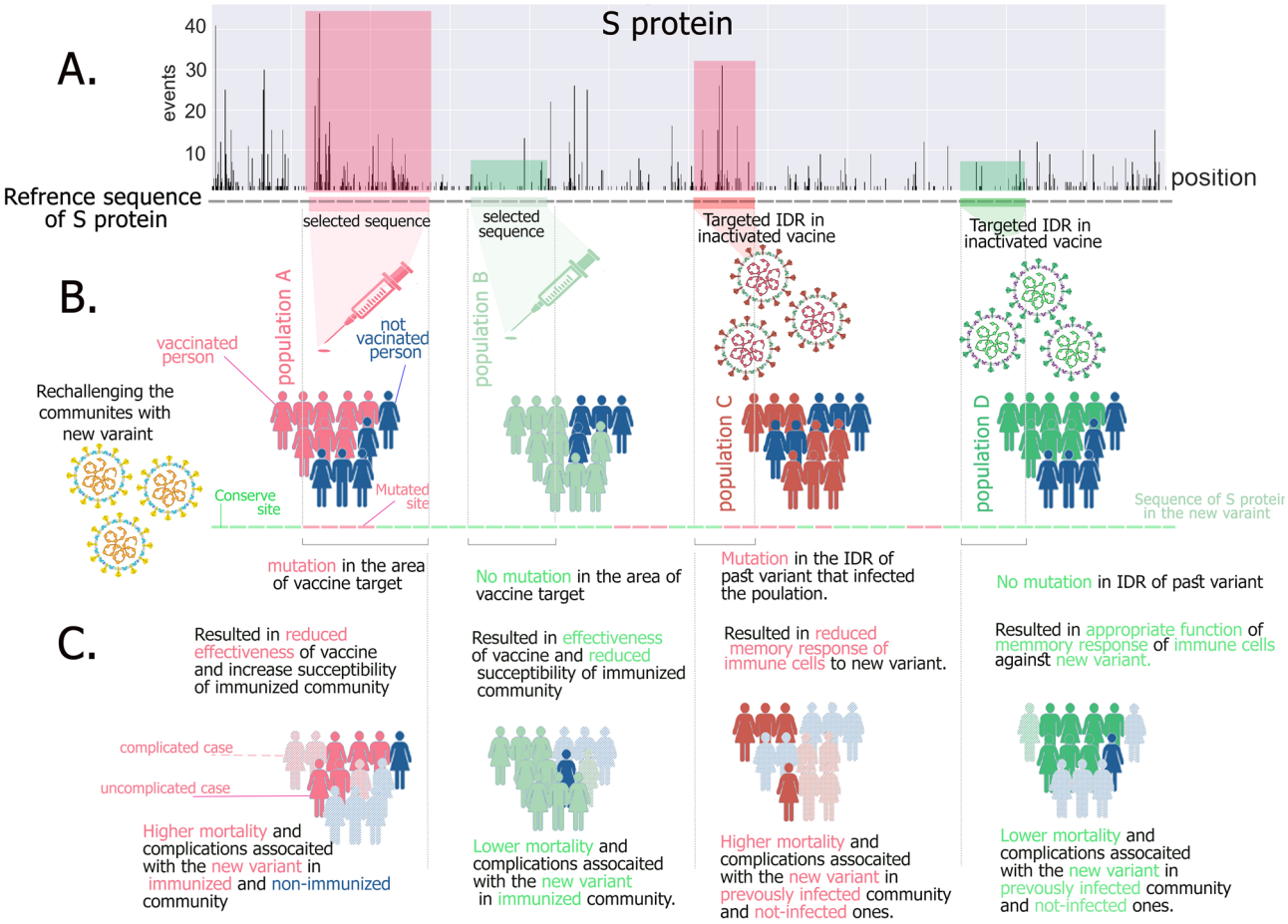

Figure 3. Different population with different exposure and vaccination. Panel (**A**) presents a mutational diagram of S protein with conserved and mutated sites and areas as demonstrated before. In panel (**B**) a schematic view of the S protein sequence with mutations and different regions with the qualitative view is illustrated. Different populations with varying types of immunization from various parts of the S protein sequence are presented. In panel (**C**), a rechallenge of all communities with a new variant is shown and the outcome of different sites of S protein that had been used for immunization is illustrated schematically. Blue dummies in the figure represent the not vaccinated population and the pale ones represent dead people in the community. Painted dummies in all colors except dark blue represent past infection or vaccination.

Suppose the population A (Fig. 3), where a significant number of people have been vaccinated with the vaccine A. At the time of designing this vaccine, there has been little data available about SARS-CoV-2 mutations. In fact, the epitope selected from the reference sequence of SARS-CoV-2 for designing vaccine A is accidentally located in the red area of the S protein, although it has significant immunogenicity. In another case, consider population B, where a significant number of people have been vaccinated with vaccine B. The epitope selected from the reference sequence of SARS-CoV-2 for designing vaccine B is accidentally located in the green area of the S protein. In population C a majority of communities have been infected with the strain C of SARS-CoV-2 previously.

This strain has an immunodominant region (IDR) in the red area of the S protein. In another hand, consider population D, where a majority of its community has been infected with the strain D of SARS-CoV-2, previously. Strain D also has one IDR in the green area of the S protein. In the dissemination and infection of all 4 populations with a new strain of SARS-CoV-2, which shares lots of mutations in all areas mainly in red areas; the difference will be more noticeable. In this example, it is somewhat obvious that populations A and D are more affected by the new strain of the virus. These populations would have higher infected patients in hospitals and higher mortality rate. With this view of the epidemic and considering the ever-changing virus, it is possible to better analyze that population is more affected by the new strains. With this in mind, it is also possible to reduce the burden of the disease on the community at risk by properly planning vaccination with the appropriate vaccine against the new species spreading. As a large number of mutations are found in different regions of the virus surface proteins, selecting areas that are less mutated is difficult to target with vaccination. But if such an area can be detected to target with the vaccination, it can be hoped that the designed vaccine will work well against the variants so far. In addition, this vaccine will also be less likely to be targeted by new mutations in the virus.

In the following, we tried to target regions in the virus genome that have appropriate immunogenicity besides acceptable conservity.

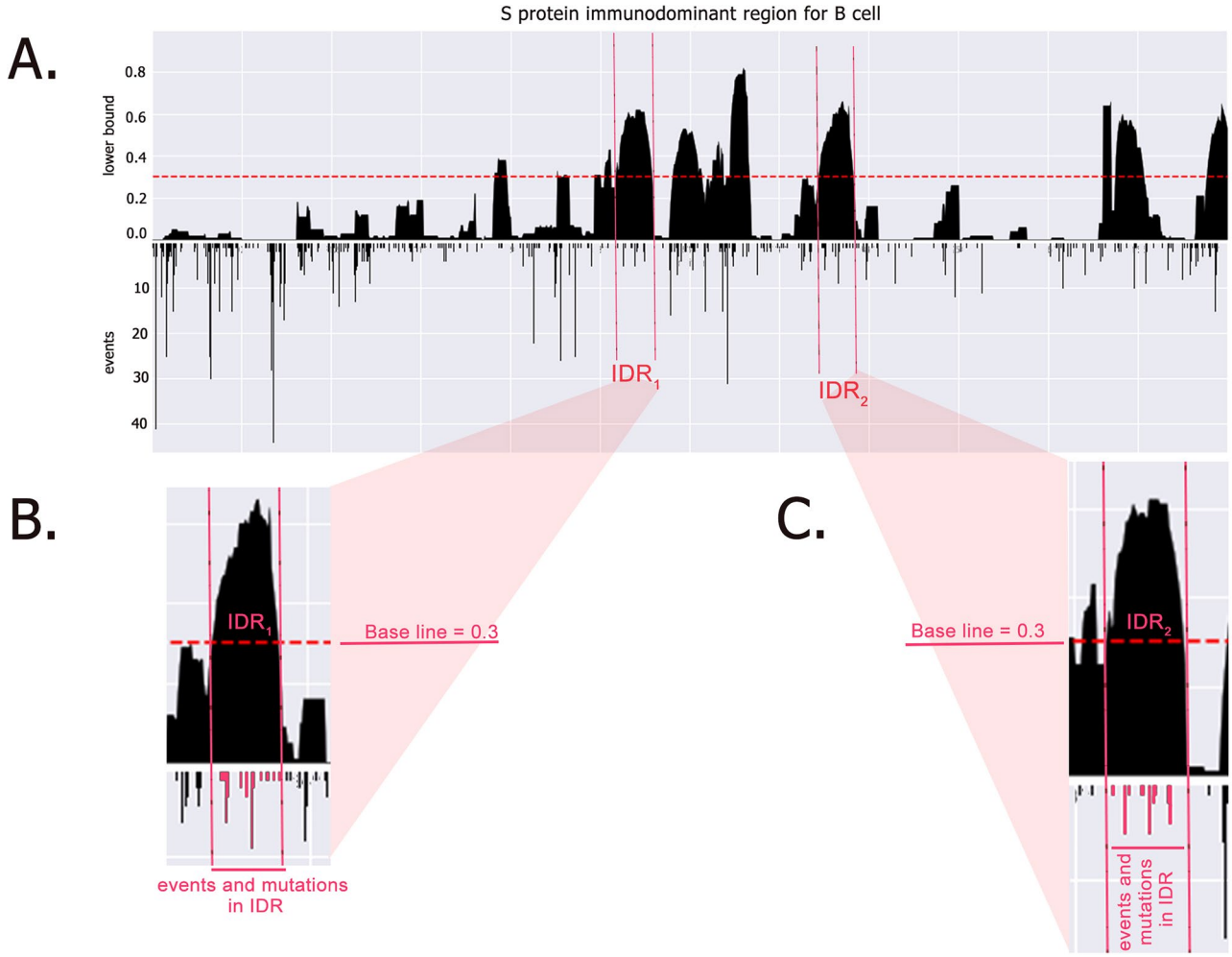

Figure 4. Example of IDRs of S protein for B cell response and events in their IDRs. In panel (**A**), two examples of IDRs (panel **B**,**C**) are magnified and events in their site (vertical bars) are shown. This is a schematic view of understanding how we calculated the number of mutations in IDRs. For example, in panel (**B**) we can see an IDR with a length of about 30 amino acids with many mutations at its site.

Investigating mutations within IDRs of SARS-CoV-2. The immune response to SARS-CoV-2 has been investigated in past years. As of May 2021, 3330 SARS-CoV2 linear epitopes were available on IEDB as they have been reported in the peer-reviewed literature. About 2/3th of them (2203 epitopes) were from S, M, N, and E proteins. These epitopes are derived from the SARS-CoV-2 genome and they have been investigated in the laboratory if they could stimulate the immune response. In this study, we have focused on S, M, N, and E protein derived B cell, HTL, and CTL epitopes (Sup. file 1).

So, we have extracted B cell, HTL, and CTL epitopes from S, M, N, and E proteins from IEDB. Here is to note that this search result is restricted to epitopes with recognition by major histocompatibility complex (MHC) class I and class II molecules. Then, extracted epitopes were mapped back to a SARS-CoV-2 reference sequence[53] using the IEDB's immunobrowser tool[57]. This tool helps to identify the IDRs by considering their response frequency (RF) score or lower bound in the diagram (where the RF score was $RF \geq 0.3$ considered as IDR) (Fig. 4). As all available records aligned along the reference sequence, the RF score was calculated by the positivity rate (positive response noted) divided by the total number of records (number of independent assays) (see the following equation)[58].

$$RF = \text{Positive response rate/Total number of records}$$

Our result demonstrated several IDRs in each protein. S protein has about 12 IDRs in terms of B cell epitopes and 5 IDRs in terms of HTL epitopes (Sup. Figs. 2 and 3). The numbers of IDRs and events in IDRs of each protein are summarized in Sup. Table 2. By merging 1st section data and this section, variations in IDRs become more prominent. As there are considerable numbers of mutations in IDRs (Fig. 4), significant immune response alteration cannot be out of mind. Because these areas (IDRs) are most likely to be identified by the immune system and result in an appropriate immune response stimulation against the virus, IDRs are suitable targets

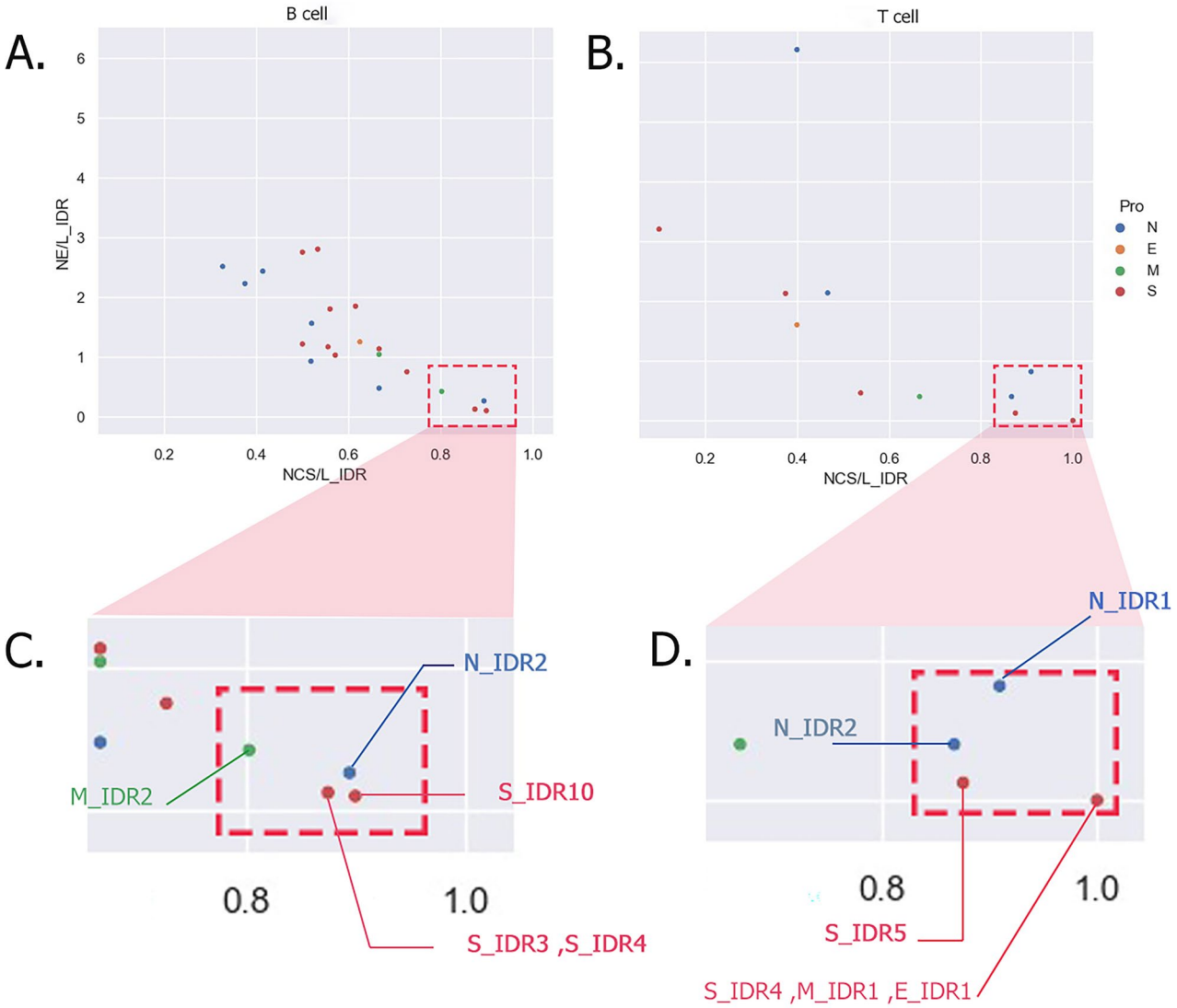

Figure 5. IDR plot. *NE* number of events, *L_IDR* length of IDR, *NCS* number of conserve sites. In panel (**A**) T cell IDRs are plotted with their number of events, their length and the number of conserve sites. As the number of conserved sites increases and the number of events decreases (Right Lower Quadrant), the IDR is more conserved in variants. Several T cell IDRs are selected as the most conserved IDRs (magnified in panel (**C**)). In panel (**B**) B cell IDRs are plotted with the same parameters. Several B cell IDRs are selected as the most conserved B cell IDRs (magnified in panel (**D**)). Here is to note that one dot in the plot can represent several IDRs that have the same normalized conserve sites and events.

for vaccine design. And also, mutations in such areas, as illustrated in Fig. 4 and Sup. Tables 1 and 2, could be a possible challenge for memory cell response.

This means even a person, who has previously been infected with the virus or has been vaccinated is still prone to be infected by new variants of SARS-CoV-2 like population C. This issue can explain to some members of VOCs that they are associated with a decrease in the efficacy of vaccines and monoclonal antibodies (such as omicron and delta variants) in the neutralization and treatment of patients suffering from COVID-19[23,59–64].

Here IDRs are plotted, to make a quantitative comparison among all IDRs in terms of several mutated sites and events. As a number of events and mutated sites are not comparable due to the different lengths of IDRs, we normalized these parameters by dividing them using the length of each IDR (Fig. 5).

As illustrated in Fig. 5, IDRs are compared in plots in terms of their conservity. There are several B cell IDRs (Fig. 5, Left graph) in S, M, and N proteins that have appropriate conservity (red box). In T cell IDRs (Fig. 5, Right graph) there are 6 IDRs seem to be conserved. This means that these selected IDRs in the plot contain epitopes with higher conservity. Thereupon, selecting epitopes from these areas in the SARS-CoV-2 genome for further vaccine design seems to be more effective against variants discovered thus far. These IDRs sequences and locations are available in the supplement (Sup. file 1).

As mentioned, each IDR includes several epitopes. According to significant mutations in IDRs, we wondered if past reported epitopes were affected by mutations in new variants of SARS-CoV-2. In the next section, we try

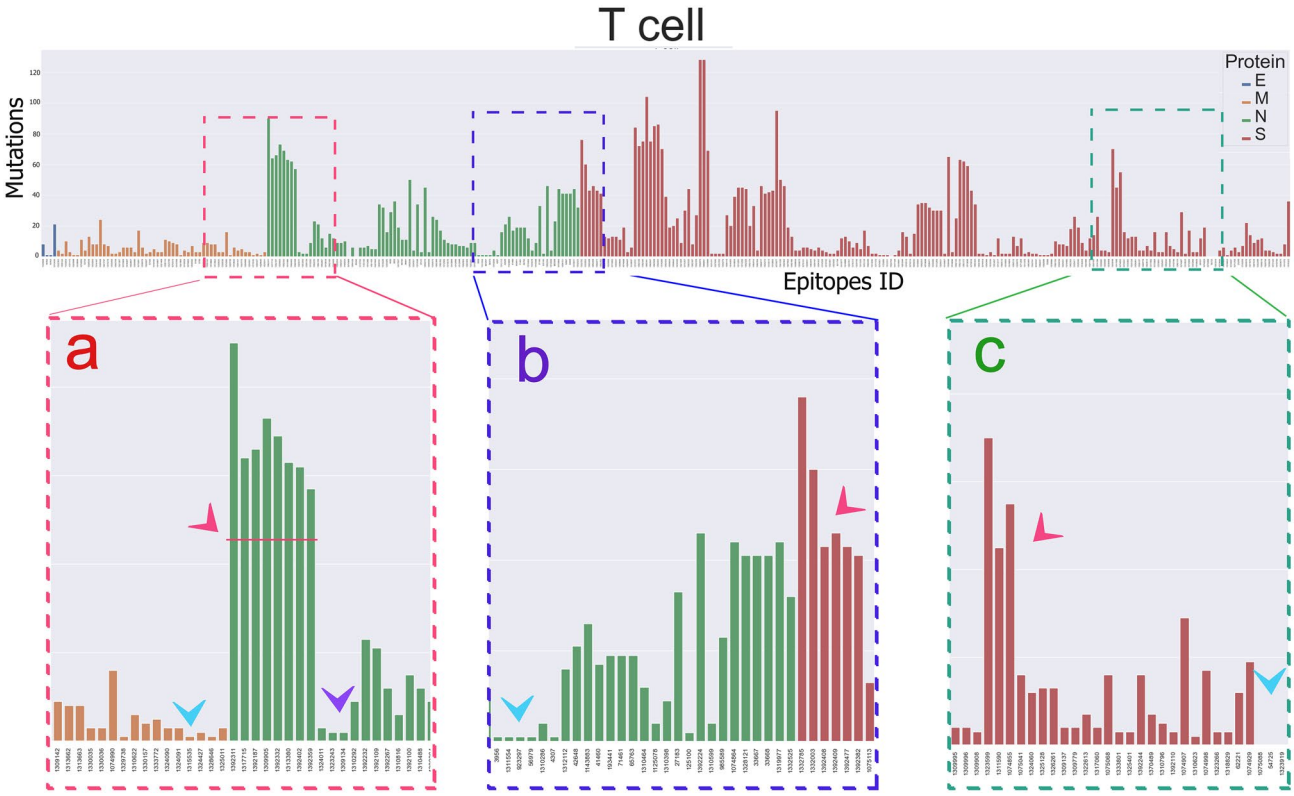

Figure 6. Mutational diagram (MD); Illustration of IEDB epitopes and their mutations. In this figure number of mutations is illustrated in vertical bars and epitopes ID is under each bar in horizontal axis. Panel (**a**–**c**) magnifies 3 different parts of MD. For example, in panel (**a**) there are epitopes from M protein (orange bars) that have a small number of mutations (sky blue arrows head) besides, epitopes from N protein (green bars) with the high number of mutations (pink arrows head). Purple arrows head point to sites with a moderate amount of mutations.

to understand how different mutations in the SARS-CoV-2 genome affected IEDB suggested epitopes in past laboratory research and also predicted ones.

Significant mutations in SARS-CoV-2 IEDB epitopes. As mentioned before, we mapped all mutations from known variants to SARS-CoV-2 M, N, E, and S protein sequences (Fig. 4). Also, we have extracted B cell, HTL, and CTL epitopes of these proteins from IEDB. Then epitopes were aligned to their reference sequence, in order to map the mutational diagram (MD) of each protein. So, we could find out where each epitope is located in its protein sequence and how this located part has undergone alterations by mutations (Sup. Fig. 4 and Fig. 6).

A list of IEDB epitopes and their identity code is available in the supplement (Sup. file 2). As demonstrated, almost all investigated IEDB epitopes of proteins have undergone mutational changes (Fig. 6 and Sup. file 3). But the other point of view, which attracts attention, is the difference in the number of mutations in different epitopes. In Fig. 6, there are epitopes with a high number of mutations (Fig. 6a red arrows head) besides, epitopes with a low number of mutations (Fig. 6b blue arrows head) or even absolutely conserved ones (Fig. 6c blue arrows head).

The results show that almost all epitopes that have been studied in the laboratory over the past years have changed in the same period. As these epitopes were evaluated and utilized for vaccine design, a reduction in the effectiveness of vaccines is very likely. As well, it could be an important alarm for the international community and clarify the necessity of research on new variants for vaccine design development. It may also be a better option to consider faster and more efficient approaches in vaccine design, such as mRNA vaccines.

Since most epitopes are mutated in the case of SARS-CoV-2, it makes sense to choose epitopes with minimal changes (fewer mutations) and maximum stability (higher conservity). While considering epitopes conservity, maintaining immunogenicity properties also should be noted. Given that, choosing the best epitope should be considered several factors that make the choice more difficult. This makes the issue more complicated than previous approaches (where just immunologic properties were important). Hence, we kept our research on new variants with a focus on how to choose the best epitope for further investigations and application in vaccine design and development.

Plotting events against immunologic properties of IEDB epitopes. We choose RF representative IEDB epitopes immunologic property as X-axis value. Besides, all mutations (without considering the site of mutation) in each epitope have been selected as events of the Y-axis. In this way Epitope Event—RF Plot (EVRP)

depicts past events to each epitope beside their immunologic property (Sup. Fig. 5). This means epitopes events can be studied retrospectively in the case of variants and mutations.

As can be seen in (Sup. Fig. 5), all epitopes are arranged according to the two factors in the EVRP. For vaccine design, epitopes are desirable that have good immunogenicity along with a low number of mutations. So, the lower right quadrant of this chart simply shows the best epitopes. It is also possible to easily compare two or more separate epitopes in this chart in terms of both factors. EVRP also shows that a large number of epitopes detected and tested in the laboratory are either weakly immunogenic (especially in the case of T-cell epitopes) or have been affected by dozens of mutations. Only a handful of them is suitable for further research in vaccine design.

In past, we had to consider thresholds for each parameter and just omit data under threshold value and then address the remaining data for evaluation with another parameter. But in the case of EVRP two parameters were illustrated simultaneously for each epitope, which facilitates comparison between epitopes. This feature allows us to choose the best among a large number of epitopes from different proteins by considering two characteristics.

Also, in the past approach, it was difficult to compare a large number of epitopes. Herein, comparing several epitopes is easy. Although EVRP is a better tool for comparison, it cannot generate an accurate valuation for each epitope based on the site of the mutation and the number of the mutations.

For instance, consider a and b epitopes with lengths of 10 amino acids (Sup. Fig. 6). Epitope a gained 10 mutations in 2 sites and b epitope gain 10 mutations in 5 sites. EVRP just illustrates them in the same position in the diagram, although epitope b's conservity is lesser than epitope a. As non-conserved sites increase in epitopes, the probability to gain further mutations rises. Thereupon, it is necessary to properly weigh each of the parameters to make epitopes comparable.

In the past sample (epitopes a and b) we compared epitopes of equal length, but in reality, we face epitopes with different lengths from different proteins. These issues make analogy difficult. So, we utilized another method for epitope evaluation to differentiate these cases.

TOPSIS is a multi-criteria decision-making approach, suitable in situations where several factors influence the decision. This technique was developed in the late twentieth century by Ching-Lai Hwang and Yoon[65]. In addition, TOPSIS is a concept that the chosen alternative (parameters) should have the shortest geometric distance from the positive ideal solution (PIS) and the longest geometric distance from the negative ideal solution (NIS).

Here, we used TOPSIS to quantify and Shannon entropy to weight events of epitopes logically considering two parameters: the sites of mutations and, the number of mutations in each site.

So, we kept RF on X-axis but we changed the Y-axis value to the TOPSIS score of mutations for epitopes. Epitope RF–TOPSIS Plot (ERTP) is designed to solve previous problems and facilitate comparison among epitopes while considering several factors (Fig. 7). By rearranging the epitopes based on the factors weighed down with TOPSIS, it is possible to make a more accurate comparison between them. The ERTP diagram clearly shows how the epitopes examined so far have been affected by mutations at different points in their sequence.

It should be noted that the lower the TOPSIS score, the more conservative the epitope is. Epitopes are worth more research in the laboratory or clinic, which are in a more appropriate area in the chart than others. We are currently facing a pandemic of a constantly evolving virus; we must use the best of the available vaccines based on the strain spreading in each region. Vaccines made from the inactivated pathogen should be screened for IDRs in the same system as introduced, and other vaccines targeting a specific epitope should be evaluated by the same system. As mentioned, mutations are inevitable in IDRs and epitopes in the case of SARS-CoV-2. But to counter the new variants and especially VOC, we can administrate immunogenic regions that have been less mutated so far for vaccine design.

So far, we have discussed the epitopes that have been discovered and tested in the laboratory. Herein, we wonder if this system could be effective to evaluate the new predicted epitopes by immunoinformatic databases. Predicting and evaluating epitopes with ERTP can facilitate designing new vaccines against the spreading variants. This not only helps to produce more effective vaccines but can also greatly reduce the cost of vaccine production and save time.

At this point we decided to run the ERTP method on SARS-CoV-2 predicted epitopes as well. To find epitopes those, despite high immunogenicity, are less mutated between different variants (conserve among variants so far).

Prediction and assessment of SARS-CoV-2 epitopes. *CTL epitopes.* By using the NetCTL v1.2 server a total of 700 CTL epitopes were predicted for 12 MHC class I supertypes for M, N, E, and S proteins. Then their immunogenicity was assessed by IEDB class I immunogenicity tool and 400 of them had been predicted to be immunogens. Out of these 400 epitopes, 233 non-toxic (by ToxinPred server) and non-allergenic (by AllerTOP 2.0 server) ones were predicted and selected for further evaluation. In the end, 114 immunogenic, non-toxic, and non-allergenic epitopes were analyzed by the Vaxijen server[66,67] in the case of antigenicity and their scores were recorded for plotting as the X-axis value (Sup. file 4).

HTL epitopes. 94 probable HTL epitopes were predicted for S, M, N, and E proteins by IEDB MHC class II allele binding prediction tool (percentile rank ≤ 0.25)[68,69]. In the case of HTL predicted epitopes, 64 epitopes were non-allergenic and nontoxic by AllerTOP 2.0[70] and ToxinPred[71,72] server, respectively. As induction of interferon (IFN)-γ, interleukin (IL)-4, and IL-10 secretion by these epitopes play a pivotal role in regulating immune response, HTL predicted epitopes were analyzed for these properties. Using the IFN epitope[73], IL4pred[74], and IL10pred[75] servers, 5 epitopes were left for the next step. Thereby, epitopes were analyzed by the VaxiJen server for antigenicity, and their scores were recorded for further analysis (Sup. file 5).

Linear B Lymphocytes (LBL) epitopes. With the aid of iBCE-EL server[76], 53 LBL epitopes were predicted for S, M, N, and E proteins. 42 of them had the appropriate length of more than 6 amino acids. Out of the 42, 20

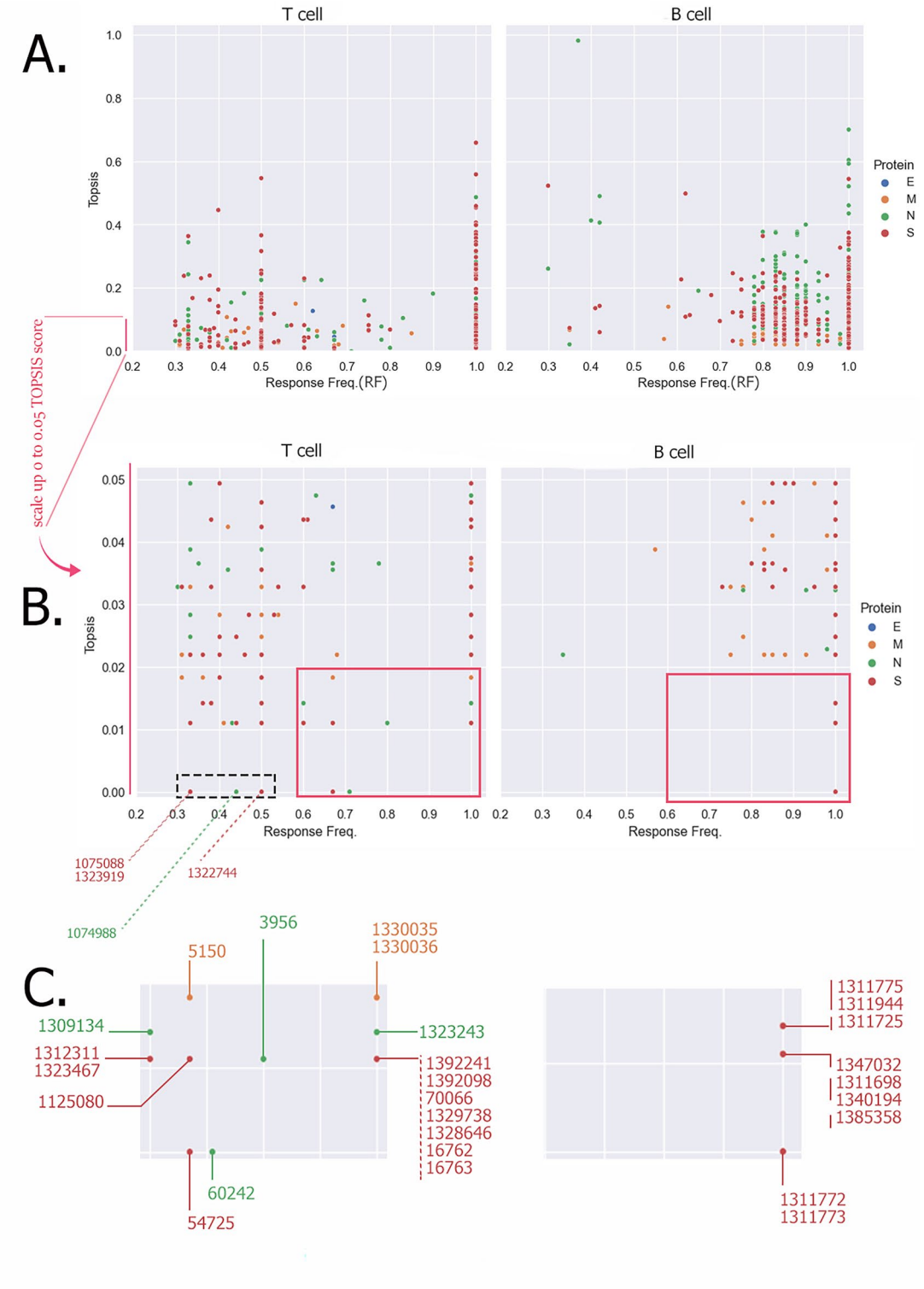

Figure 7. ERTP illustrates IEDB B cell and T cell epitopes immunologic property (by RF) along with conservity by TOPSIS scoring. In Panel (**A**) a logical comparison between IEDB epitopes has made through TOPSIS score (vertical axis) and Response Frequency (horizental axis). It helps to understand difference between IEDB epitopes with the same number of events but a distinct number of mutated sites. Panel (**B**) scales up vertical axis (TOPSIS score) for better illustration of position of different dots in the diagram. Right lower quadrant presents epitopes with higher immunologic properties besides higher conservity (the red boxes). In panel (**C**) red boxes are magnified and ID of each epitope (dot) is noted. It should be note that some dots represent for several epitopes with the same RF and TOPSIS scoring.

of these epitopes had been predicted to be non-toxic (by ToxinPred server) and non-allergenic (by AllerTOP 2.0 server). Like other predicted epitopes, VaxiJen scores of antigenicity were calculated for 15 predicted LBL epitopes (Sup. file 6).

All predicted epitopes mapped back to their protein reference sequence. In order to find mutations of predicted epitopes, we mapped the location of these epitopes on their reference protein sequence. As, we have sites of mutations in proteins in "Sect. 3" section, mutations in each epitope could be identified by merging these data (Sup. Fig. S7). In this way, the number of mutations and sites of mutations in each predicted epitope was understood. As demonstrated in (Sup. Fig. S7), even the predicted epitopes have significant mutations. There is a range from 0 up to more than 200 events in some cases. As discussed before, we need to select the best ones in terms of conservity and immunogenicity for further investigation. Therefore, we continued this study with the next step and plotting the predicted data.

Plotting the predicted epitopes with TOPSIS scoring. The Vaxijen antigenicity score of each epitope was considered as an immunological factor for the value of the Y-axis of the plot. Then, we calculated the TOPSIS scores of predicted epitopes in terms of sites and number of mutations due to Shannon entropy. In the following, the X-axis value stands for their TOPSIS score (Fig. 8 and Sup. file 7). This is clearly demonstrated in Antigenicity Score-TOPSIS Plot (ASTP) is the lower TOPSIS score in predicted epitopes in comparison with IEDB epitopes in ERTP. This may indicate that there are more appropriate options in the predicted epitopes that need further consideration. Also, proper dispersion of predicted epitopes with this method makes it easy to select the best ones based on their location in the plot. According to the plot, in the case of HTL predicted epitopes there is almost no suitable option. Either, LBL predicted epitopes offer limited options. In contrast, CTL-predicted epitopes have many appropriate choices. The closer we get to the lower right corner of the plot, the better the options.

As illustrated in Fig. 8, there are more suitable options in terms of CTL epitopes than HTL and LBL ones. In the case of HTL and LBL, there are just three epitopes with desirable properties. According to this method, the best epitopes in terms of mutation are shown in Fig. 8. Selected epitopes (Fig. 8 and Sup. Table 4) are the best epitopes ever discovered and predicted due to their immunogenicity and SARS-CoV-2 mutations.

Discussion

Lots of mutations have been identified in the SARS-CoV-2 genome, which resulted in new variants. Potentially developing mutations can result in structural changes in key proteins involved in the pathogenesis and spread of the virus, a point that has also been shown in studies[18,20,47]. This means regardless of various approaches to vaccine design; the vaccine target needs to be selected precisely with two major points; acceptable antigenicity and appropriate conservity.

In this study, at first, we detected mutations in the SARS-CoV-2 genome by considering 3369 sequences, extracted from the GISIAD. For this purpose, first, all the available sequences were compared with the reference sequence and then the phylogenetic tree was drawn. In this way, the number of mutations and the site of them, in the variants and sequences were examined. Our approach is similar to CoVariants and Nextstrain sites to depict diversities and mutations in the genome of different viruses.

So far, several methods have been explored to find different mutations and variants. Hassan et al. and Almubaid et al. considered a method similar to what we did in finding mutations but using different software and modules[77,78]. However, other studies such as Kames et al. and another article by Hassan et al. have used other methods, by the capabilities of NCBI site and blasting tools or using MATLAB software to achieve this goal[79,80]. It was important to our study, to find the number of mutations besides the sites of mutations for further scoring and analysis. This issue needed further programming with Python to improve modules function and find custom results that are similar to Schrors et al. study[81].

Then investigated IDRs of SARS-CoV-2 S, M, N, and E proteins have been considered for mutations. To find IDRs, IEDB immunebrowser tool was administrated with a threshold of 0.3. This approach is similar to Grifoni et al. Mukherjee et al. and S Zhuang et al. studies[82,83].

We found that most of IDRs have mutated, just 8 of them have acceptable conservity till now. The high rate of mutations in IDRs is in accordance with several past studies but with different methods[84–87]. Schrors et al. investigated large-scale mutations in S protein and considered the variants according to diversities in this area[81]. In another study by Zhuang et al. also, the IDRs mutations located in S protein were investigated for mutations[83]. A large number of mutations in both last articles is acceptable according to our result[81,83]. Until now, IDRs have not been extensively and comprehensively examined for mutations in SARS-CoV-2. In this study for the first time, all of these areas were examined with a comprehensive scientific approach. Furthermore, IDRs were plotted with normalized factors to compare these areas, which had not been addressed before. This plot simplifies the selection of the best IDRs to target.

Furthermore, we presented evidence of significant mutations in B cell linear epitopes and HTL and CTL epitopes from S, M, N, and E proteins of SARS-CoV-2, which are discussed in peer reviews extracted from IEDB. Mutations in epitopes have also been presented in other studies[88,89]. In previous studies, only one or several epitopes were examined for mutations, but in this study, with a comprehensive view, all IEDB epitopes were examined for this purpose.

Finding mutation itself is not enough to find the best epitopes. Thereupon in a step beyond, we scored their events with the TOPSIS method to quantify their conserved manner. In the setting of this algorithm and plotting with their immunogenicity properties the best epitopes were selected for further analysis in the laboratory. In Mullick et al. study, they investigated S protein mutations hotspot through Shannon entropy and K-means

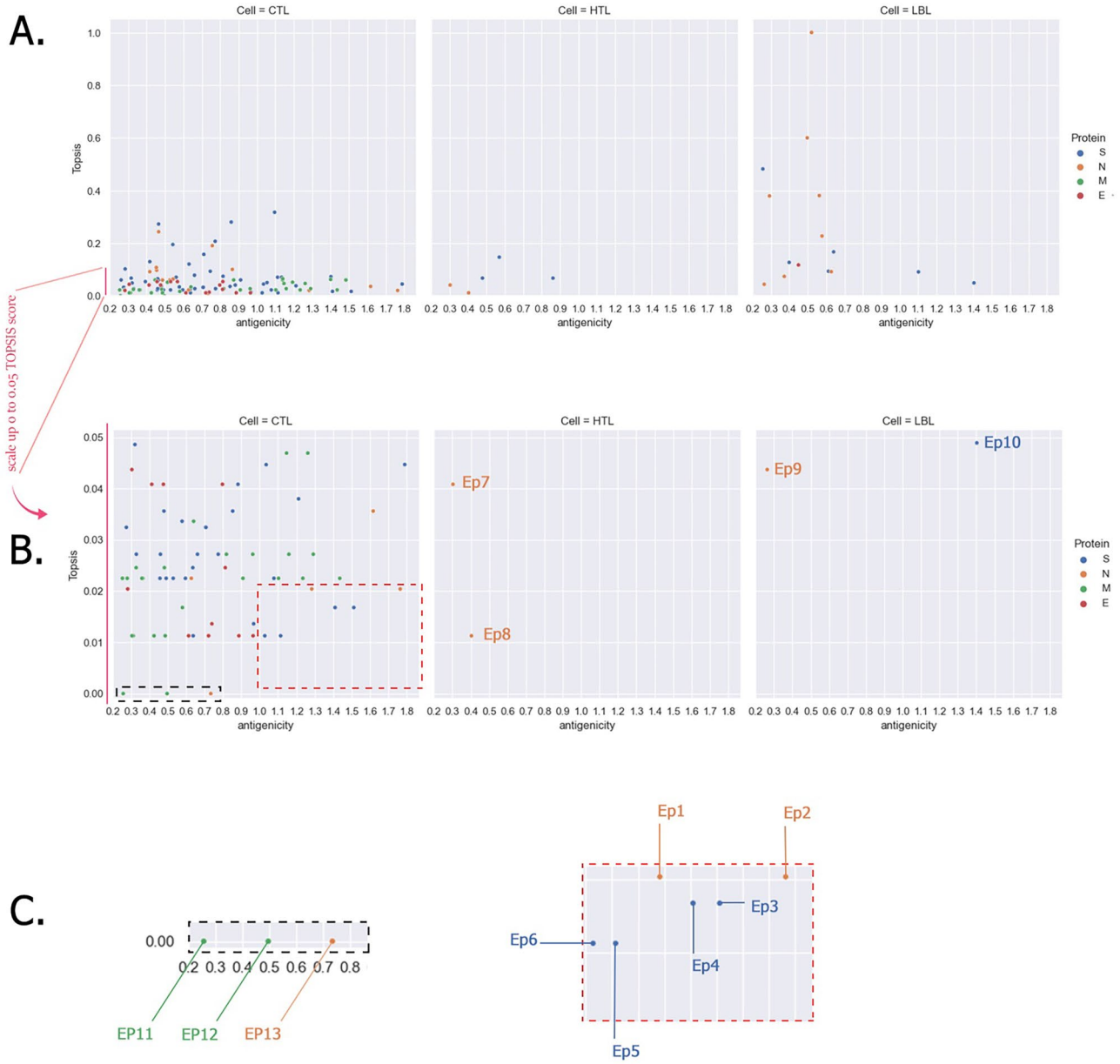

Figure 8. ASTP of epitopes from two perspectives. In Panel (**A**) predicted epitopes goes through a logical comparison due to TOPSIS score (vertical axis) and antigenicity (horizontal axis). This helps to differentiate between predicted epitopes with the same number of events but a distinct number of mutated sites. Again, right lower quadrant presents epitopes with higher immunologic properties besides higher conservity. Panel (**B**) scales up vertical axis (TOPSIS score) for better illustration of position of different dots (CTL, HTL, and LBL) in ASTP. The red and black boxes in CTL panel and EP7-EP10 of HTL and LBL epitopes are the most conserved ones with higher antigenicity. In panel (**C**) CTL epitopes boxes are magnified and ID of each epitope (dot) is noted. Sequence of each epitope is available in Sup. Table 4.

clustering[90]. Their approach in terms of scoring is different from ours, but their efforts to find a suitable way to score areas with high mutations as hotspots are admirable.

As our results showed, almost all epitopes have changed through mutations, which challenge the immune response against SARS-CoV-2 over time. This could also explain the generation of new and more dangerous variants or VOCs. In Garrett et al. study, S protein was investigated through variants with the Phage-DMS approach in the laboratory. They have shown that antibody neutralization has also changed as a result of IDRs mutation[91]. This confirms our hypotheses in this article in terms of the huge number of mutations in IDRs and epitopes that challenge the immune system and the importance of finding conserve epitopes to target.

There are several different scoring methods. Here we needed to weigh each factor (number of mutations and sites of mutations) in terms of conservity unequally with Shannon entropy. In this case, TOPSIS was preferred as the best choice for further scoring.

Also, we have tested this approach (TOPSIS scoring and plotting) in predicted epitopes. To predict the epitopes there are different databases and approaches. In this article, we have predicted epitopes with a platform designed in Ahammad et al. study[92]. After prediction of epitopes, we ran Shannon entropy and TOPSIS scoring and plotting through antigenicity and conservity, successfully for identification of best epitopes with two factors. In addition, having a quantitative view in terms of conservity and antigenicity with efficient illustration and plotting can help to simplify rational decision-making in a variety of epitopes.

The next and final section was to select the best epitopes. Illustration of epitopes scores (by plotting) helps to understand their different properties on a large scale. Besides, having a quantitative measurement in terms of conservity can help to make rational decisions. This plot provides a roadmap understanding antigens and epitopes alteration in variants. Also, this could help to select epitopes by considering different criteria. This method is an open-source multi-criteria platform for deciding on the best epitopes.

Our findings indicated the rapid and significant changes in IDRs and epitopes. This highlights the importance of a system for the efficient development of vaccines against new strains and variants. As this platform retrospectively finds previous changes and mutations in epitopes and IDRs, it could help to understand the development of recent emerging variants.

The limitation of our study. There are two main categories of epitopes: linear epitopes and conformational or 3 d epitopes. T cells recognize linear epitopes and B cells can recognize conformational epitopes and linear ones[93–95]. As we aimed to understand changes in the sequence of proteins, we focused on linear epitopes of IEDB and predicted ones. Of course, mutations can make conformational changes in epitopes and antigens tertiary structure and affects B cells' recognition of epitopes. This issue (considering conformational changes in epitopes with mutations) is important to be addressed by further research.

In this study, we faced with the limitation of computing power, so instead of using all the genomes found in the world, we tried to use those that have the most differences and located in different components of the phylogeny tree. Although this problem has been solved with a scientific and algorithmic approach (Sup. Algorithm 1), it should be taken into consideration by other researchers.

Our platform is not able to predict future variations and it is based on past mutations that happened in the SARS-CoV-2 genome. It will help to understand mutations retrospectively and select appropriate epitopes against recent emerging variants. This platform needs recent new variants' sequences data to be up to date for further analysis.

As we aimed to find a method for the selection of epitopes with different properties, we didn't focus on vaccine design. In the end, we just mention several epitopes as a potential potential candidate for vaccine design. Of course, these epitopes need further investigation like; vaccine constructs design, molecular docking, prediction of population coverage, and laboratory studies.

Conclusion

As we face a pandemic caused by an ever-changing virus; providing a scientific and practical approach to select epitopes with appropriate conservity and immunogenicity seems to be crucial.

Also, we found some predicted epitopes are more conserved and immunogen as a potential candidate for vaccine design (Sup. Table 4). According to our findings (Sup. Table 4 and Fig. 8), LSPRWYFYY and DLSPRWYFY predicted CTL epitopes from N protein, and VVFLHVTYV, GVVFLHVTY, VRFPNITNL, and PYRVVVLSF predicted CTL epitopes from S protein are highly conserved and immunogen and suitable for vaccine design. Using an AAY linker between CTL epitopes, we can design a multiple-epitope vaccine. Also, WPQIAQFAPSASAFF and QIAQFAPSASAFFGM predicted HTL epitopes from N protein and AGLPYGANK predicted LBL epitopes from N protein can be added to CTL epitopes through linkers (like GPGPG). There are other choices for vaccine design between predicted epitopes and IEDB epitopes that can be added to this vaccine. Here we selected the 7 best epitopes for vaccine construct design.

At the end, we introduce a scientific method, with the hope that this protocol will aid in the development of vaccine design against SARS-CoV-2, particularly the VOC.

Methods

Detecting site and number of mutations in variants. In order to detect mutations made over the past year in the SARS-CoV-2 genome, the following procedure was done (Fig. 1). First, SARS-CoV-2 complete genome sequences have been collected from the global initiative on sharing avian flu data (GISAID) database (http://www.gisaid.org). Then, the collected sequences were aligned with the reference genome (hCoV-19/Wuhan/Hu-1/2019) by Multiple Alignment using Fast Fourier Transform [MAFFT; a multiple sequence alignment program (cbrc.jp)]. In the following, sequences with more than 3000 not read bases (Ns) and gaps ('–'), wrong dates before 2019, and sequences shorter than 1000 base pairs (bp) were excluded from the analysis.

Herein, using IQ-TREE[56] [Cibiv/IQ-TREE: Efficient phylogenomic software by maximum likelihood (github.com)] phylogenetic tree was constructed for all remaining sequences. In the next step phylogenetic tree needed to reroot and resolve polytomies. Therefore, TreeTime [neherlab/treetime: Maximum likelihood inference of time stamped phylogenies and ancestral reconstruction (github.com)] infers the tree's internal nodes dates and prone sequences.

Afterward, using the augur Ancestral module [augur/ancestral.py at master next strain/augur (github.com)], mutations in genomes, and the phylogenetic tree's internal nodes were inferred in their sequence.

In order to translate sequences and their mutations into Amino Acids, we used the augur translate module [augur/translate.py at master nextstrain/augur (github.com)]. Employing predefined Amino acids mutations

(mutations in protein sequence) as input, we calculated the diversity of each SARS-CoV-2 genome region by Python programming (version 3.8).

Exploring mutations within IDRs of SARS-CoV-2. B cell, HTL, and CTL epitopes from S, M, N, and E proteins were extracted from IEDB. Then search result was restricted to epitopes with recognition by human leukocyte antigen (HLA) that is the human version of MHC molecules, In the following, extracted epitopes were mapped back to a SARS-CoV-2 reference sequence[53] using the IEDB's immunobrowser tool[57]. As all available records aligned along the reference sequence, the RF score was calculated by the positivity rate (positive response noted) divided by the total number of records (number of independent assays) (see the following equation) [58].

$$\mathrm{RF} = \text{Positive response rate/Total number of records}$$

In this way, IDRs were identified by considering their RF score or lower bound in the diagram (where the RF score was RF ≥ 0.3 considered as IDR) (Fig. 4, Sup. Figs. 2.1 and 2.2).

Discovering mutations in SARS-CoV-2 IEDB epitopes. Extracted B cell, HTL, and CTL epitopes of M, N, E, and S proteins from IEDB were aligned to their reference sequence, in order to map the mutational diagram of each protein. As we have the number of events and mutations in each site of the reference sequence in step 2, we could find out where each epitope is located in its protein sequence and how this located part has undergone alterations by mutations.

Plotting events against immunologic properties of IEDB epitopes. We chose RF of IEDB epitopes as the X-axis value. Besides, all mutations (without considering the site of mutation) in each epitope have been selected as events of epitopes for the Y-axis. In this way EVRP clearly depicts past events to each epitope besides their immunologic property.

Epitope prediction. *CTL epitope prediction.* NetCTL v1.2 server is a powerful free source for predicting CTL epitopes[71]. This server integrates the prediction score of MHC class I binding peptides and; proteasomal C terminal cleavage with transporter associated with antigen processing (TAP) transport efficiency score; to deliver an integrated score for CTL epitope prediction from a sequence. In this study, 9-mer CTL epitopes were predicted by using the NetCTL v1.2 server for 12 MHC class I supertypes (A1, A2, A3, A24, A26, B7, B8, B27, B39, B44, B58, and B62)[72]. Epitopes that reached above 0.75 scores as threshold was selected for the next step[92]. There is no reference and definite value for thresholds. Thereupon, we followed other articles used similar methods to find the value of threshold (> 0.75)[92,96,97].

To predict the immunogenicity of the CTL epitopes, the Class I immunogenicity tool of the IEDB Analysis Resource was administrated[40]. Epitopes with a positive value for immunogenicity were selected for the next steps (A percentile rank score ≤ 2).

Here is to note that there is no reference value for the percentile rank. According to similar studies and our study design, percentile rank ≤ 0.25 was considered for this study[92,98].

HTL epitope prediction. To predict 15-mer HTL epitopes and their MHC class II alleles, a consensus algorithm of the IEDB MHC class II binding tool was administrated[41,99,100]. Epitopes with percentile rank ≤ 0.25 were considered for the next steps. It is important to note that there is no reference value for the percentile rank. According to similar studies and our study design, percentile rank ≤ 0.25 was considered for this study[92,98].

Prediction of linear B cell epitopes. For the prediction of Linear B cell epitopes iBCE-EL server was used[76,101]. Epitopes with positive values were selected for further analysis.

Allergenicity prediction. Allergenicity of epitopes was assessed using AllerTOP 2.0 server[70]. The non-allergic epitopes were subjected to the next steps for further analysis.

Toxicity prediction. The toxicity of selected epitopes was evaluated using the ToxinPred server[102,103]. The non-toxic epitopes were subjected to the next steps for further analysis.

TOPSIS scoring method and Shannon entropy. TOPSIS is one of the best multiple decision-making methods. In this method, 'i' is the number of alternatives that can be evaluated by the number of attributes 'j'. In the decision matrix, (epitopes) were considered as alternatives (the total number of mutations), and (the number of sites of mutations) in each epitope was considered as attributes. Attributes were weighted by Shannon entropy. Then, the decision matrix was normalized. The normalized decision matrix (N) was multiplied by a diagonal matrix of attributes weights (Wj × j). The positive ideal solution (Vj+) and negative ideal solution (Vj−) were determined by a weighted normalized decision matrix (V). The difference between each attribute of epitopes from positive and negative ideal solutions (Vj, Vj−) was calculated. The relative closeness of each epitope to the ideal solution was determined. By sorting epitopes into rating order, the epitopes with fewer scores were detected as more conserve ones. It should be mentioned that the epitope, which gets a score of zero in the TOPSIS method, may not be absolutely conserved. The score of zero stands for more conserved epitopes than the other evaluated epitopes. TOPSIS matrix equation and Shannon entropy equation are noted in the supplementary file equation part.

Predicted epitopes events and plotting with TOPSIS. We mapped the location of predicted epitopes on their reference protein sequence, in order to find mutations of predicted epitopes. As we have sites of mutations in proteins in past sections, mutations in each predicted epitope could be identified by merging these data.

The Vaxijen antigenicity score of each epitope was considered as the value of the Y-axis of the plot. Then, we calculated the TOPSIS scores of predicted epitopes in terms of sites and number of mutations. In the following, the Y-axis value stands for their TOPSIS score for plotting. The same has been done for IEDB epitopes.

Ethics approval. No human or animal models were utilized in this investigation. All experiments were performed according to the guidelines of the Medical Ethics Committee of the Jahrom University of Medical Sciences (IR.JUMS.REC.1399.090).

Data availability

Our data and codes are available through corresponding author if asked.

Received: 10 April 2022; Accepted: 5 August 2022
Published online: 18 August 2022

References

1. Organization, W. H. in *daily* 1 (WHO, 2021).
2. Organization, W. H. *Origin of SARS-CoV-2* (World Health Organization, 2020).
3. Hasanzadeh, A. *et al.* Nanotechnology against COVID-19: Immunization, diagnostic and therapeutic studies. *J. Control Release* **336**, 354–374. https://doi.org/10.1016/j.jconrel.2021.06.036 (2021).
4. Huang, C. *et al.* Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* **395**, 497–506 (2020).
5. Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C. & Garry, R. F. The proximal origin of SARS-CoV-2. *Nat. Med.* **26**, 450–452 (2020).
6. Ingraham, N. E. *et al.* Immunomodulation in COVID-19. *Lancet Respir. Med.* **8**, 544–546 (2020).
7. Zheng, H. *et al.* Virulence and pathogenesis of SARS-CoV-2 infection in rhesus macaques: A nonhuman primate model of COVID-19 progression. *PLoS Pathog.* **16**, e1008949 (2020).
8. Hu, Y. *et al.* Prevalence and severity of corona virus disease 2019 (COVID-19): A systematic review and meta-analysis. *J. Clin. Virol.* **127**, 104371 (2020).
9. Su, L. *et al.* The different clinical characteristics of corona virus disease cases between children and their families in China—The character of children with COVID-19. *Emerg. Microbes Infect.* **9**, 707–713 (2020).
10. Yang, L. & Tu, L. Implications of gastrointestinal manifestations of COVID-19. *Lancet Gastroenterol. Hepatol.* **5**, 629–630 (2020).
11. Asadi-Pooya, A. A. & Simani, L. Central nervous system manifestations of COVID-19: A systematic review. *J. Neurol. Sci.* **413**, 116832 (2020).
12. *COVID-19 Vaccine Tracker and Landscape*. https://www.who.int/publications/m/item/draft-landscape-of-covid-19-candidate-vaccines.
13. Li, Z. *et al.* Active case finding with case management: The key to tackling the COVID-19 pandemic. *Lancet* **396**, 63–70 (2020).
14. Di Domenico, L., Pullano, G., Sabbatini, C. E., Boëlle, P.-Y. & Colizza, V. Modelling safe protocols for reopening schools during the COVID-19 pandemic in France. *Nat. Commun.* **12**, 1–10 (2021).
15. De Giorgi, V. *et al.* Naturally acquired SARS-CoV-2 immunity persists for up to 11 months following infection. *J. Infect. Dis.* **224**, 1294–1304. https://doi.org/10.1093/infdis/jiab295 (2021).
16. Dodd, R. H. *et al.* Concerns and motivations about COVID-19 vaccination. *Lancet Infect. Dis.* **21**, 161 (2021).
17. Carl Zimmer, J. C. a. S.-L. W. (New York Times, 2021).
18. Allen, H. *et al.* Household transmission of COVID-19 cases associated with SARS-CoV-2 delta variant (B. 1.617. 2): National case-control study. *Lancet Reg. Health-Eur.* **12**, 102 (2021).
19. Garcia-Beltran, W. F. *et al.* Multiple SARS-CoV-2 variants escape neutralization by vaccine-induced humoral immunity. *Cell* **184**, 2372–2383 (2021).
20. Zhou, B. *et al.* SARS-CoV-2 spike D614G change enhances replication and transmission. *Nature* **592**, 122–127 (2021).
21. Prevention, C. F. D. C. A. *SARS-CoV-2 Variant Classifications and Definitions—CDC*. https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-classifications.html (2022).
22. CDC. (The United States Centers for Disease Control and Prevention (CDC), 2021).
23. Wang, P. *et al.* Increased resistance of SARS-CoV-2 variant P. 1 to antibody neutralization. *Cell Host Microbe* **29**, 747–751 (2021).
24. Yadav, P. D. *et al.* Neutralization of Beta and Delta variant with sera of COVID-19 recovered cases and vaccinees of inactivated COVID-19 vaccine BBV152/Covaxin. *J. Travel Med.* **28**, taab104 (2021).
25. Hodcroft, E. B. *et al.* Spread of a SARS-CoV-2 variant through Europe in the summer of 2020. *Nature* **595**, 707–712 (2021).
26. Singh, R., Bhardwaj, V. K., Sharma, J., Kumar, D. & Purohit, R. Identification of potential plant bioactive as SARS-CoV-2 Spike protein and human ACE2 fusion inhibitors. *Comput. Biol. Med.* **136**, 104631 (2021).
27. Singh, R., Bhardwaj, V. K., Sharma, J., Purohit, R. & Kumar, S. In-silico evaluation of bioactive compounds from tea as potential SARS-CoV-2 nonstructural protein 16 inhibitors. *J. Tradit. Complement. Med.* **12**, 35–43 (2022).
28. Bhardwaj, V. K. *et al.* Identification of bioactive molecules from tea plant as SARS-CoV-2 main protease inhibitors. *J. Biomol. Struct. Dyn.* **39**, 3449–3458 (2021).
29. Yang, Z., Bogdan, P. & Nazarian, S. An in silico deep learning approach to multi-epitope vaccine design: A SARS-CoV-2 case study. *Sci. Rep.* **11**, 1–21 (2021).
30. Rahman, M. S. *et al.* Epitope-based chimeric peptide vaccine design against S, M and E proteins of SARS-CoV-2, the etiologic agent of COVID-19 pandemic: An in silico approach. *PeerJ* **8**, e9572 (2020).
31. Ostaszewski, M. *et al.* COVID-19 disease map, building a computational repository of SARS-CoV-2 virus-host interaction mechanisms. *Sci. Data* **7**, 1–4 (2020).
32. Moreno-Eutimio, M. A., Lopez-Macias, C. & Pastelin-Palacios, R. Bioinformatic analysis and identification of single-stranded RNA sequences recognized by TLR7/8 in the SARS-CoV-2, SARS-CoV, and MERS-CoV genomes. *Microbes Infect.* **22**, 226–229 (2020).
33. Singh, R., Bhardwaj, V. K., Das, P. & Purohit, R. A computational approach for rational discovery of inhibitors for non-structural protein 1 of SARS-CoV-2. *Comput. Biol. Med.* **135**, 104555 (2021).
34. Sharma, J. *et al.* An in-silico evaluation of different bioactive molecules of tea for their inhibition potency against non structural protein-15 of SARS-CoV-2. *Food Chem.* **346**, 128933 (2021).
35. Singh, R., Bhardwaj, V. K. & Purohit, R. Potential of turmeric-derived compounds against RNA-dependent RNA polymerase of SARS-CoV-2: An in-silico approach. *Comput. Biol. Med.* **139**, 104965 (2021).

36. Bhardwaj, V. K. *et al.* Bioactive molecules of Tea as potential inhibitors for RNA-dependent RNA polymerase of SARS-CoV-2. *Front. Med.* https://doi.org/10.3389/fmed.2021.684020 (2021).
37. Tahir ul Qamar, M. *et al.* Designing of a next generation multiepitope based vaccine (MEV) against SARS-COV-2: Immunoinformatics and in silico approaches. *PLoS ONE* **15**, e0244176 (2020).
38. Abraham Peele, K., Srihansa, T., Krupanidhi, S., Ayyagari, V. S. & Venkateswarulu, T. Design of multi-epitope vaccine candidate against SARS-CoV-2: A in-silico study. *J. Biomol. Struct. Dyn.* **39**, 3793–3801 (2021).
39. Ferrarini, M. G. *et al.* Genome-wide bioinformatic analyses predict key host and viral factors in SARS-CoV-2 pathogenesis. *Commun. Biol.* **4**, 1–15 (2021).
40. Calis, J. J. *et al.* Properties of MHC class I presented peptides that enhance immunogenicity. *PLoS Comput. Biol.* **9**, e1003266. https://doi.org/10.1371/journal.pcbi.1003266 (2013).
41. Wang, P. *et al.* Peptide binding predictions for HLA DR, DP and DQ molecules. *BMC Bioinform.* **11**, 1–12 (2010).
42. Sidney, J. *et al.* Quantitative peptide binding motifs for 19 human and mouse MHC class I molecules derived using positional scanning combinatorial peptide libraries. *Immunome Res.* **4**, 1–14 (2008).
43. Davies, N. G. *et al.* Estimated transmissibility and impact of SARS-CoV-2 lineage B. 1.1. 7 in England. *Science* **372**, eabg3055 (2021).
44. Li, Q. *et al.* The impact of mutations in SARS-CoV-2 spike on viral infectivity and antigenicity. *Cell* **182**, 1284–1294 (2020).
45. Wang, P. *et al.* Antibody resistance of SARS-CoV-2 variants B. 1.351 and B. 1.1. 7. *Nature* **593**, 130–135 (2021).
46. Weissman, D. *et al.* D614G spike mutation increases SARS CoV-2 susceptibility to neutralization. *Cell Host Microbe* **29**, 23–31 (2021).
47. Jangra, S. *et al.* SARS-CoV-2 spike E484K mutation reduces antibody neutralisation. *Lancet Microbe* **2**, e283–e284 (2021).
48. Domingo, E. & Holland, J. RNA virus mutations and fitness for survival. *Ann. Rev. Microbiol.* **51**, 151–178 (1997).
49. Xia, H. *et al.* An evolutionary NS1 mutation enhances Zika virus evasion of host interferon induction. *Nat. Commun.* **9**, 1–13 (2018).
50. Callaway, E. Making sense of coronavirus mutations. *Nature* **585**, 174–177 (2020).
51. Bekker, L.-G. *et al.* The complex challenges of HIV vaccine development require renewed and expanded global commitment. *Lancet* **395**, 384–388 (2020).
52. Burton, D. R. *et al.* HIV vaccine design and the neutralizing antibody problem. *Nat. Immunol.* **5**, 233–236 (2004).
53. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269 (2020).
54. Blanco-Melo, D. *et al.* Imbalanced host response to SARS-CoV-2 drives development of COVID-19. *Cell* **181**, 1036–1045 (2020).
55. Gao, Y. *et al.* Structure of the RNA-dependent RNA polymerase from COVID-19 virus. *Science* **368**, 779–782 (2020).
56. Chernomor, O., von Haeseler, A. & Minh, B. Q. Terrace aware data structure for phylogenomic inference from supermatrices. *Syst. Biol.* **65**, 997–1008. https://doi.org/10.1093/sysbio/syw037 (2016).
57. Dhanda, S. K. *et al.* Development of a novel clustering tool for linear peptide sequences. *Immunology* **155**, 331–345 (2018).
58. Dhanda, S. K. *et al.* Development of a strategy and computational application to select candidate protein analogues with reduced HLA binding and immunogenicity. *Immunology* **153**, 118–132 (2018).
59. Ai, J. *et al.* Omicron variant showed lower neutralizing sensitivity than other SARS-CoV-2 variants to immune sera elicited by vaccines after boost. *Emerg. Microbes Infect.* **11**, 337–343 (2022).
60. Davis, C. *et al.* Reduced neutralisation of the Delta (B. 1.617. 2) SARS-CoV-2 variant of concern following vaccination. *PLoS Pathog.* **17**, 22 (2021).
61. Dejnirattisai, W. *et al.* Reduced neutralisation of SARS-CoV-2 omicron B. 1.1. 529 variant by post-immunisation serum. *Lancet* **399**, 234–236 (2022).
62. Wilhelm, A. *et al.* Reduced neutralization of SARS-CoV-2 Omicron variant by vaccine sera and monoclonal antibodies. *MedRxiv* https://doi.org/10.1101/2021.12.07.21267432 (2021).
63. VanBlargan, L. A. *et al.* An infectious SARS-CoV-2 B. 1.1. 529 Omicron virus escapes neutralization by several therapeutic monoclonal antibodies. *BioRxiv* https://doi.org/10.1101/2021.12.15.472828 (2021).
64. VanBlargan, L. A. *et al.* An infectious SARS-CoV-2 B. 1.1. 529 Omicron virus escapes neutralization by therapeutic monoclonal antibodies. *Nat. Med.* **28**, 490–495 (2022).
65. Hwang, C.-L., Lai, Y.-J. & Liu, T.-Y. A new approach for multiple objective decision making. *Comput. Oper. Res.* **20**, 889–899. https://doi.org/10.1016/0305-0548(93)90109-V (1993).
66. Doytchinova, I. A. & Flower, D. R. VaxiJen: A server for prediction of protective antigens, tumour antigens and subunit vaccines. *BMC Bioinform.* **8**, 4. https://doi.org/10.1186/1471-2105-8-4 (2007).
67. Doytchinova, I. A. & Flower, D. R. Identifying candidate subunit vaccines using an alignment-independent method based on principal amino acid properties. *Vaccine* **25**, 856–866 (2007).
68. Paul, S. *et al.* Development and validation of a broad scheme for prediction of HLA class II restricted T cell epitopes. *J. Immunol. Methods* **422**, 28–34 (2015).
69. Paul, S., Sidney, J., Sette, A. & Peters, B. TepiTool: A pipeline for computational prediction of T cell epitope candidates. *Curr. Protoc. Immunol.* **114**, 18–19 (2016).
70. Dimitrov, I., Bangov, I., Flower, D. R. & Doytchinova, I. AllerTOP v. 2—A server for in silico prediction of allergens. *J. Mol. Model.* **20**, 1–6 (2014).
71. Larsen, M. V. *et al.* Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction. *BMC Bioinform.* **8**, 424. https://doi.org/10.1186/1471-2105-8-424 (2007).
72. Larsen, M. V. *et al.* Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction. *BMC Bioinform.* **8**, 1–12 (2007).
73. *IFNepitope*. https://webs.iiitd.edu.in/raghava/ifnepitope/index.php.
74. *IL4pred*. https://webs.iiitd.edu.in/raghava/il4pred/design.php.
75. *IL-10Pred*. https://webs.iiitd.edu.in/raghava/il10pred/index.html.
76. Manavalan, B., Govindaraj, R. G., Shin, T. H., Kim, M. O. & Lee, G. iBCE-EL: A new ensemble learning framework for improved linear B-cell epitope prediction. *Front. Immunol.* https://doi.org/10.3389/fimmu.2018.01695 (2018).
77. Almubaid, Z. & Al-Mubaid, H. Analysis and comparison of genetic variants and mutations of the novel coronavirus SARS-CoV-2. *Gene Rep.* **23**, 101064 (2021).
78. Hassan, S. S., Choudhury, P. P., Basu, P. & Jana, S. S. Molecular conservation and differential mutation on ORF3a gene in Indian SARS-CoV2 genomes. *Genomics* **112**, 3226–3237 (2020).
79. Hassan, S. S. *et al.* Emergence of unique SARS-CoV-2 ORF10 variants and their impact on protein structure and function. *Int. J. Biol. Macromol.* **194**, 128–143 (2022).
80. Kames, J. *et al.* Sequence analysis of SARS-CoV-2 genome reveals features important for vaccine design. *Sci. Rep.* **10**, 1–11 (2020).
81. Schrörs, B. *et al.* Large-scale analysis of SARS-CoV-2 spike-glycoprotein mutants demonstrates the need for continuous screening of virus isolates. *PLoS ONE* **16**, e0249254 (2021).
82. Dai, Y. *et al.* Immunodominant regions prediction of nucleocapsid protein for SARS-CoV-2 early diagnosis: A bioinformatics and immunoinformatics study. *Pathog. Glob. Health* **114**, 463–470 (2020).
83. Zhuang, S. *et al.* Bioinformatic prediction of immunodominant regions in spike protein for early diagnosis of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). *PeerJ* **9**, e11232 (2021).

84. Maitra, A. *et al.* Mutations in SARS-CoV-2 viral RNA identified in Eastern India: Possible implications for the ongoing outbreak in India and impact on viral structure and host susceptibility. *J. Biosci.* **45**, 1–18 (2020).
85. Mohammadi, M., Shayestehpour, M. & Mirzaei, H. The impact of spike mutated variants of SARS-CoV2 [Alpha, Beta, Gamma, Delta, and Lambda] on the efficacy of subunit recombinant vaccines. *Braz. J. Infect. Dis.* **25**, 101606 (2021).
86. Harvey, W. T. *et al.* SARS-CoV-2 variants, spike mutations and immune escape. *Nat. Rev. Microbiol.* **19**, 409–424 (2021).
87. Greaney, A. J. *et al.* Comprehensive mapping of mutations in the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human plasma antibodies. *Cell Host Microbe* **29**, 463–476 (2021).
88. Issa, E., Merhi, G., Panossian, B., Salloum, T. & Tokajian, S. SARS-CoV-2 and ORF3a: Non-synonymous mutations and poly-proline regions. *bioRxiv* https://doi.org/10.1101/2020.03.27.012013 (2020).
89. Mohammadi, E. *et al.* Novel and emerging mutations of SARS-CoV-2: Biomedical implications. *Biomed. Pharmacother.* **139**, 111599 (2021).
90. Mullick, B., Magar, R., Jhunjhunwala, A. & Farimani, A. B. Understanding mutation hotspots for the sars-cov-2 spike protein using shannon entropy and k-means clustering. *Comput. Biol. Med.* **138**, 104915 (2021).
91. Garrett, M. E. *et al.* High-resolution profiling of pathways of escape for SARS-CoV-2 spike-binding antibodies. *Cell* **184**, 2927–2938 (2021).
92. Ahammad, I. & Lira, S. S. Designing a novel mRNA vaccine against SARS-CoV-2: An immunoinformatics approach. *Int. J. Biol. Macromol.* **162**, 820–837 (2020).
93. Cyster, J. G., Shotton, D. M. & Williams, A. F. The dimensions of the T lymphocyte glycoprotein leukosialin and identification of linear protein epitopes that can be modified by glycosylation. *EMBO J.* **10**, 893–902 (1991).
94. Jespersen, M. C., Peters, B., Nielsen, M. & Marcatili, P. BepiPred-2.0: Improving sequence-based B-cell epitope prediction using conformational epitopes. *Nucleic Acids Res.* **45**, W24–W29 (2017).
95. Steers, N. J. *et al.* Designing the epitope flanking regions for optimal generation of CTL epitopes. *Vaccine* **32**, 3509–3516 (2014).
96. Dong, R., Chu, Z., Yu, F. & Zha, Y. Contriving multi-epitope subunit of vaccine for COVID-19: Immunoinformatics approaches. *Front. Immunol.* **11**, 1784 (2020).
97. Kalita, P., Padhi, A. K., Zhang, K. Y. & Tripathi, T. Design of a peptide-based subunit vaccine against novel coronavirus SARS-CoV-2. *Microb. Pathog.* **145**, 104236 (2020).
98. Grifoni, A. *et al.* A sequence homology and bioinformatic approach can predict candidate targets for immune responses to SARS-CoV-2. *Cell Host Microbe* **27**, 671–680 (2020).
99. Wang, P. *et al.* Peptide binding predictions for HLA DR, DP and DQ molecules. *BMC Bioinform.* **11**, 568. https://doi.org/10.1186/1471-2105-11-568 (2010).
100. Reynisson, B., Alvarez, B., Paul, S., Peters, B. & Nielsen, M. NetMHCpan-4.1 and NetMHCIIpan-4.0: Improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res.* **48**, W449–W454. https://doi.org/10.1093/nar/gkaa379 (2020).
101. Manavalan, B., Govindaraj, R. G., Shin, T. H., Kim, M. O. & Lee, G. iBCE-EL: A new ensemble learning framework for improved linear B-cell epitope prediction. *Front. Immunol.* **9**, 1695 (2018).
102. Gupta, S. *et al.* Peptide toxicity prediction. *Methods Mol. Biol. (Clifton, N.J.)* **1268**, 143–157. https://doi.org/10.1007/978-1-4939-2285-7_7 (2015).
103. Gupta, S. *et al.* In silico approach for predicting toxicity of peptides and proteins. *PLoS ONE* **8**, e73957 (2013).

Author contributions

M.A.Bz. and M.I. gathered the data and analyzed the data. M.A.Bz. and M.A.M.J. wrote the main manuscript text and M.A.Bz. prepared and designed figuers, M.A,Bz, M.I. and M.A.M.J. had designed this study. M.P. and K.B. revised the article and checked its technical issues. M.I coded Python for this study. Finally M.A.M.J and M.A.Bz proofed this aricle.

Funding

Mirza Ali Mofazzal Jahromi was financially supported by Jahrom University of Medical Sciences, Grant 99000127.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-18152-5.

Correspondence and requests for materials should be addressed to M.A.M.J. or M.P.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.