Computational Structural and Functional Analyses of ORF10 in Novel Coronavirus SARS-CoV-2 Variants to Understand Evolutionary Dynamics

Evolutionary Bioinformatics
Volume 18: 1–14

DOI: 10.1177/11769343221108218

Seema Mishra

Department of Biochemistry, School of Life Sciences, University of Hyderabad, Hyderabad, Telangana, India.

ABSTRACT

INTRODUCTION: In an effort to combat SARS-CoV-2 through multi-subunit vaccine design, during studies using whole genome and immunome, ORF10, located at the 3′ end of the genome, displayed unique features. It showed no homology to any known protein in other organisms, including SARS-CoV. It was observed that its nucleotide sequence is 100% identical in the SARS-CoV-2 genomes sourced worldwide, even in the recent-most VoCs and VoIs of B.1.1.529 (Omicron), B.1.617 (Delta), B.1.1.7 (Alpha), B.1.351 (Beta), and P.1 (Gamma) lineages, implicating its constant nature throughout the evolution of deadly variants.

AIM: The structure and function of SARS-CoV-2 ORF10 and the role it may play in the viral evolution is yet to be understood clearly. The aim of this study is to predict its structure, function, and understand evolutionary dynamics on the basis of mutations and likely heightened immune responses in the immunopathogenesis of this deadly virus.

METHODS: Sequence analysis, ab-initio structure modeling and an understanding of the impact of likely substitutions in key regions of protein was carried out. Analyses of viral T cell epitopes and primary anchor residue mutations was done to understand the role it may play in the evolution as a molecule with likely enhanced immune response and consequent immunopathogenesis.

RESULTS: Few amino acid substitution mutations are observed, most probably due to the ribosomal frameshifting, and these mutations may not be detrimental to its functioning. As ORF10 is observed to be an expressed protein, ab-initio structure modeling shows that it comprises mainly an α-helical region and maybe an ER-targeted membrane mini-protein. Analyzing the whole proteome, it is observed that ORF10 presents amongst the highest number of likely promiscuous and immunogenic CTL epitopes, specifically 11 out of 30 promiscuous ones and 9 out of these 11, immunogenic CTL epitopes. Reactive T cells to these epitopes have been uncovered in independent studies. Majority of these epitopes are located on the α-helix region of its structure, and the substitution mutations of primary anchor residues in these epitopes do not affect immunogenicity. Its conserved nucleotide sequence throughout the evolution and diversification of virus into several variants is a puzzle yet to be solved.

CONCLUSIONS: On the basis of its sequence, structure, and epitope mapping, it is concluded that it may function like those mini-proteins used to boost immune responses in medical applications. Due to the complete nucleotide sequence conservation even a few years after SARS-CoV-2 genome was first sequenced, it poses a unique puzzle to be solved, in view of the evolutionary dynamics of variants emerging in the populations worldwide.

KEYWORDS: SARS-CoV-2, ORF10, genomics, immunoinformatics, structure-function, ab initio structural model, CTL epitopes, miniproteins, mutations and evolution

RECEIVED: February 24, 2022. **ACCEPTED:** June 1, 2022.

TYPE: Original Research

FUNDING: The author received no financial support for the research, authorship, and/or publication of this article.

DECLARATION OF CONFLICTING INTERESTS: The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Seema Mishra, Department of Biochemistry, School of Life Sciences, University of Hyderabad, Hyderabad-500046, Telangana, India.
Email: seema_uoh@yahoo.com

Introduction

Novel coronavirus (SARS-CoV-2, also known previously as 2019-nCoV) is a highly contagious virus first emerging in the population in December 2019. Its infection results in Covid-19 disease with significant mortality around the world, along with the occurrence of re-infections. WHO has declared it as a pandemic with even the second, third, and fourth waves occurring[1] and there are concerted efforts toward its prevention and treatment. While vaccines based on a plethora of platforms and design strategies are available, protein subunit-based vaccines, such as Novavax, are still under development or waiting to be administered to the population at a larger and much wider scale, even though vaccines based on protein subunit platform, such as Hepatitis B vaccine based on HBsAg, have been demonstrated to have lesser side-effects. Based on the newly available SARS-CoV-2 genome sequences, studies were undertaken to design potential vaccine candidates for multi-subunit vaccine.[2-4] On the way to ranking and designing these epitopes on the basis of relevant biological pathways and immunological parameters, viz., MHC class-I and class-II antigen processing and presentation pathways and immunogenicity, several useful insights into the genomic contribution to the deadly nature of this pathogen were found along the way. One of the insights led to the crucial leads on the role of an unknown gene

sequence, called ORF10, in viral immunopathogenesis and evolution. As ORF10 was found to harbor a large number of highly promiscuous epitopes, and was not homologous to any known sequence outside SARS-CoV-2 viruses, interest in this genetic sequence gained ground. The aim of this study is to understand the structural and functional aspects of ORF10 gene-encoded protein in several variants, and its evolutionary dynamics on the basis of the predicted mutations in the key regions of its structure. A few studies specific to ORF10 have emerged only after the publication of these preprints from the author's lab at the time of first submission to ChemRxiv, preprint, March to April 2020,[2,3] and therefore, to the best of this author's knowledge, this study is the first one on analyzing the likely functional role of ORF10 based on its sequence conservation, structural analyses, and epitope compositions. Moreover, this study gains importance in view of the fact that this region is 100% conserved at nucleotide sequence level across geographical boundaries and major lineages, even in the WHO specified recent Variants of Concern (VOCs) and Variants of Interest (VOIs) including the dominant B.1.1.529, B.1.617, B.1.1.7, P.1, and B.1.351 lineages, and therefore, needs attention as to its likely functional role in evolution. The methodological approaches used in these studies are presented as a flow-chart in Figure 1. In an integrative study, utilizing phylogenetic and other Bioinformatics pipelines as well as mathematical models, the biological effects of the mutation S:T1117I on the function of the spike protein were studied.[5] It was found that B.1.1.389 lineage harbors this mutation, and it is surmised to be a positive or adaptive selection product. The strategy developed in this paper can also be used to study other specific genes like ORF10 in several different lineages. Because there is altogether no conservation of ORF10 sequence, or structure, it may be presented as a novel protein to the immune system. Further, the human body may not have been able to utilize any memory B and T cells generated against other microorganisms to target ORF10 and fight this pathogen, contributing to its deadly, contagious nature. Therefore, this study was undertaken to understand the sequence-structure-function relationship of ORF10 and the potential role it may play in viral evolution and immunopathogenesis.

Results

Open Reading Frame 10 (ORF10) is a novel expressed protein with no homologs

ORF10, an encoded protein, has been cloned and expressed.[6] Being 38 amino acids in length, it is located after the nucleocapsid region toward the C-terminal end of the SARS-CoV-2 genome. As per the GenBank sequence accession number MT106054.1 (earliest version submitted in February 2020/RefSeq NC_045512.1), the complete genome is 29 882 nucleotides in length and location of the ORF10 gene region is from 29 558 to 29 674 nucleotides. The encoded protein sequence is as follows: MGYINVFAFPFTIYSLLLCRMNSRNYI AQVDVVNFNLT. It has no known structure and function, and not much relevant information is available on it. Translation initiation signals in ORF10 have been found through ribosome footprinting indicating an expressed protein.[7] As per Swiss-Model webpage harboring the structures of all SARS-CoV-2 proteins (https://swissmodel.expasy.org/repository/species/2697049), HHblits identified zero significant templates for ORF10 and hence the model could not be built. HHblits is an iterative protein sequence search tool using HMM-HMM alignment model. It is the only protein among all the SARS-CoV-2 proteins/putative ORFs, having no homologs for building even a low quality model. BLASTp and PSI-BLAST searches against "nr" database, too, identified no homologs. This is an interesting observation, given that this sequence is located toward the C-terminal end of the SARS-CoV-2 genome. As is already known, the order of genes in RNA-virus genomes is highly conserved to ensure tight regulation and continuous gene expression.[8] Had it been located in-between anywhere else along the genome, it would have been called a DNA-like insertion sequence or a result of possible gene rearrangement.[9] Hence, its genome location shows that its occurrence may be a novel event distinct from events like standard DNA rearrangements comprising of transposition and recombination. This is also evident from the fact that it is a 38-amino acids small protein coded by 117 nucleotides in the SARS-CoV-2 genome, while transposons and insertion sequences are much longer sequences, characterized by sizes ranging from 700 bps to several thousands bps. Further, there is no possibility of occurrence of long terminal repeat sequences, since the ORF10 protein or nucleotide sequence is not a repetitive sequence, nor the sequences flanking it. This can be observed from GenBank sequence accession number MT106054.1/RefSeq sequence NC_045512.2 harboring sequences from 2 different geographical sources. Protein family search in UniProt or in MobiDB database for disordered protein function identified no known protein family. Hence, its origin is difficult to determine and needs to be further studied.

ORF10 is constant in its nucleotide sequence across SARS-CoV-2 genomes sourced from widespread geographical locations and lineages around the world

Further, it is of particular interest to note that this ORF10 nucleotide sequence (from MT106054.1/RefSeq sequence ID NC_045512.2, China, lineage A), using BLASTn tool, was found to be 100% identical to all SARS-CoV-2 ORF10 sequences from different geographical regions and major lineages (as observed from PANGO lineages, https://cov-lineages.org/), from USA (MT106054.1, lineage A), India (MT012098.1, lineage B), Turkey (MT327745.1, lineage B.4), South Korea (MT304474.1, lineage B.41), Iran (MT320891.2, lineage B.4), Taiwan (MT066175.1, lineage A), Israel (MT276597.1, lineage B), Nepal (MT072688.1,

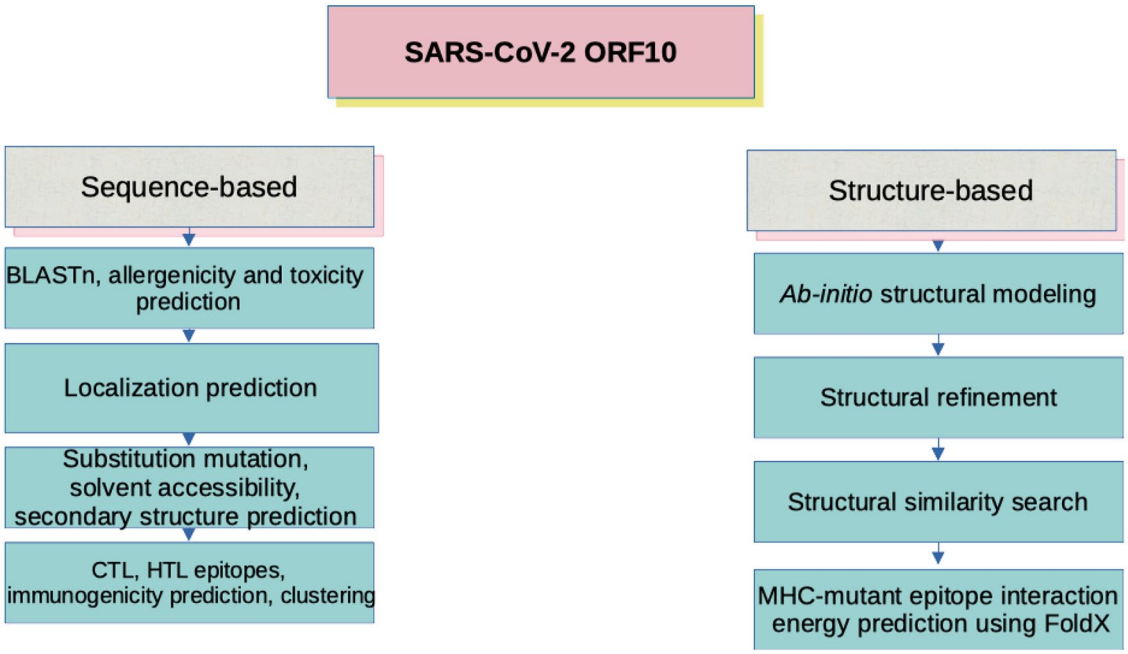


Figure 1. Schematic flowchart to depict the methodologies used in this study.

lineage B), Vietnam (MT192772.1, lineage B), China (MT291828. lineage B), Sweden (MT093571, lineage B), Greece (MT328032, lineage B.59), Italy (MT077125, lineage B), France (MT320538, lineage B.1.211), and Spain (MT292569, lineage B.1.610). Even though the newly evolving B.1.1.7 lineage of new variant harbors non-synonymous mutations in ORF1ab, spike, ORF8, and N protein (https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563) as well as several other regions in the viral genome found to harbor mutations as seen from the literature, the latest update as per NCBI Virus resource and GISAID again shows 100% identical alignment of ORF10 region in new genomes sequenced across geographical boundaries (Latest GISAID data accessed on 05-02-2022 to include Omicron variant sequences, Supplemental Table S1 shows NCBI Virus resource). Taking sequences from GISAID, this ORF10 nucleotide sequence is 100% conserved even in the recent-most B.1.1.529 (Omicron variant) and B.1.617 (Delta variant) lineages and in its sub-lineages for example, B.1.1.529 (hCoV-19/India/DL-AIIMS-COVGE27430/2021|EPI_ISL_9549954|2021-12-21, hCoV-19/South Africa/CERI-KRISP-K034200/2021|EPI_ISL_9423098|2021-12-24, hCoV-19/South Africa/CERI-KRISP-K034200/2021|EPI_ISL_9423098|2021-12-24, hCoV-19/South Africa/SU-NHLS_3942/2022|EPI_ISL_9413816|2022-01-17); B.1.617.1 (hCoV-19/India/MH-ILSGS01663/2021), and in other VOCs and VOIs such as B.1.1.7 (hCoV-19/England/SHEF-10EA7AA/2021), P.1 (hCoV-19/Brazil/PR-FUNED-411548404/2021), B.1.351 (hCoV-19/SouthAfrica/NICD-R01869/2021), and B.1.429 (hCoV-19/USA/CA-CDC-QDX24603415/2021) as well. This again points to a striking level of conservation of ORF10 nucleotide sequence in all SARS-CoV-2 genomes, interestingly, even after a year since it originated. This leads to the theory that this sequence is subjected to very tight evolutionary constraint. A strong functional role of ORF10-encoded protein in viral pathogenesis is, hence, bound to occur.

Prediction of allergenicity, toxicity, and localization of ORF10

This protein is predicted to be a possible non-allergen by AllerTOP version2 and AllergenFP version 1.0 allergenicity prediction tools. The peptides were further predicted to be non-toxic using the ToxinPred tool.

Different sub-cellular localization tools for viral proteins, Phobius (https://phobius.sbc.su.se/index.html), TMHMM (http://www.cbs.dtu.dk/services/TMHMM/), and MSLVP (https://bioinfo.imtech.res.in/manojk/mslvpred/index.php) were used to predict ORF10 localization. While Phobius and TMHMM results described it as a predicted non-cytoplasmic protein with no transmembrane helices, MSLVP described it as single-pass membrane protein in virus-infected cells, at both the 90% and 30% identity levels (Supplemental Figure S1).

Substitution mutations in full-length ORF10 are likely to occur due to ribosomal frameshifting

Mutations affect a protein's function and dynamics, and if these occur anywhere in the sequence, these may indicate the amino acid residues important or vital to ORF10 functioning. Even though there is remarkably high fidelity in nucleotide sequence conservation, substitution mutations in amino-acid sequence are likely to occur owing to ribosomal frameshifting. To gain an

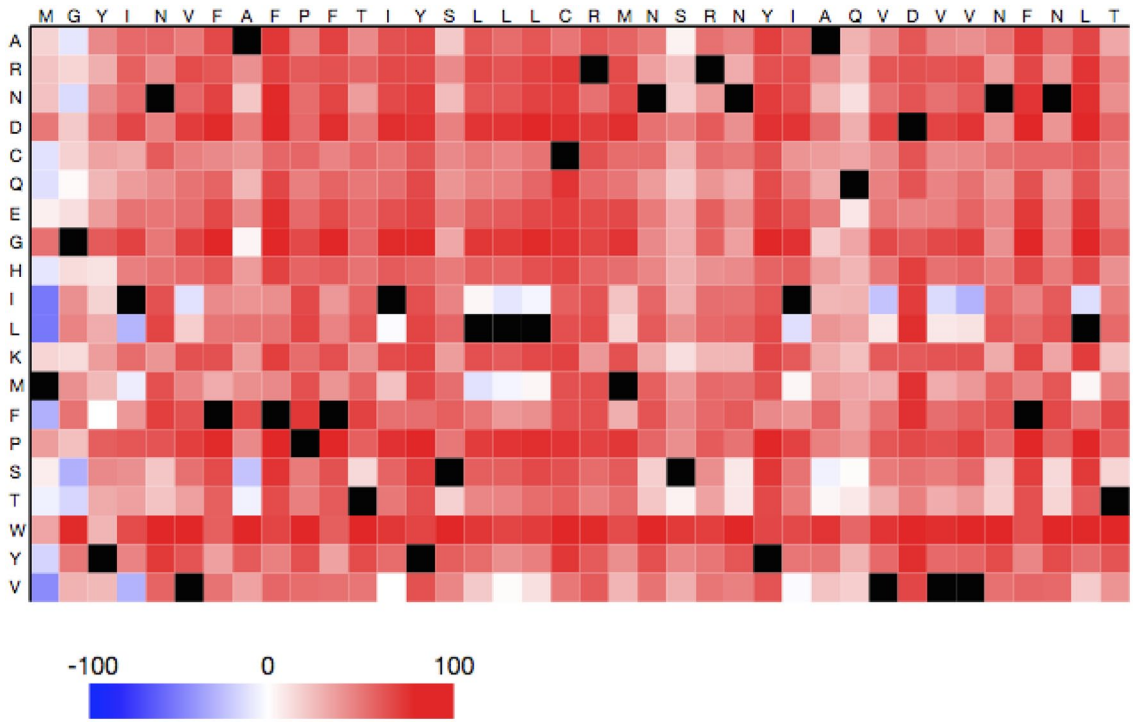


Figure 2. Heatmap depicting the predictive impact of substitution mutations for each residue of ORF10 amino-acid sequence shown at the top label of the figure. As seen from the scoring bar below the heatmap, dark red square (score >50) indicates a high score for the strong effect of a substitution mutation, white-colored square indicates weak signals (–50 <score <50), meaning that there may be an effect, and blue-colored square indicates a low score (score < 50), meaning that this substitution mutation is neutral or has no effect. Black-colored squares indicate the corresponding wild-type residues.

understanding of likely mutations disrupting ORF10 function and playing a potential role in the acceleration of evolution, if any, computational substitution mutation analysis using SNAP2 algorithm in PredictProtein was carried out. SNAP2 predicts functional effects of mutations by taking a variety of features such as evolutionary information, secondary structures, solvent accessibility among others into account.

Heatmap analyses showed that all of the amino acid residues have a higher probability of substitution mutations strongly impacting the protein's function, except those present at positions 1-3, 15, 22, 23, 25, 28-29 which have a lesser probability of impact of mutation effects (Figure 2). Therefore, these delineated regions may be required to be fully conserved in ORF10, although the effects of any protein mutations on virus biology will be required to be investigated further.

As has been noted above, from the alignments of ORF10 nucleotide sequences harvested at early time point, viz., January to March 2020 and at later months (July 2020-May 2021) across geographical boundaries (sequences from NCBI Virus and GISAID databases), rather surprisingly, it was observed that all the studied sequences harbored no changes in the ORF10 nucleotide sequences at all. The evolutionary mechanism by which Nature has retained this whole sequence is worth exploring into, since conservation of a sequence over time implies an important role in a protein's function, and thereby virus survival. It may be the case that the proofreading activity of 3′-5′-exonuclease and/or RdRp[10] is robust enough to withstand any changes, if at all occurring, in the ORF10 nucleotide sequence. Ribosomal frameshifting may be the cause of protein substitutions as observed above.

Ab-initio structural model of ORF10 consists mainly of an α-helical region

As there was no homology observed with any protein from BLASTp search, and no domains/family were predicted, the sequence was subjected to de novo/ab-initio modeling. For this, one of the best ab-initio prediction servers from Critical Assessment of Structure Prediction (CASP) experiments (number 1 in CASP10), QUARK (https://zhanglab.ccmb.med.umich.edu/QUARK/) was chosen. Structural modeling using QUARK (Figure 3a) showed that the structure mostly comprised a middle α-helical region with disordered regions at the terminal ends. Model quality assessment check using UCLA SAVES version 6.0 (https://saves.mbi.ucla.edu/) showed that ERRAT calculated the overall quality factor as 96.6%, higher than the 91% cutoff, while PROCHECK statistics for Ramachandran plot showed 91.2%, 5.9%, and 2.9% residues in the most favored, additionally allowed and generously allowed regions, respectively. This shows high structural stability of this model after refinement. Further, as observed from the link given: https://zhanglab.ccmb.med.umich.edu/COVID-19/, I-TASSER, another prediction

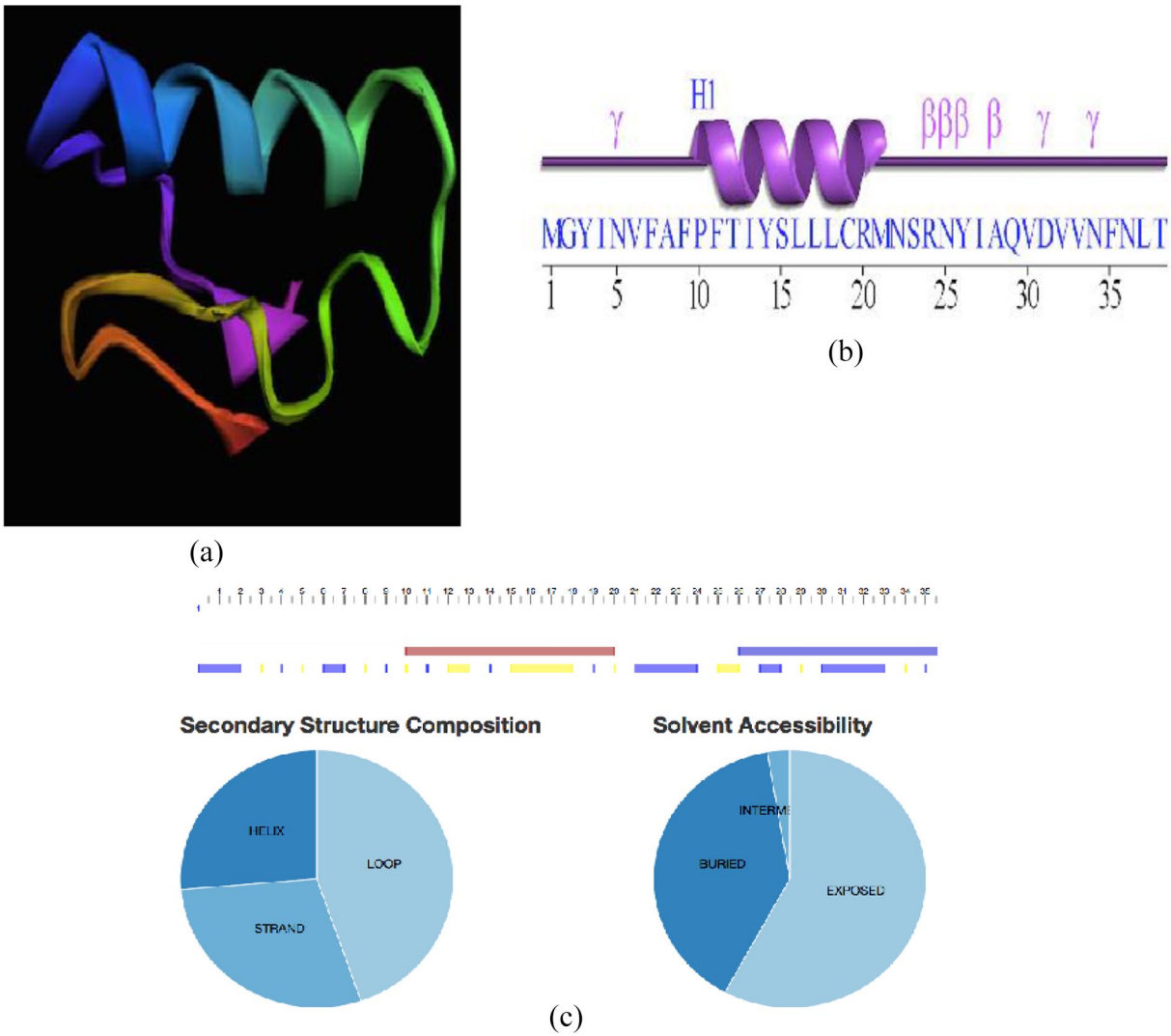


Figure 3. (a) ORF10 protein structure modeled using ab-initio modeling webserver QUARK, (b) secondary structure plot of ORF10 protein, motifs: β denotes β-turn, while γ denotes γ-turn, and (c) secondary structure composition and solvent accessibility of residues generated by PredictProtein. Orange and blue colored horizontal bars in the first panel depict helix and strand, respectively. Yellow and blue colored horizontal bars in the second panel depict buried and exposed regions, respectively. Secondary structure composition and solvent accessibility are also shown as pie-charts.

tool has been used and the structures from both QUARK and I-TASSER are found to be in congruence. This protein has been cloned and expressed alongwith other SARS-CoV-2 proteins and has also been found to harbor translation initiation signals.[6,7] Its secondary structure (Figure 3b) consists of α-helix, β-turn, and γ-turn motifs.

Secondary structure composition and solvent accessibility of residues as determined by PredictProtein also shows the same overall structural topology as generated by QUARK (Figure 3c). It should be noted that most secondary structure prediction tools do not predict turns, instead they predict only helix, sheets, and coils. QUARK can model a β-turn also, in addition to these. Therefore, a β-turn or a β-strand at this region is indicated without changing the overall topology. Solvent accessibility analysis shows that most of the residues are exposed, and residues positioned at 13 to 20 are buried.

Structural similarity search using PDBeFold (https://www.ebi.ac.uk/msd-srv/ssm/) was done for pairwise comparison and 3D alignment for similarity, if any, to other proteins in PDB. This search yielded several miniproteins/peptide fragments with fold homology. The first hit generated with highest Q-score was rat synaptotagmin-II, a membrane protein with PDB ID: 4ISR, other hits were membrane-bound as well (Supplemental Table S4). It must be noted that it is quite obvious that the proteins with only an alpha-helical fold and with a higher frequency of hydrophobic residues, will most probably be localized to the membrane, as is observed above. These observations lead to the surmise that it must be an organelle-targeted membrane-localized miniprotein. This miniprotein might function as viroporin consequent to oligomerization or modulate biological processes by binding to cellular or other viral proteins as most other viral miniproteins

do function in the same manner,[11] and are relatively easy to construct.

Open Reading Frame 10 (ORF10) may be the reason for the contagious nature of this virion

Sequences of all of the 10 SARS-CoV-2 proteins, including ORF1ab replicase complex, were used to predict promiscuous cytotoxic T lymphocyte (CTL) and helper T lymphocyte (HTL) epitopes with high immunogenicity and conserved regions across SARS family of viruses.[2-4]

During the above studies toward designing multi-subunit vaccines, upon identification of promiscuous epitopes binding to all 12 HLA-I supertypes, a total of 9621 nonameric CTL epitopes were generated across 10 SARS-CoV-2 proteins, including ORF1ab polyprotein.[4] Further analyses of these studies showed that, within our dataset selected with the criteria of promiscuity and being amongst the common and top-scoring in the results of 2 different prediction algorithms, ORF10 harbored the highest number of promiscuous epitopes, among all proteins, apart from nsp7 of ORF1ab polyprotein region (Figure 4a). These 11 out of 30 CTL epitopes generated for ORF10 were predicted to have higher TAP transporter binding, higher proteasomal cleavage as well as HLA-I binding capacity (see Mishra[2,4] for amino acid sequences of selected T cell epitopes belonging to ORF10). Further, out of these 11 epitopes, 9 epitopes were predicted to be highly immunogenic among the top ranked candidates.[2-4] Sequence analysis showed that out of all the promiscuous CTL and HTL epitopes selected in the case of ORF10,[3] most of the epitopes, either in full or in parts, belonged to the N-terminal α-helical part of the structure (Figure 4b and c). It has been widely recognized that T cell epitopes are mostly found in α-helical regions of a protein.[12,13] Further, this region has been shown to be more susceptible to substitution mutations (Figure 2) and therefore, may be essential to the ORF10 function. In contrast, spike, membrane, and nucleocapsid proteins had comparatively lower number of such epitopes in the selected list. The finding in this work (Figure 4a and Mishra[2,4]) that nucleocapsid protein harbors lower number of promiscuous, immunogenic epitopes, 3 epitopes in total, is also corroborated by recent studies that also surprisingly found only 1 HLA-I epitope from N protein from biochemical binding assays, even though it was the most abundant viral protein inside cells.[14] Of remarkable note, through these Immunoinformatics studies, 1 epitope, 269-YLQPRTFLL-277, out of only 2 spike protein's promiscuous and immunogenic epitopes that were zeroed in, was subsequently found to be the most frequently observed reactive CTL epitope, generating the strongest CD8+ T cell responses in multiple independent studies using blood samples from convalescent patients.[15-17]

Upon further ranking of the peptide epitopes, using immunogenicity screening, 9 out of the mentioned 11 CTL epitopes from ORF10 were found to be the highest in number among high scoring epitopes. HLA-II binding studies across all proteins showed that all the epitopes of ORF10 binding to key DRB1 alleles were predicted to be weak binders, although were predicted to be immunogenic.[4] HTL epitopes are required for helper T cells to boost antibody and CTL responses. In the case of ORF10, the presence of weak HTL binders may also be a contributing factor toward low neutralizing antibody levels in SARS-CoV-2 infected patients in the initial days of infection. This low level of neutralizing antibodies in initial infection days has been reported in the literature.[18,19]

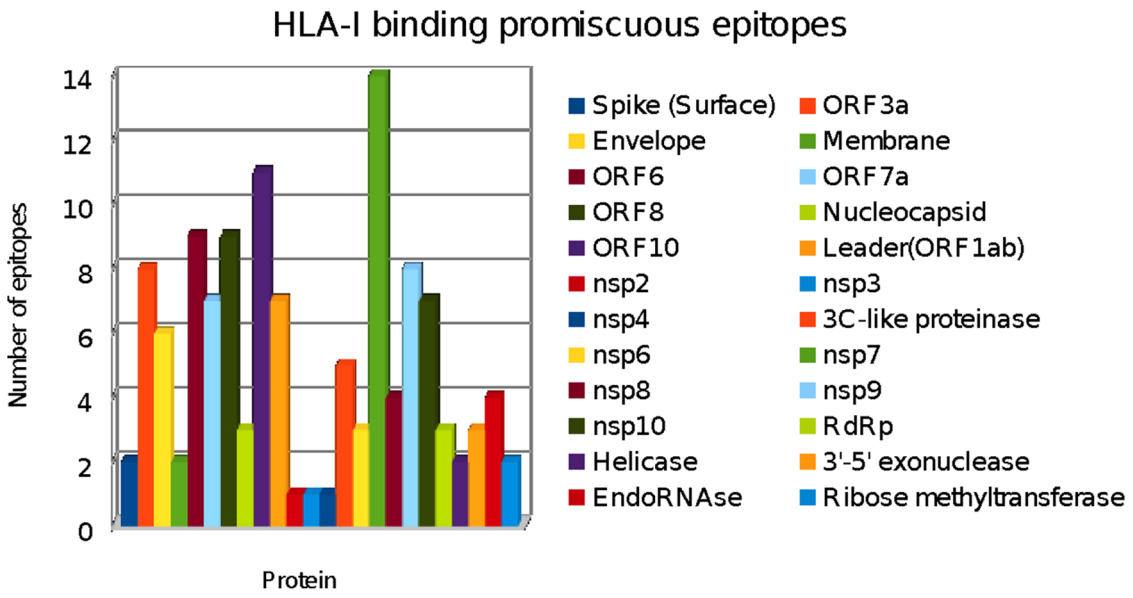
It should be noted that while the other proteins larger than ORF10 may provide more epitope sequences which may or may not be promiscuous, each and every nonameric part of ORF10 provides a promiscuous epitope (Figure 4b). This preponderance of epitopes along the complete ORF10 amino acid sequence indicates that ORF10 may play a much bigger role in contributing to immunopathology. While most of the MHC-peptide binding and T cell stimulation studies from donors (healthy and convalescent) are focused using epitopes from structural proteins of SARS-CoV-2, more attention should also be focused on non-structural proteins such as ORF10, and on their in vivo analyses.

ORF10 amino acid substitution mutations in anchor and secondary anchor residues of T cell epitopes with P10S primary anchor residue substitution occuring in the highest number of sequences

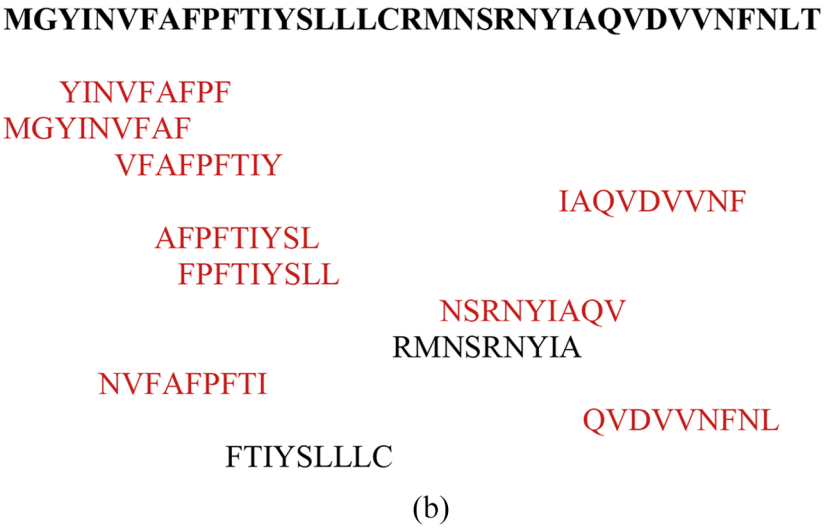
As noted above, while the nucleotide sequence of ORF10 has been strictly conserved, few amino acid substitutions have been observed, and are thought presumably to occur through ribosomal frameshifting.

Therefore, studies were designed to assess the impact, if any, of such substitutions on HLA-I binding and immunogenicity. Anchor residues at positions P2 and P9 in a nonamer peptide and secondary anchors at positions P1, P3, P5, and P7 are strongly involved in HLA-I binding,[20,21] and therefore substitutions at these positions may decrease or enhance the binding with consequent effect on immunogenicity.

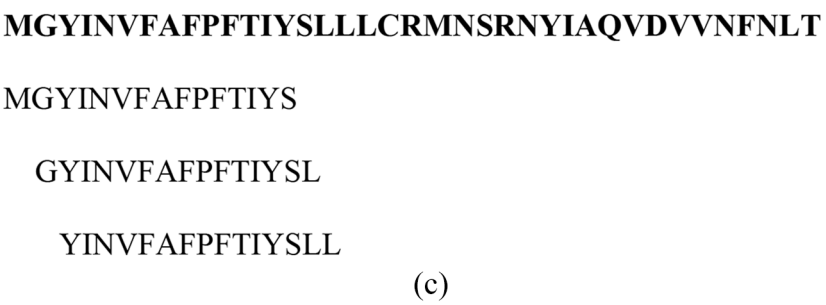
As seen from CoV-GLUE (http://cov-glue.cvr.gla.ac.uk/#/home), which is a database of amino acid variations observed in GISAID EpiCoV™ sequences, the most replaced amino acid in ORF10 was V30L, in about 135 413 sequences (Table 1 and Supplemental Table S2) followed by P10S (2693 sequences) till August 2021, both replacements also show an effect (probable impact on protein function) in SNAP2 predictions (Figure 2). The percent accuracy of its predictions can be observed from Supplemental Table S3, V30L and P10S are predicted to have an impact with 53% and 75% accuracy, respectively. Interestingly, P10S mutation is present in the sequences from the highly transmissible B.1.1.7 lineage. V30L mutation is present majorly in sequences from Europe, whereas P10S mutation is present in sequences from Europe as well as in other



(a)



(b)



(c)

Figure 4. (a) Number of promiscuous HLA-I binding epitopes across SARS-CoV-2 proteins studied. Labeling of protein names in the respective bars starts from the first column of names continuing to the next column, (b) location of selected promiscuous (11 in number) and immunogenic (9 in number, in red fonts) CTL epitopes in ORF10 amino acid sequence predicted through NetCTLpan, PickPocket, and IEDB immunogenicity prediction tools, and (c) location of selected promiscuous (observed through NetMHCIIpan analysis) and immunogenic (observed through both CD4episcore and ITcell analysis) HTL epitopes in ORF10 amino acid sequence.

Table 1. Substitution mutations in ORF10 amino acid sequence (anchor and secondary anchor positions) as taken from CoV-GLUE (http://cov-glue.cvr.gla.ac.uk/#/home), a database of amino acid variations observed in GISAID EpiCoV™ sequences.

MUTATION	NUMBER OF SEQUENCES IN WHICH THIS MUTATION IS FOUND
P10S	2693
I4V	1566
L37F	1133
S23F	1024
D31Y	1004
R24C	1003
R24L	774
A8V	764
T12M	350
P10L	331
L17P	223
F11S	170
F7L	145
I4L	132
F35S	124
M1I	115
R24H	114
A8S	97
S15G	97
L17F	89
D31N	78
I13L	73
F11L	73
I13V	70
A28S	68
I27T	67
Y14C	65
I13M	64
Y3C	61
Y26H	60
D31H	59
I4T	58
F9S	52
L37I	45
A8T	41
V33F	39
I13T	38

(continued)

Table 1. (Continued)

V30L, despite topping the number of sequences, is neither an anchor nor a secondary anchor residue mutation, so it is not included in this table which displays only such mutations.

lineages from North America, Asia, and Canada. For other mutations, readers can refer to Supplemental Table S3. These mutations may have occurred through ribosomal frameshifting, and studies on these can help improve effectiveness of antiviral preventive and treatment methods.

Taking from this analysis (Table 1 and Supplemental Table S2), the substitution mutations in ORF10 were ranked in the order of decreasing frequency and V30L (13 543 sequences) was the highest in the frequency of occurrence, although it is neither a primary nor a secondary anchor in any of the epitopes selected (Figure 4b). Among anchor residues, the topmost primary anchor residue substitution is P10S (2693 sequences) followed by I4V (1566 sequences), L37F (1133 sequences), S23F (1024 sequences), L17P (223 sequences), I13L (73 sequences), and A28S (68 sequences). D31Y (1004 sequences) and R24C (1003 sequences) were among secondary anchor substitutions.

All the topmost frequent mutations, where the number of sequences harboring these was very high, were introduced in the respective epitope sequences, and the epitopes were modeled onto MHC-I alleles to predict and compare the binding affinities through interaction energy calculations using FoldX (Table 2). Lower the interaction energy, higher is the binding strength/affinity. In terms of the largest decrease in binding affinity, L37F mutation had a major impact across the 2 MHC alleles present in DockTope. V30L mutation, although neither a primary nor a secondary anchor residue mutation, also displayed a decrease in binding affinity while S23F mutation, in contrast, led to an increase in binding affinity but no change in immunogenicity score.

Comparison of immunogenicity scores using IEB immunogenicity tool revealed that most of these top-ranked substitution mutants had lower predicted immunogenicity than the wild type (Table 2), while the scores for a few did not change after the introduction of mutation. Interestingly, in primary anchor residue mutations, at positions 2 and 9, there was no change in immunogenicity while secondary anchor residue mutations, R24C and D31Y, showed a drastic lowering of the immunogenicity score. V30L-harboring epitopes had the

Table 2. Immunogenicity ranking of wild type (in red fonts) and mutant (in black fonts) epitopes, wild type immunogenicity scores are taken from Mishra.[2-4]

NONAMER EPITOPE	IMMUNOGENICITY SCORE	FOLDX INTERACTION ENERGY (KCAL/MOL) FOR HLA-A*0201 - BOUND EPITOPES	FOLDX INTERACTION ENERGY (KCAL/MOL) FOR HLA-B*2705 - BOUND EPITOPES
YINVFAFPF	0.28259	–5.52	–10.5
Y***V***NVFAFPF (I4V)	0.28259	–8.73	–7.57
QVDVVNFNL	0.17787	–11.47	–8.01
QVDVVNFN***F*** *(L37F)*	0.17787	–3.85	–2.91
NSRNYIAQV	0.09731	1.91	–3.4
N***F***RNYIAQV (S23F)	0.09731	–3	–4.76
IAQVDVVNF	0.09546	–7.24	–10.09
IAQV***Y***VVNF (D31Y)	0.07026	–12.52	–5.99
NS***C***NYIAQV (R24C)	0.06301	–4.22	–2.3
FPFTIYSLL	0.05708	NA	–7.85
F***S***FTIYSLL (P10S)	0.05708	–8.88	–15.62
IAQ***L***DVVNF (V30L)	0.04276	–9.45	–9.3

Abbreviation: NA, not available as DockTope job failed.
Substitution mutations are in bold and italics.

lowest immunogenicity among the mutants. L37F mutation, although lowering the interaction energy, did not possess any change in immunogenicity score as compared to the wild type epitope. These observations are mostly consistent with PredictProtein heatmap analysis in Figure 2 above on the impact of mutations on protein activity.

ORF10 may function like miniproteins used in medical applications

Clustering analyses utilizing both HLA-I and HLA-II binding epitopes showed that all of those CTL epitopes from ORF10 that were high scoring in immunogenicity prediction, had higher number of clusters with HTL epitopes. This led to a high number of consensus epitopes of ORF10 amongst the proteins incorporating both CTL and HTL epitopes (Figure 5). Many epitopes belonging to other proteins were not a part of a cluster. Taken together, these analyses show that the CTL and HTL epitopes of ORF10 may be highly immunogenic. As far as in vitro immune response of ORF10 is considered, in 1 recent paper,[22] immunogenic HLA-DR T cell epitope derived from ORF10 (INVFAFPFTIYSLLL, among the dominant T cell epitopes also identified in this paper, Figure 4c) was validated as a naturally occurring T cell epitope by using in vitro amplified T cells taken from convalescent SARS-CoV-2 patients through Interferon (IFN)-γ ELISpot screening. There was no detection of ORF10 T cell epitopes by T cells from unexposed individuals. However, in another paper,[23] SARS-CoV-2 reactive CD4+ T cells against ORF10 were identified from both unexposed donors and Covid19 patients. Membrane proteins do not follow the usual antigen processing and presentation pathway. Since membrane proteins can be processed for antigen presentation only when these are bound to endoplasmic reticulum,[24] as suggested from the presence and validation of likely immunogenic epitopes and localization prediction above, it is surmised that ORF10 localizes to ER and eventually gets processed into immunogenic epitopes. Indeed, ORF10 co-localization with ER was observed from ORF10-expressing plasmid transfection and immunofluorescent assay.[25] In vivo, ORF10 may well function like those miniprotein scaffolds that are used to boost the immune response by displaying binding epitopes in medical applications.[26,27] Like these miniproteins, which are generally less than 40 to 50 amino acids in length, ORF10 also has a well defined hydrophobic core and an alpha helical segment harboring immunogenic epitopes. Hence, together with the non-conservation of ORF10 at the sequence and structural level with other organisms, and 100% conservation within SARS-CoV-2 sourced from humans across geographical locations, this may all be the reason the human body may mount a high immune response to SARS-CoV-2 resulting in immunopathological conditions and subsequent consequences.

Furthermore, there is no complete sequence conservation of ORF10 in other closely related SARS and MERS

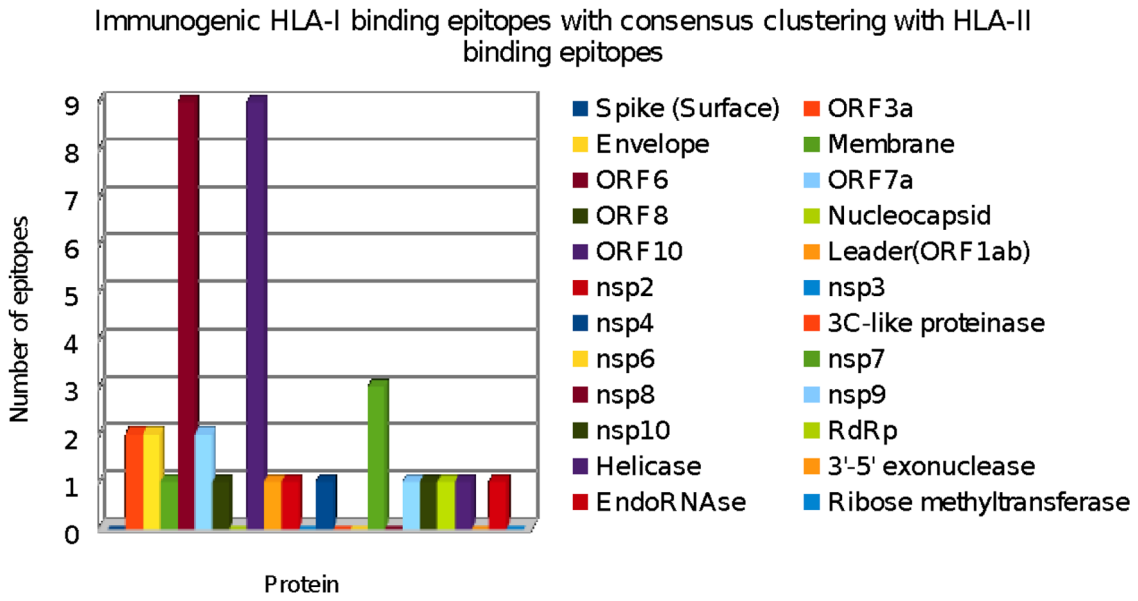


Figure 5. Number of promiscuous immunogenic HLA-I binding epitopes across SARS-CoV-2 proteins studied clustered with HLA-II binding epitopes. Labeling of protein names in the respective bars starts from the first column of names continuing to the next column.

viral species. It should be mentioned, however, that while the initial analysis in BLASTn search using default analysis did not inform anything, this author searched, found, and aligned using BLASTn "align two sequences" field, this region with a similar looking region in previous SARS-CoV sourced from human with RefSeq accession number NC_004718.3. Out of 117 nucleotides in each of the 2, there were 7 mutations (Supplemental Figure S2). Further, probing of this sequence alignment with sequence of SARS-CoV sourced from bat (accession ID: KY417142) showed the same mutations alongwith 11 more (Supplemental Figure S3).

This sequence in SARS-CoV sourced from humans may not produce a correct protein structure because it has 1 ambiguous position marked with an asterisk as seen in EMBOSS Translate Tool output. The translated amino acid sequence in SARS-CoV is MGYVNVFAIPFTIHSLLLCR MNSRN*TAQVGLVNFNLT, which is different from the ORF10 amino acid sequence from SARS-CoV-2, mentioned in the beginning of this paper.

In view of these analyses, this region may be considered an inserted tail region in the SARS-CoV-2 genome which may be responsible for the contagious nature of SARS-CoV-2. The events leading to the accumulation of such nucleotides appear to be novel/different and therefore, need to be studied in greater detail, in terms of the viral molecular evolution. Further, the human body may not have been able to utilize any memory B and T cells generated against other microorganisms to target ORF10 and fight this pathogen, contributing to its deadly, contagious nature. These key theoretical studies await further confirmation by in vitro and in vivo experiments as to the involvement of ORF10 in SARS-CoV-2 immunopathology and evolution.

Discussion

This paper discusses the possible functional role of ORF10 in SARS-CoV-2 pathogenesis. There is as yet no clear information on the likely function or regulation of ORF10, and no homology to any known protein exists in order to gain an evolutionary perspective. The findings in this paper implicate ORF10 in immunopathogenesis due to the presence of promiscuous, immunogenic CTL epitopes more frequent in number than those of the other proteins in the viral proteome across MHC alleles, at least in our dataset. ORF10, being 38 amino acids-long, may be thought of as a miniprotein, with the smallest discovered miniprotein composed of only 18 amino acid residues.[28] The function of miniproteins among several other functions, is to protect bacteria from heat and regulate/bolster the human immune system response. ORF10 is surmised to function in the same manner, with its α-helix region providing key immunogenic CTL and HTL epitopes resulting in possible immune system hyperactivation.

Structural alignments with other proteins in PDB database as well as subcellular localization predictions point to the likelihood of ORF10 being a membrane protein targeted to an organelle, ER, which has been shown experimentally. These epitope-based findings are open to further investigation in an in vitro and in vivo experimental setting and also to elucidate structure-function relationship. Collectively speaking, these theoretical studies provide key hypotheses driving the experiments to assess the biological relevance of ORF10 in viral pathogenesis and evolution of variants.

ORF10: Non-coding RNA or encoded protein?

Several contradictory papers have pointed to ORF10 either being a non-coding RNA molecule or an encoded protein. The evidence that it may not code for a protein is based on the basis of just a single fact that ORF10 is represented by only 1 read in DNB data, and it does not have sequence homology with known proteins,[29] while another paper[30] concludes that ORF10 is dispensable for SARS-CoV-2 replication and cannot be a protein coding gene, on the basis of only 2 human samples, which is insufficiently powered. Another recent paper observes through comparative genomics among sarbecoviruses, the sub-genus to which SARS-CoV-2 belongs, that ORF10 may be a non-coding transcript in sarbecoviruses because of no protein-coding constraint and may be a part of 3′-UTR.[29] In contrast to groups propounding non-coding RNA annotation, more number of papers have observed ORF10 to be an encoded protein,[6,7,22,23,31] on the basis of the following: (1) It has been cloned and expressed as a protein, and several plasmids containing ORF10 gene are commercially available (from Addgene plasmid repository). (2) Through affinity purification, mass spectrometry and immunoblotting,[31] it was found that ORF10 binds to $CRL2^{ZYG11B}$, a cellular E3 ubiquitin ligase, implying it to be an expressed protein. (3) Translation initiation signals in ORF10 have been found through ribosome footprinting indicating an expressed protein.[6] T cell responses from convalescent patients have been observed to one of the peptide epitopes of ORF10[22] and in unexposed donors as well,[23] leading to an understanding that in vivo, the peptide may be processed for presentation to T cells. One study[14] was not able to determine ORF10 epitopes in mass spectrometry analysis. In contrast, some recent studies have been able to identify ORF10 as an expressed protein, to affinity-purify ORF10 and identify its interacting partner by mass spectrometry[32] and elucidate T cell epitopes as noted above.[22,23] To be even more conclusive, observations from this paper conclude that ORF10 may constitute an immunogenic mini-protein because of its shorter length, with likely T cell epitopes being mostly in alpha-helical region, and with a well-defined hydrophobic region just like mini-protein scaffolds are. Abundance of ORF10 in vivo awaits further studies, and it is quite possible that this miniprotein due to its localization within ER, is processed rapidly into T cell epitopes, and is subjected to rapid degradation inside cells. The limitations of this study are that further experimental studies are required to corroborate these findings, in order to make these more conclusive.

Conclusions

ORF10 is 100% identical in nucleotide sequence in several SARS-CoV-2 genomes sourced all over the world and even in recent WHO-specified VOCs and VOIs. Comprising mostly of α-helix generated through ab-initio modeling, it has no known homolog. Its sequence, structure, and epitope mapping analyses indicate that its likely key function is acting as an immunogenic viroporin and in playing a role in T cell hyperactivation and concomitant immunopathogenesis.

Materials and Methods

Genome sequences

NCBI RefSeq sequences of all ten SARS-CoV-2 ORFs/proteins were retrieved. Specifically, the accession numbers were as follows: ORF10 (YP_009725255.1), nucleocapsid phosphoprotein (YP_009724397.2), ORF8 (GenBank: QID21074.1), ORF7a (YP_009724395.1), ORF6 (YP_009724394.1), membrane glycoprotein (YP_009724393.1), envelope protein (YP_009724392.1), ORF3a (YP_009724391.1), surface glycoprotein (YP_009724390.1), ORF1ab (polyprotein accession number YP_009724389.1 and the proteins therein).

Structural modeling and analyses

Ab-initio modeling conditions were applied to the ORF10 RefSeq sequence in view of no sequence or structural homology. The web server QUARK (https://zhanglab.ccmb.med.umich.edu/QUARK/,[33]) predicts a structure de novo. QUARK is one of the best ab-initio prediction servers as seen from Critical Assessment of Structure Prediction (CASP) experiments, ranked number 1 in CASP10. As per their paper,[33] an amino acid sequence is passed through several steps: multiple sequence alignment using PSI-BLAST, secondary structure prediction using PSSpred, followed by solvent accessibility, Φ and Ψ torsion angles, β-turn positions calculations. Replica-exchange Monte Carlo (REMC) simulations and refinements are then applied.

A total of 11 terms comprise the total energy in the QUARK force field, as taken from Xu and Zhang[33] as follows:

$$E_{tot} = E_{prm} + w_1E_{prs} + w_2E_{ev} + w_3E_{hb} + w_4E_{sa} + w_5E_{dh} + w_6E_{dp} + w_7E_{rg} + w_8E_{bab} + w_9E_{hp} + w_{10}E_{bp}$$

where, $w_{1,2,3...}$ are the weighting factors, and E_{prm}, E_{prs}, and E_{ev} are the atomic-level terms, E_{hb}, E_{sa}, E_{dh}, and E_{dp} are the residue-level terms, and E_{rg}, E_{bab}, E_{hp}, and E_{bp} are the topology-level terms.

Secondary structure plot was generated by PDBsum[34] and the first structural model obtained from QUARK was used as an input after refinement using ModRefiner (https://zhanggroup.org/ModRefiner/) for further analyses.

Sequence and structural analyses

PredictProtein (https://open.predictprotein.org/,[35]) was utilized to generate secondary structure information, solvent accessibility of residues and substitution mutation plots for

ORF10. Secondary structure and solvent accessibility information is generated by a neural network-based algorithm. Substitution mutation effects are calculated by SNAP2 algorithm which uses previously generated secondary structure and solvent accessibility and other information to predict the likelihood of a substitution mutation to alter or have an impact on a protein's function.

Structural similarity search using PDBeFold (https://www.ebi.ac.uk/msd-srv/ssm/) was done with the refined model generated from QUARK as an input. Subcellular localization predictions were done using ORF10 sequence as an input with 3 different prediction tools: PHOBIUS (https://phobius.sbc.su.se/index.html,[36]), TMHMM (http://www.cbs.dtu.dk/services/TMHMM/,[37]), and MSLVP (https://bioinfo.imtech.res.in/manojk/mslvpred/index.php,[38]). While PHOBIUS and TMHMM predict the presence of transmembrane helices, MSLVP predicts subcellular localization in virus-infected cells.

Cytotoxic T cell epitope prediction

Nonameric peptide epitopes were selected using NetCTLpan version 1.1 (http://www.cbs.dtu.dk/services/NetCTLpan/[39]) and PickPocket version 1.1 (http://www.cbs.dtu.dk/services/PickPocket/[40]) with default parameters. NetCTLpan neural network algorithm uses a combined version of 3 methods to identify epitopes: HLA-I binding, TAP transporter binding, and C-terminal cleavage predictions. Its prediction value is defined as a weighted sum of all 3 of these prediction values. To make these values optimal, relative weights were assigned to TAP transport efficiency and proteasomal cleavage prediction values in contrast to original values, based on the average AUC value per HLA-I bound to an epitope.

PickPocket works on the basis of position-specific weight matrices. Epitopes from NetCTLpan were ranked according to the combined score, and epitopes from PickPocket algorithm were sorted by affinity (IC_{50} values in nM). In order to increase prediction accuracy, high scoring epitopes common to both these algorithms (among top 10 in PickPocket and same epitopes among high scoring ones in NetCTLpan) were fished out. 12 HLA supertypes as present in both algorithms were used.[2-4] For ORF1ab proteins, promiscuous epitopes were selected among top 30 candidates, as not many epitopes could be found common to NetCTLpan and PickPocket results among top-scorers.

Helper T cell epitope prediction

NetMHCIIpan version 3.2 (http://www.cbs.dtu.dk/services/NetMHCIIpan/[41]) was used to predict and design helper T cell epitopes across several HLA-DRB1 alleles. Quantitative MHC-peptide binding affinity data obtained from the Immune Epitope Database is used. From a consensus list of 15 amino acids long epitopes, top ranked epitopes were sorted using descending order of percent rank. As per this tool paper,[41] those epitopes with %rank $<2\%$ and $<10\%$ are considered strong and weak binders, respectively.

Immunogenicity prediction

All the nonameric CTL epitopes were predicted for their immunogenicity using IEDB Immunogenicity tool which is validated for 9-mer epitopes.[42] Physicochemical properties of amino acids and their positions in the predicted peptide, including amino acids with large and aromatic side chains and positions 4 to 6 are used to predict potential epitopes. Ranking was done from higher to lower immunogenicity score as per the authors' guidelines.

The given formula[42] is used to calculate the immunogenicity score:

$$S(H,L)=\sum_{p=1}^{9} E_{A(L,p)} \times I_p \times M(H,p)$$

Where S represents score and L represents epitope ligand which is presented on an HLA molecule, H. This formula calculates the log enrichment score E for the amino acid at that position $A(L,p)$ for every position p in L. This is further weighted by position importance denoted by I_p and summed up.

For HTL epitopes, immunogenicity was assessed by 2 independent tools, CD4episcore, and ITcell with default parameters.[2-4]

Clustering

IEDB epitope cluster analysis tool was applied to group all HLA-I and HLA-II epitopes in clusters.[43] Minimum sequence identity threshold was 70% and cluster-break algorithm was applied.

Interaction energy calculations for substitution mutants

MHC-peptide complexes were generated using IEDB DockTope, with the provided MHC-I alleles, namely, HLA-A*0201 and HLA-B*2705, for both the wild type and mutant sequences. The complexes were then subjected to RepairPDB option of FoldX plugin in YASARA visualization tool, and thereafter, the interaction energy of molecules was calculated by FoldX AnalyseComplex tool in YASARA.[44]

All the prediction tools used in this paper and their URLs are provided in Table 3 below.

Table 3. Names, URLs, and description of prediction tools used throughout this paper.

S. NUMBER	NAME	URL	PREDICTION TOOL DESCRIPTION
Structural analyses			
1.	QUARK	https://zhanglab.ccmb.med.umich.edu/QUARK/	Ab-initio structural modeling
2.	3Drefine	http://sysbio.rnet.missouri.edu/3Drefine/	Protein structure refinement server
3.	PDBsum	http://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/pdbsum/GetPage.pl?pdbcode=index.html	Secondary structure plot
4.	PredictProtein	https://open.predictprotein.org/	Structural and mutation analyses
5.	PDBeFold	https://www.ebi.ac.uk/msd-srv/ssm/	Structural similarity search
6.	FoldX in YASARA	https://foldxyasara.switchlab.org/index.php?title=FoldX_plugin_for_YASARA	MHC-peptide interaction energy prediction
Sequence-based analyses			
7.	AllerTOP	http://www.ddg-pharmfac.net/Aller TOP	Prediction of allergenicity
8.	AllergenFP	http://ddg-pharmfac.net/Aller genFP /	Prediction of allergenicity
9.	ToxinPred	https://webs.iiitd.edu.in/raghava/toxinpred/multisubmit.php	Prediction of toxicity
10.	PHOBIUS TMHMM and MSLVP	https://phobius.sbc.su.se/index.html http://www.cbs.dtu.dk/services/TMHMM/ https://bioinfo.imtech.res.in/manojk/mslvpred/index.php	Subcellular localization prediction tools
11.	NetCTLpan version 1.1 and PickPocket version 1.1	http://www.cbs.dtu.dk/services/NetCTLpan/ http://www.cbs.dtu.dk/services/PickPocket/	CTL epitope prediction
12.	NetMHCIIpan version 3.2	http://www.cbs.dtu.dk/services/NetMHCIIpan	HTL epitope prediction
13.	IEDB	http://tools.iedb.org/immunogenicity/ http://tools.iedb.org/cluster/	Immunogenicity of CTL epitopes and clustering prediction
14.	CD4episcore	http://tools.iedb.org/CD4episcore/	Immunogenicity of HTL epitopes
15.	ITcell	http://salilab.org/itcell	Immunogenicity of HTL epitopes
16.	DockTope	http://tools.iedb.org/docktope/source.php	MHC-peptide modeling

Acknowledgements

This author acknowledges help in the form of nucleotide/protein sequences deposited in GenBank by several groups.

Author Contribution

SM: conception or design of the work; acquisition, analysis, and interpretation of data; drafted the work; wrote the final version of this paper.

ORCID iD

Seema Mishra https://orcid.org/0000-0002-4093-7899

Data Availability Statement

All supporting data are included within the main article and its Supplemental files.

Supplemental Material

Supplemental material for this article is available online.

REFERENCES

1. WHO. Coronavirus disease (COVID-19) pandemic. https://www.who.int/emergencies/diseases/novel-coronavirus-2019
2. Mishra S. T cell epitope-based vaccine design for pandemic novel coronavirus 2019-nCoV. *ChemRxiv Preprint*. 2020. doi:10.26434/chemrxiv.12029523.v2
3. Mishra S. ORF10: molecular insights into the contagious nature of pandemic novel coronavirus 2019-nCoV. *ChemRxiv Preprint*. 2020. doi:10.26434/chemrxiv.12118839.v3
4. Mishra S. Designing of cytotoxic and helper T cell epitope map provides insights into the highly contagious nature of the pandemic novel coronavirus SARS-CoV-2. *R Soc Open Sci*. 2020;7:201141. doi:10.1098/rsos.201141
5. Molina-Mora JA. Insights into the mutation T1117I in the spike and the lineage B.1.1.389 of SARS-CoV-2 circulating in Costa Rica. *Gene Rep*. 2022;27:101554.
6. Gordon DE, Jang GM, Bouhaddou M. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature*. 2020;583:459-468. doi:10.1038/s41586-020-2286-9
7. Finkel Y, Mizrahi O, Nachshon A, et al. The coding capacity of SARS-CoV-2. *Nature*. 2021;589:125-130.
8. Wertz GW, Perepelitsa VP, Ball LA. Gene rearrangement attenuates expression and lethality of a nonsegmented negative strand RNA virus. *Proc Natl Acad Sci USA*. 1998;95:3501-3506.
9. Makałowski W, Gotea V, Pande A, Makałowska I. Transposable elements: classification, identification, and their use as a tool for comparative genomics. *Methods Mol Biol*. 2019;1910:177-207.

10. Denison MR, Graham RL, Donaldson EF, Eckerle LD, Baric RS. Coronaviruses: an RNA proofreading machine regulates replication fidelity and diversity. *RNA Biol*. 2011;8:270-279.
11. DiMaio D. Viral miniproteins. *Annu Rev Microbiol*. 2014;68:21-43.
12. Berkower I, Buckenmeyer GK, Berzofsky JA. Molecular mapping of a histocompatibility-restricted immunodominant T cell epitope with synthetic and natural peptides: implications for T cell antigenic structure. *J Immunol*. 1986;136:2498-2503.
13. Yusim K, Kesmir C, Gaschen B, et al. Clustering patterns of cytotoxic T-lymphocyte epitopes in human immunodeficiency virus type 1 (HIV-1) proteins reveal imprints of immune evasion on HIV-1 global variation. *J Virol*. 2002;76:8757-8768.
14. Weingarten-Gabbay S, Klaeger S, Sarkizova S, et al. Profiling SARS-CoV-2 HLA-I peptidome reveals T cell epitopes from out-of-frame ORFs. *Cell*. 2021;184:3962-3980.e17.
15. Habel JR, Nguyen THO, van de Sandt CE, et al. Suboptimal SARS-cov-2–specific cd8$^+$ T cell response associated with the prominent HLA-A*02:01 phenotype. *Proc Natl Acad Sci*. 2020;117:24384-24391.
16. Ferretti AP, Kula T, Wang Y, et al. Unbiased screens show CD8+ T cells of COVID-19 patients recognize shared epitopes in SARS-CoV-2 that largely reside outside the spike protein. *Immunity*. 2020;53:1095-1107.e3.
17. Shomuradova AS, Vagida MS, Sheetikov SA, et al. SARS-CoV-2 epitopes are recognized by a public and diverse repertoire of human T-cell receptors. *Immunity*. 2020;53:1245-1257.
18. Wu F, Wang A, Liu M, et al. Neutralizing antibody responses to SARS-CoV-2 in a COVID-19 recovered patient cohort and their implications. *MedRxiv*. 2020. doi:10.2139/ssrn.3566211
19. Haveri A, Smura T, Kuivanen S, et al. Serological and molecular findings during SARS-CoV-2 infection: the first case study in Finland, January to February 2020. *Euro Surveill*. 2020;25:2000266.
20. Mishra S, Sinha S. Immunoinformatics and modeling perspective of T cell epitope-based cancer immunotherapy: a holistic picture. *J Biomol Struct Dyn*. 2009;27:293-305.
21. Mishra S, Sinha S. Prediction and molecular modeling of T-cell epitopes derived from placental alkaline phosphatase for use in cancer immunotherapy. *J Biomol Struct Dyn*. 2006;24:109-121.
22. Nelde A, Bilich T, Heitmann JS, et al. SARS-CoV-2-derived peptides define heterologous and COVID-19-induced T cell recognition. *Nat Immunol*. 2021;22:74-85.
23. Bacher P, Rosati E, Esser D, et al. Low-Avidity CD4+ T cell responses to SARS-CoV-2 in unexposed individuals and humans with severe COVID-19. *Immunity*. 2020;53:1258-1271.e5.
24. Bianchi F, Textor J, van den Bogaart G. Transmembrane helices are an overlooked source of major histocompatibility complex class I epitopes. *Front Immunol*. 2017;8:1118.
25. Zhang J, Cruz-cosme R, Zhuang MW, et al. A systemic and molecular study of subcellular localization of SARS-CoV-2 proteins. *Signal Transduct Target Ther*. 2021;6:192.
26. Zoller F, Haberkorn U, Mier W. Miniproteins as phage display-scaffolds for clinical applications. *Molecules*. 2011;16:2467-2485.
27. Skerra A. Engineered protein scaffolds for molecular recognition. *J Mol Recognit*. 2000;13:167-187.
28. Slavoff SA, Mitchell AJ, Schwaid AG, et al. Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nat Chem Biol*. 2013;9:59-64.
29. Kim D, Lee J-Y, Yang J-S, Kim JW, Kim VN, Chang H. The architecture of SARS-CoV-2 transcriptome. *Cell*. 2020;181:914-921.e10.
30. Pancer K, Milewska A, Owczarek K, et al. The SARS-CoV-2 ORF10 is not essential in vitro or in vivo in humans. *PLoS Pathog*. 2020;16:e1008959.
31. Jungreis I, Sealfon R, Kellis M. SARS-CoV-2 gene content and COVID-19 mutation impact by comparing 44 *Sarbecovirus* genomes. *Nat Commun*. 2021;12:2642.
32. Mena EL, Donahue CJ, Vaites LP, et al. ORF10–Cullin-2–ZYG11B complex is not required for SARS-CoV-2 infection. *Proc Natl Acad Sci*. 2021;118:e2023157118.
33. Xu D, Zhang Y. Toward optimal fragment generations for ab initio protein structure assembly. *Proteins*. 2013;81:229-239.
34. Laskowski RA, Chistyakov VV, Thornton JM. PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids. *Nucleic Acids Res*. 2005;33:D266-D268.
35. Yachdav G, Kloppmann E, Kajan L, et al. PredictProtein—an open resource for online prediction of protein structural and functional features. *Nucleic Acids Res*. 2014;42:W337-W343.
36. Käll L, Krogh A, Sonnhammer EL. Advantages of combined transmembrane topology and signal peptide prediction—the phobius web server. *Nucleic Acids Res*. 2007;35:W429-W432.
37. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol*. 2001;305:567-580.
38. Thakur A, Rajput A, Kumar M. MSLVP: prediction of multiple subcellular localization of viral proteins using a support vector machine. *Mol Biosyst*. 2016;12:2572-2586.
39. Stranzl T, Larsen MV, Lundegaard C, Nielsen M. NetCTLpan: pan-specific MHC class I pathway epitope predictions. *Immunogenetics*. 2010;62:357-368.
40. Zhang H, Lund O, Nielsen M. The PickPocket method for predicting binding specificities for receptors based on receptor pocket similarities: application to MHC-peptide binding. *Bioinformatics*. 2009;25:1293-1299.
41. Karosiene E, Rasmussen M, Blicher T, Lund O, Buus S, Nielsen M. NetMHCIIpan-3.0, a common pan-specific MHC class II prediction method including all three human MHC class II isotypes, HLA-DR, HLA-DP and HLA-DQ. *Immunogenetics*. 2013;65:711-724.
42. Calis JJA, Maybeno M, Greenbaum JA, et al. Properties of MHC class I presented peptides that enhance immunogenicity. *PLoS Comput Biol*. 2013;9:e1003266.
43. Dhanda SK, Vaughan K, Schulten V, et al. Development of a novel clustering tool for linear peptide sequences. *Immunology*. 2018;155:331-345.
44. Van Durme J, Delgado J, Stricher F, Serrano L, Schymkowitz J, Rousseau F. A graphical interface for the FoldX forcefield. *Bioinformatics*. 2011;27:1711-1712.