Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: http://www.elsevier.com/locate/compbiomed

A unique view of SARS-CoV-2 through the lens of ORF8 protein

Sk Sarif Hassan [a], Alaa A. A. Aljabali [b], Pritam Kumar Panda [c], Shinjini Ghosh [d], Diksha Attrish [e], Pabitra Pal Choudhury [f], Murat Seyran [g], Damiano Pizzol [h], Parise Adadi [i], Tarek Mohamed Abd El-Aziz [j,k], Antonio Soares [k], Ramesh Kandimalla [l], Kenneth Lundstrom [m], Amos Lal [n], Gajendra Kumar Azad [o], Vladimir N. Uversky [p], Samendra P. Sherchan [q], Wagner Baetas-da-Cruz [r], Bruce D. Uhal [s], Nima Rezaei [t], Gaurav Chauhan [u], Debmalya Barh [v], Elrashdy M. Redwan [w], Guy W. Dayhoff II [x], Nicolas G. Bazan [y], Ángel Serrano-Aroca [z], Amr El-Demerdash [aa], Yogendra K. Mishra [ab], Giorgio Palu [ac], Kazuo Takayama [ad], Adam M. Brufsky [ae], Murtaza M. Tambuwala [af,*]

[a] Department of Mathematics, Pingla Thana Mahavidyalaya, Maligram, 721140, India
[b] Department of Pharmaceutics and Pharmaceutical Technology, Yarmouk University-Faculty of Pharmacy, Irbid, 566, Jordan
[c] Condensed Matter Theory Group, Materials Theory Division, Department of Physics and Astronomy, Uppsala University, Box 516, SE-751 20, Uppsala, Sweden
[d] Department of Biophysics, Molecular Biology and Bioinformatics, University of Calcutta, Kolkata, 700009, West Bengal, India
[e] Dr. B. R. Ambedkar Centre for Biomedical Research (ACBR), University of Delhi (North Campus), Delhi, 110007, India
[f] Applied Statistics Unit, Indian Statistical Institute, Kolkata, 700108, West Bengal, India
[g] Doctoral Studies in Natural and Technical Sciences (SPL 44), University of Vienna, Austria
[h] Italian Agency for Development Cooperation - Khartoum, Sudan Street 33, Al Amarat, Sudan
[i] Department of Food Science, University of Otago, Dunedin, 9054, New Zealand
[j] Zoology Department, Faculty of Science, Minia University, El-Minia, 61519, Egypt
[k] Department of Cellular and Integrative Physiology, University of Texas Health Science Center at San Antonio, 7703 Floyd Curl Dr, San Antonio, TX, 78229-3900, USA
[l] CSIR-Indian Institute of Chemical Technology Uppal Road, Tarnaka, Hyderabad, 500007, Telangana State, India
[m] PanTherapeutics, Rte de Lavaux 49, CH1095, Lutry, Switzerland
[n] Division of Pulmonary and Critical Care Medicine, Mayo Clinic, Rochester, MN, USA
[o] Department of Zoology, Patna University, Patna, 800005, Bihar, India
[p] Department of Molecular Medicine, Morsani College of Medicine, University of South Florida, Tampa, FL, 33612, USA
[q] Department of Environmental Health Sciences, Tulane University, New Orleans, LA, 70112, USA
[r] Translational Laboratory in Molecular Physiology, Centre for Experimental Surgery, College of Medicine, Federal University of Rio de Janeiro (UFRJ), Rio de Janeiro, Brazil
[s] Department of Physiology, Michigan State University, East Lansing, MI, 48824, USA
[t] Research Center for Immunodeficiencies, Pediatrics Center of Excellence, Children's Medical Center, Tehran University of Medical Sciences, Tehran, Iran and Network of Immunity in Infection, Malignancy and Autoimmunity (NIIMA), Universal Scientific Education and Research Network (USERN), Stockholm, Sweden
[u] School of Engineering and Sciences, Tecnologico de Monterrey, Av. Eugenio Garza Sada 2501, Sur, 64849, Monterrey, NL, Mexico Tecnológico De Monterrey, Campus Monterrey, Monterrey, Nuevo León, Mexico
[v] Centre for Genomics and Applied Gene Technology, Institute of Integrative Omics and Applied Biotechnology (IIOAB), PatnaPatna, India
[w] King Abdulazizi University, Faculty of Science, Department of Biological Science, Saudi Arabia
[x] Department of Chemistry, College of Art and Sciences, University of South Florida, Tampa, FL, 33620, USA
[y] Neuroscience Center of Excellence, School of Medicine, Louisiana State University Health New Orleans, New Orleans, LA, 70112, USA
[z] Biomaterials and Bioengineering Lab, Translational Research Centre San Alberto Magno, Catholic University of Valencia San Vicente Mártir, C/Guillem de Castro 94, 46001, Valencia, Spain
[aa] Natural Products and Medicinal Chemistry Department, Institute de Chimie des Substances Naturelles, Gif-sur-Yvette, France
[ab] University of Southern Denmark, Mads Clausen Institute, NanoSYD, Alsion 2, 6400 Sønderborg, Denmark
[ac] Department of Molecular Medicine, University of Padova, Italy
[ad] Center for IPS Cell Research and Application, Kyoto University, Kyoto, 606-8397, Japan
[ae] University of Pittsburgh School of Medicine, Department of Medicine, Division of Hematology/Oncology, UPMC Hillman Cancer Center, Pittsburgh, PA, USA
[af] School of Pharmacy and Pharmaceutical Science, Ulster University, Coleraine BT52 1SA, Northern Ireland, UK

[*] Corresponding author.
E-mail address: m.tambuwala@ulster.ac.uk (M.M. Tambuwala).

https://doi.org/10.1016/j.compbiomed.2021.104380
Received 2 March 2021; Received in revised form 1 April 2021; Accepted 2 April 2021
Available online 15 April 2021

ARTICLE INFO

Keywords:
ORF8
Phylogenetics
Mutational hotspots
Physicochemical properties
ORF8 evolution
SARS-CoV-2

ABSTRACT

Immune evasion is one of the unique characteristics of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) attributed to its ORF8 protein. This protein modulates the adaptive host immunity through down-regulation of MHC-1 (Major Histocompatibility Complex) molecules and innate immune responses by surpassing the host's interferon-mediated antiviral response. To understand the host's immune perspective in reference to the ORF8 protein, a comprehensive study of the ORF8 protein and mutations possessed by it have been performed. Chemical and structural properties of ORF8 proteins from different hosts, such as human, bat, and pangolin, suggest that the ORF8 of SARS-CoV-2 is much closer to ORF8 of Bat RaTG13-CoV than to that of Pangolin-CoV. Eighty-seven mutations across unique variants of ORF8 in SARS-CoV-2 can be grouped into four classes based on their predicted effects (Hussain et al., 2021) [1]. Based on the geo-locations and timescale of sample collection, a possible flow of mutations was built. Furthermore, conclusive flows of amalgamation of mutations were found upon sequence similarity analyses and consideration of the amino acid conservation phylogenies. Therefore, this study seeks to highlight the uniqueness of the rapidly evolving SARS-CoV-2 through the ORF8.

1. Introduction

Severe acute respiratory syndrome-coronavirus-2 (SARS-CoV-2) is a novel coronavirus whose first outbreak was reported in December 2019 in Wuhan, China, where a cluster of pneumonia cases was detected. In March 11, 2020, WHO declared this outbreak a pandemic [2–5]. As of March 30, 2021, a total of 127.8 million confirmed COVID-19 cases had been reported worldwide, with 2.8 million deaths (World Health

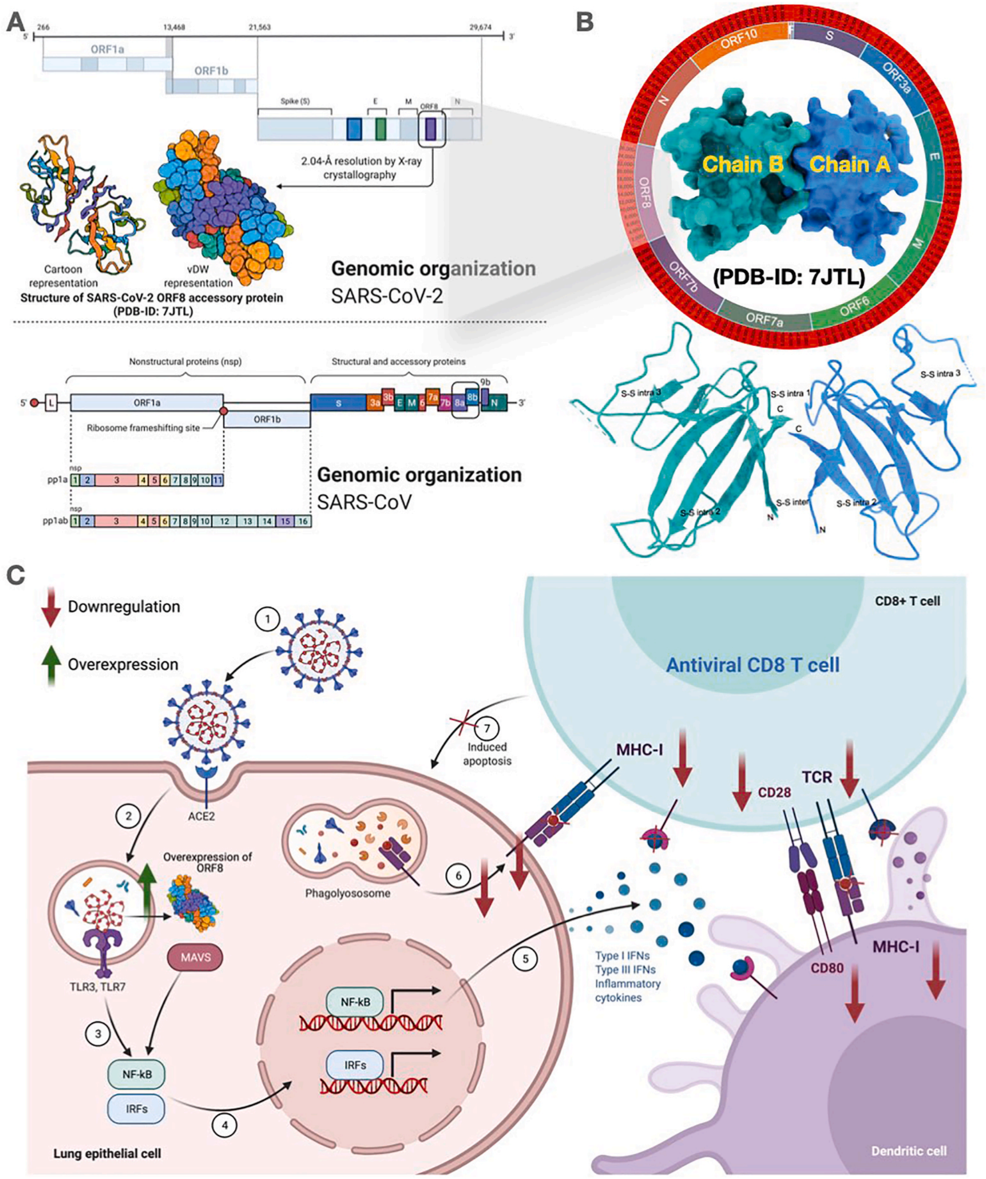

Fig. 1. Landscape of ORF8 protein. (A) Genomic organization of SARS-CoV-2 genome highlighting the ORF8 region with vdW and cartoon representation of the ORF8 structure. **(B)** The SARS-CoV-2 ORF8 protein structure (surface representation) showing 2.04 Å resolution by X-ray crystallography bearing PDB ID: 7JTL. (Below) Schematic illustration of the ORF8 dimer structure (Chain A (blue) and B (turquoise)) depicting disulfide bonds showing both intermolecular and intramolecular bond pairing. **(C)** Schematic illustration of immune invasion mechanism of ORF8 overexpression that modulates downregulation MHC-1 complex. This figure was created with Biorender.com.

Organization. Coronavirus disease 2019 (COVID-19) situation). SARS-CoV-2 belongs to the family Coronaviridae and has 55% nucleotide similarity and 30% protein sequence similarity with SARS-CoV, which caused the outbreak of SARS in 2002 [6–8]. SARS-CoV-2 is an enveloped, single-stranded RNA virus of positive polarity whose genome is approximately 30 kb in length and encodes 16 non-structural proteins, four structural, and six accessory proteins [9–11], ORF3a, ORF6, ORF7a, ORF7b, ORF8, and ORF10 (Fig. 1A) [12–16]. Among these accessory proteins, SARS-CoV-2 ORF8 is a complete protein, as it is different from any other known coronavirus ORF8 and thereby can be associated with SARS-CoV-2 pathogenicity [17,18]. The SARS-CoV-2 ORF8 displays arrays of functions; inhibition of interferon 1, promotion of viral replication, induction of apoptosis, and modulation of the ER stress [19–21].

The SARS-CoV-2 ORF8 is a 121 amino acid (aa) long protein, which has an N-terminal hydrophobic signal peptide (1–15 aa), and an ORF8 chain (16–121 aa) bearing dimer crystallography determined to 2.04 Å (PDB-ID:7JTL) (Fig. 1B) [22,23]. The functional motif (VLVVL) of SARS-CoV ORF8b, responsible for the induction of cell stress pathways and activation of macrophages, is absent from the SARS-CoV-2 ORF8 protein [24]. In the later stages of the SARS-CoV epidemic, it was found that a 29 nucleotide deletion in the ORF8 protein caused it to split into ORF8a (39 aa) and ORF8b (84aa), rendering it functionless [25]. Although such deletions have not been reported for SARS-CoV-2, a 382-nucleotide deletion variant (Δ382) was identified in Singapore and other countries, which caused the deletion of the entire ORF8 protein [26]. Patients with the Δ382 variant exhibited less severe symptoms, including milder hypoxic conditions and low cytokine activity compared to patients infected with the wildtype virus [26]. Also, the SARS-CoV-2 ORF8 functions in interspecies transmission and viral replication efficiency as the Δ382 deletion variant resulted in a reduced viral replication ability in human cells [27]. However, the SARS-CoV-2 ORF8 mainly acts as an immune-modulator by down-regulating MHC class I molecules, thereby shielding the infected cells against cytotoxic T cells, killing the target cells (Fig. 1C). Simultaneously, it is a potent inhibitor of the type 1 interferon signaling pathway, a key component of antiviral host immune response [28,29]. The ORF8 also regulates unfolded protein response (UPR) induced due to the ER stress by triggering the ATF-6 activation, thus enhancing the survivability of infected cells [30]. Since this protein impacts various host processes and develops various strategies for evading the host immune responses, it is essential to study the ORF8 mutations (natural variability) to understand better the viral infectivity and development of potent antiviral drugs against SARS-CoV-2 [31].

The present study identified a set of distinct mutations across unique variants of the SARS-CoV-2 ORF8 and classified them according to their predicted effect on the host (i.e., disease or neutral) and their consequences for protein structural stability. Furthermore, a comparison of the ORF8 protein of SARS-CoV-2 with Bat-RaTG13-CoV and Pangolin-CoV ORF8 was conducted to determine the evolutionary relationships regarding sequence similarity and originality of these paralogues. Similarly, a hydropathy and charge examination of the SARS-CoV-2 ORF8 mutations in distinct domains were executed to explore the possible effect on functionality changes. A possible flow of mutations scales simultaneously concerning the different geographical locations and chronological time has been depicted through phylogenetic analysis. Hence, validating the proposed sequence-based and amino acid conservation-based phylogeny is critical.

2. Results

2.1. Structural view of SARS-CoV-2 ORF8 protein in comparison to SARS-CoV

The SARS-CoV-2 ORF8 protein (YP 009724396) is a 121-amino-acid-long protein, which has an N-terminal hydrophobic signal peptide (1–15 aa) and an ORF8 chain (16–121 aa). Fig. S1 shows a schematic representation of ORF8 (SARS-CoV-2). In this protein, the total number of hydrophilic residues (63) was more extensive than that of the hydrophobic residues (58). The ORF8 protein of SARS-CoV-2 has only 55.4% nucleotide and 30% amino acid similarity with SARS-CoV, as shown in Fig. S2. Although the SARS-CoV-2 ORF8 has different genome characteristics, it exhibits high functional similarity with SARS-CoV ORF8ab. The ORF8 of SARS-CoV-2 consists of a 60-residue core similar to SARS-CoV-2 ORF7a (PDB-ID:6W37), with the addition of two dimerization interfaces unique to SARS-CoV-2 ORF8 (Fig. 2A). The superimposition of ORF7a, ORF8a, and ORF8b of SARS-CoV revealed significant insights into the SARS-CoV-2 ORF8 protein architecture. The Root Mean Square Deviation and (Secondary structure matching (Q score) [32] depicted a deviation of 3.206 Å, 2.301 Å, 1.007 Å and 0.078, 0.036, and 0.621 when the SARS-CoV proteins were superimposed with ORF8 protein of SARS-CoV-2. From the aforementioned analysis, SARS-CoV ORF8b showed a high degree of similarity as a greater Q score represents high similarity, whereas ORF7a and ORF8a consist of less similarity with the ORF8 of SARS-CoV-2. The SARS-CoV ORF8ab original protein possesses an N-terminal hydrophobic signal sequence, which directs its transport to the endoplasmic reticulum (ER). However, after deleting the 29 nucleotides, which splits the ORF8ab protein into ORF8a and ORF8b, only ORF8a can translocate to the ER, and ORF8b remains distributed throughout the cell. Likewise, the SARS-CoV-2 ORF8 protein also contains an N-terminal hydrophobic signal peptide (1–15 aa), which is involved in the same function. The ER has an internal oxidative environment akin to other organelles, necessary for correct protein folding and oxidation processes. Due to this oxidative environment, the formation of intra or intermolecular disulfide bonds between unpaired cysteine residues can occur as the SARS-CoV ORF8ab protein is an ER-resident protein. There are ten cysteine residues present in ORF8 of SARS-CoV, which can be involved in disulfide linkages leading to the formation of homomultimeric complexes in the ER. Similarly, the ORF8 of SARS-CoV-2 also has seven cysteines, which may be expected to form these types of disulfide linkages.

Upon inspection of the SARS-CoV-2 ORF8 protein, it was found that it consists of two domains, named D1 and D2, in which D1 consists of a signal peptide and D2 consists of the ORF8 domain. The most conserved region in the ORF8 protein of SARS-CoV-2 is “PFTINCQE” (highlighted in green) which is present in the catalytic core of the protein (Fig. 2B). The dimeric form of the protein consists of intermolecular disulfide bonds formed by Cys20 (yellow color) of each monomer (Chain A and B) (Fig. 2C). The ORF8 monomer also comprises of two-antiparallel β-sheets (smallest sheet with β2, β5, and β6 and the larger one with β3, β4, β7, and β8 where β8 is linked to β1). The dimer structure of SARS-CoV-2 ORF8 i.e., Chain A, interfaces with Chain B involving 1 disulfide bond, 4 salt bridges, 12 hydrogen bonds, and 70 non-bonded contacts. Most of the interface residues that are involved in interactions are aliphatic amino acids (grey color) followed by positive amino acids (blue color) (Fig. 2C). The SARS-CoV ORF8ab is characterized by an asparagine residue at position 81 with the Asn-Val-Thr motif responsible for the N-linked glycosylation SARS-CoV-2 ORF8 has an N-linked glycosylation site at Asn78, and its glycosylation motif is Asn-Tyr-Thr (Fig. 3A). Val77 in SARS-CoV ORF8b has been pointed out to play a critical role in the induction of the intracellular aggregation, lysosomal stress, and interleukin-mediated inflammatory responses by activating NLRP3 inflammasomes [33]. Furthermore, Val77 is conserved in SARS-CoV-2 (Wuhan strain) and the recently identified British variation SARS-CoV-2/B.1.1.7 [1,34], contributing to the pathological manifestation of infections. SARS-CoV-2/B.1.1.7 may have an evolutionary advantage over SARS-CoV-2/Wuhan based not only on antigenic changes in the spike and ORF8 proteins but also on enhanced cytokine-mediated inflammatory responses because of intracellular aggregation in host cells. The SARS-CoV-2 ORF8 is engaged in protein-protein and protein-DNA interactions, while SARS-CoV ORF8ab shows only protein-protein interactions [35]. Most conserved regions in SARS-CoV-2 ORF8 lie around the helix-coil and strand-coil junctions,

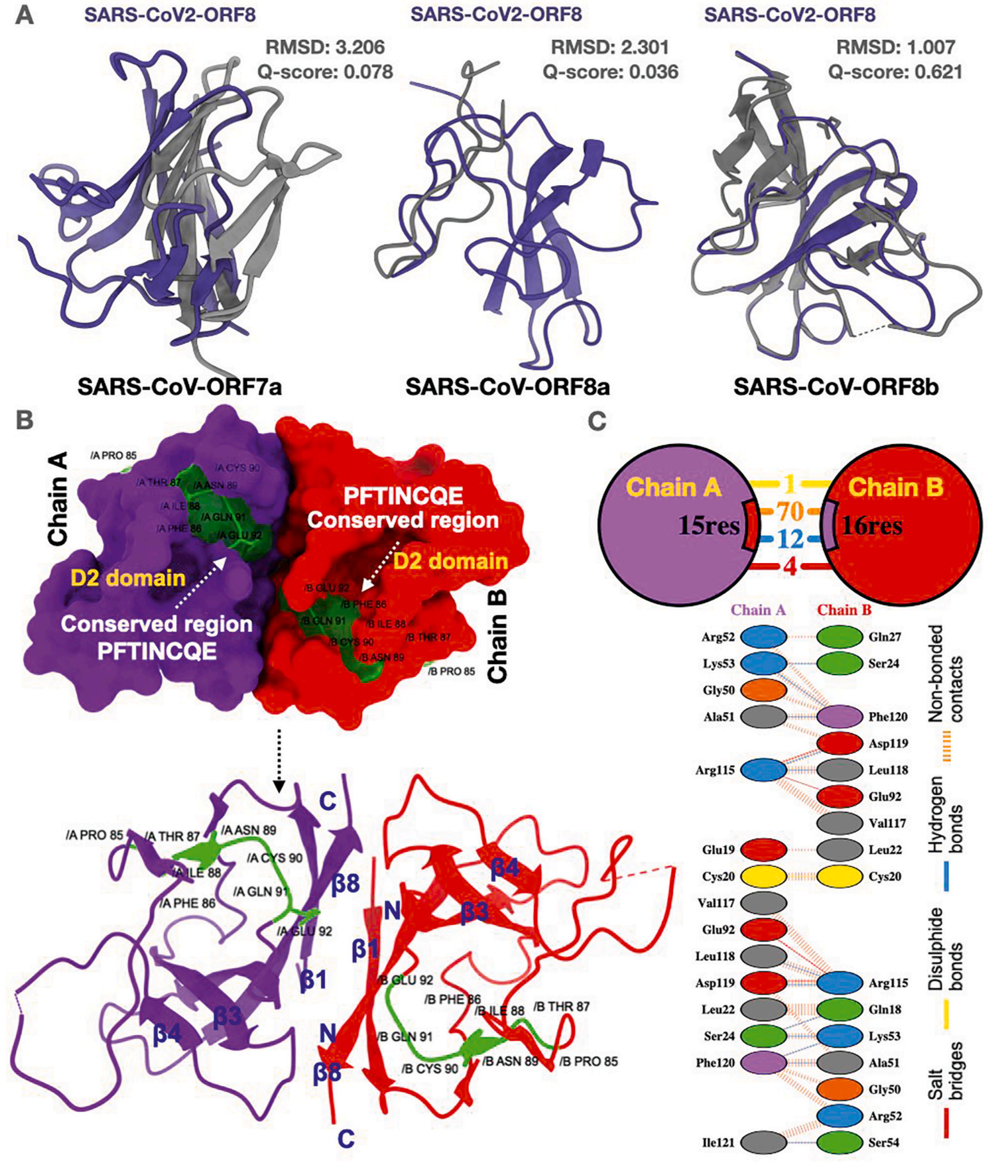

Fig. 2. Structure-based alignment of SARS-CoV-2 ORF8. (A) Superimposition of SARS-CoV-2 ORF8 with SARS-CoV ORF7a, SARS-CoV ORF8a and SARS-CoV ORF8b protein structures illustrating Q score and RMSD. **(B)** SARS-CoV-2 ORF8 surface structure bearing conserved region and D2 domain (protein dimer chains A (violet) and B (red)). (Below) Schematic illustration of the SARS-CoV-2 ORF8 structure depicting anti-parallel β sheets (β1-β8). N and C termini are labeled accordingly. The green color shows the conserved region of the SARS-CoV-2 ORF8 protein. **(C)** Interaction between the interface residues between the two chains A and B (ORF8 dimer) showing bonding patterns.

signifying these regions' functional importance. Predicted α-helical regions were also found to have conserved amino acids. It could be hypothesized that these junctions are involved in protein-protein interactions, and consequently, these regions are naturally conserved (Table S1 and Fig. S3). Our findings agree with recently published data on conserved regions in the SARS-CoV-2 ORF8 [36].

2.2. *Proximal evolutionary origin of the SARS-CoV-2 ORF8 in comparison to bat-RaTG13-CoV-ORF8 and pangolin-CoV-ORF8*

The ORF8 protein sequences of Pangolin-CoV and Bat-RaTG13-CoV were aligned against SARS-CoV-2 ORF8 amino acid sequence to understand the proximal evolutionary origin of the SARS-CoV-2 ORF8 protein. The unique ORF8 variants of Pangolin-CoV sequences (QIA48620.1, QIA48638.1, QIA48647.1, and QIQ54055.1) were aligned, which suggested that the Pangolin-CoV ORF8 protein is conserved. There are three available ORF8 sequences of Bat-RaTG13-CoV (AVP78048.1, AVP78037.1, and QHR63307.1), of which two variants (QHR63307.1 and AVP78048.1) have turned out to be characterized by 96% sequence similarity, where mutations L3F, T14A, K44R, F104Y, and V114I were embedded in the ORF8 sequences of Bat-RaTG13-CoV (Fig. 3B and Fig. S4). Furthermore, the ORF8 protein of SARS-CoV-2 is very much similar (95%) to that of Bat-RaTG13-CoV based on sequence similarity as well as phylogenetic relationships (Fig. 6). Sequence alignment indicates six amino acid differences between the ORF8 from SARS-CoV-2 and Bat-RaTG13-CoV (Fig. S5).

We have also aligned the Pangolin-CoV ORF8 (QIQ54055.1 ORF8 protein) sequence with SARS-CoV-2 (YP 009724396.1 ORF8 protein) and found that there is a sequence similarity of 88%, as depicted in Fig. S6. We observed a difference of 15 amino acid residues between the Pangolin-CoV and SARS-CoV-2 ORF8. It was established that in both the Bat-RaTG13-CoV and Pangolin-CoV ORF8 proteins, the mutations L10I, V65A, and S84L were present. So, it can be hypothesized that the SARS-CoV-2 ORF8 may have originated from Pangolin-CoV or Bat-RaTG13-CoV ORF8.

In terms of structural alignment, when the SARS-CoV-2 ORF8 protein structure was superimposed against Bat-RaTG13-CoV and Pangolin-CoV ORF8, it was observed that both Bat-RaTG13-CoV and Pangolin-CoV ORF8 structure contains a small helix which was not observed in the SARS-CoV-2 ORF8 structure (PDB-ID:7JTL) (Fig. 4A). The helix region's difference consists of Val49, Gly50, and Ala51, which has evolved as a beta-hairpin structure in the SARS-CoV-2 ORF8 structure. The amino

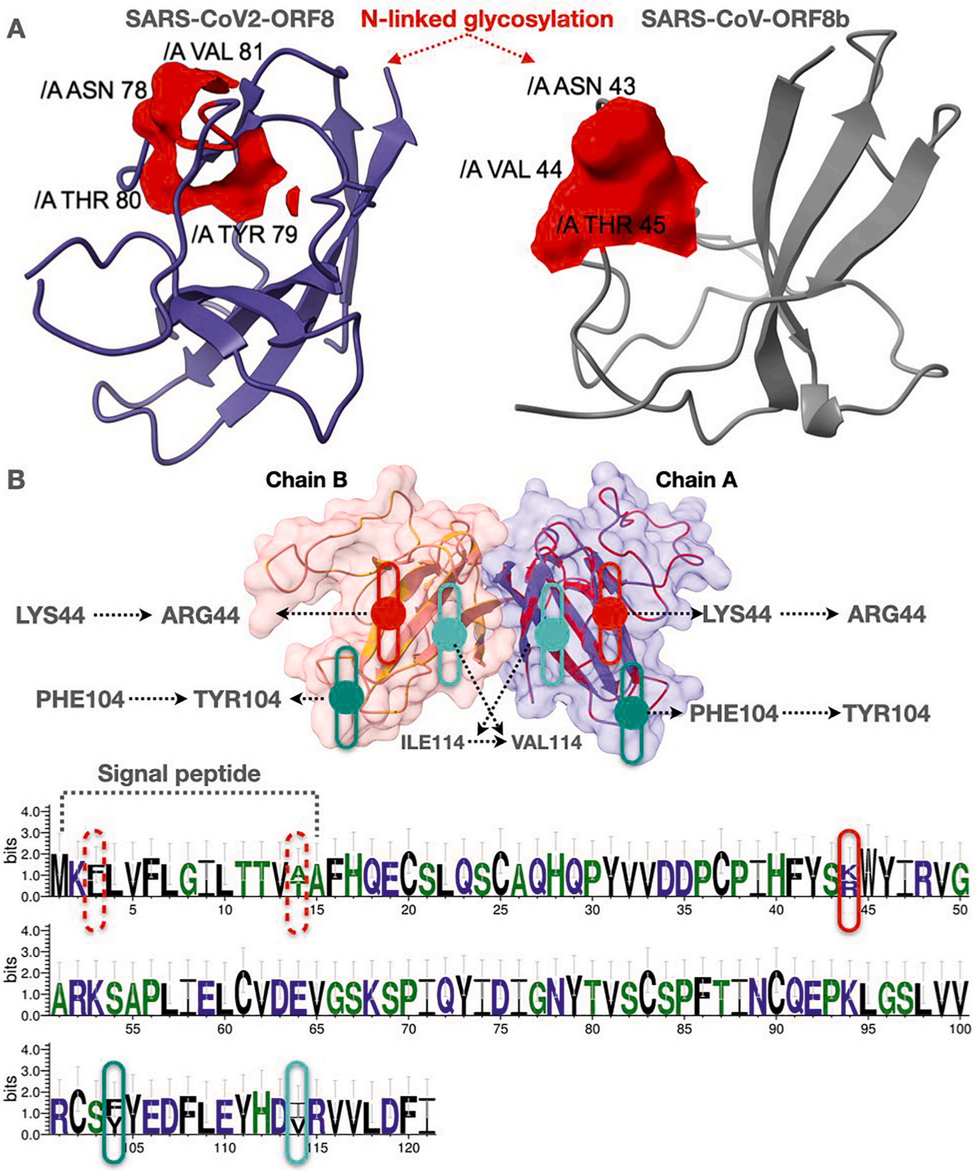

Fig. 3. (A) SARS-CoV-2 ORF8 monomer and SARS-CoV ORF8b showing N-linked glycosylation sites analyzed through NetNGlyc 1.0. The N-linked glycosylation sites are marked red. **(B)** Structural alignment of two ORF8 sequences (116 among 121 residues was identical) of Bat-RaTG13-CoV QHR63307.1 and AVP78048.1, illustrating mutations at particular sites. The same presentation using Web Logo server.

acid difference between all three sequences depicted in Fig. 4B shows that the conserved region PFTINCQE has been conserved in all three species. Comparison of the secondary structure of ORF8 proteins from SARS-CoV-2, Bat-RaTG13-CoV, and Pangolin-CoV indicates changes at four different locations (Table S2). From Table S2, it is inferred that the secondary structures of ORF8 (SARS-CoV-2) and Bat-RaTG13-CoV are closely related compared to the ORF8 of Pangolin-CoV. Based on the sequence alignment, the ORF8 of SARS-CoV-2 differs substantially from the ORF8 of Pangolin-CoV in terms of a greater number of amino acid differences (mutations). It can be hypothesized that the SARS-CoV-2 ORF8 protein is using ORF8 of Bat RaTG13-CoV as a blueprint of its structure.

2.3. *Evaluating the propensity of various ORF8 proteins for intrinsic disorder*

The differences between the ORF8 of SARS-CoV-2, Bat-RaTG13-CoV, and Pangolin-CoV can be further demonstrated by analyzing the per-residue intrinsic disorder predispositions of these proteins. Results of this analysis shows the results in Fig. 4C, which illustrates that the intrinsic disorder propensity of the ORF8 from SARS-CoV-2 is closer to that of the ORF8 from Bat-RaTG13-CoV than to the disorder potential of ORF8 from Pangolin-CoV. This agrees with the results of the analyses mentioned above conducted in this study. Because SARS-COV-2 ORF8 is closer to Bat-RaTG13-CoV, we then analyzed the variants of SARS-CoV-2 ORF8 itself (96 variants) to shed light on the possible effect of mutations on disorder profiles within the variants of ORF8. When all the 96 variant sequences were aligned, it was observed that 6.6% of the region is 100% evolutionary conserved across all 96 distinct variants of the 121-amino-acid-long ORF8 protein, including the largest conserved region PFTINCQE' (in D2 domain) (Fig. S2) in the ORF8 protein of SARS-CoV-2. The intrinsic disorder profile analysis revealed that the intrinsic disorder predispositions could vary significantly, especially in highly and moderately flexible regions (among 96 variants) (Fig. 4D). Although many mutations are disorder-silent, some increase the local disorder propensity, whereas others cause a noticeable decrease in disorder

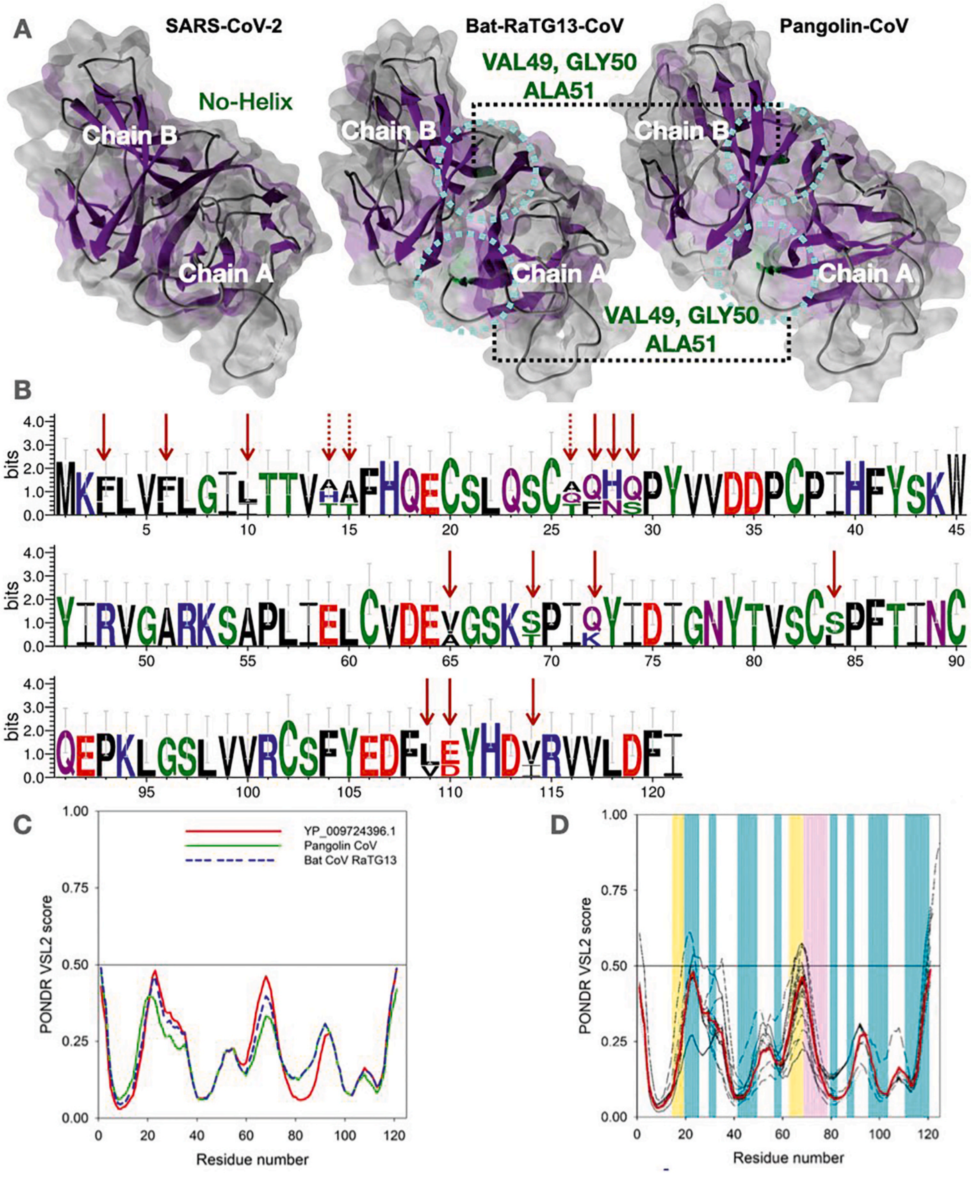

Fig. 4. Comparative sequence and structural analysis of SARS-CoV-2, Bat-CoV-RaTG13, and Pangolin-CoV ORF8. **(A)** Secondary structure analysis of ORF8 protein structures. Rounded circle represents the helix region (green color). The β sheets are illustrated using ChimeraX (violet color). **(B)** Web logo presentation of the species' aligned sequences mentioned above of ORF8 amino acid sequences depicting the mismatches (arrows). The dotted arrows indicates all mismatches across the aligned sequences. **(C)** Comparison of the intrinsic disorder predisposition of the reference ORF8 protein (YP 009724396.1) of the NC 045512 SARS-CoV-2 genome from Wuhan, China (bold red curve) with disorder predispositions of ORF8 from the Pangolin-CoV (QIA48620.1) and Bat-RaTG13-CoV (QHR63307.1). **(D)**. Analysis of the intrinsic disorder predisposition of the unique variants of the SARS-CoV-2 ORF8 in comparison with the reference ORF8 protein (YP 009724396.1) from the NC 045512 SARS-CoV-2 genome from Wuhan, China (bold red curve). Analysis was conducted using the PONDR® VSL2 algorithm [37], one of the more accurate standalone disorder predictors [37–40]. A disorder threshold is indicated as a thin line (at score = 0.5). Residues/regions with disorder scores >0.5 are considered as disordered. Light cyan vertical bars represent positions of β-strands, whereas light pink and light-yellow vertical bars show regions with missing electron density in the crystal structures of the ORF8 protein from SARS-CoV-2 (PDB ID: 7JILB and 7JX6, respectively).

predisposition. For example, local disorder predisposition in the vicinity of residue 20 was increased in the variants QKV39247.1 and QLC47867.1. Variants QMT50144.1 and QLH01532.1 have a prominent new peak in the vicinity of residue 30, where most other variants have a shoulder. Although variant QJS56890.1 also has a prominent peak in residue 30 and has one of the highest peaks in the vicinity of residue 50, the intensity of its peak in the vicinity of residue 20 is noticeably decreased. Variant QMT96539.1 showed higher disorder propensity in the vicinity of residue 110. Variants QMT49652.1 and QMT54388.1 have the lowest disorder predisposition in residue 50, whereas the lowest disorder propensity in the vicinity of residue 70 is found in variants QKV06506.1 and QKQ29929.1. Finally, although variant QMU91370.1 is almost indistinguishable from variant QJS56890.1 within the first 40 residues, its intrinsic disorder predisposition in the vicinity of residue 70 is one of the lowest among all proteins analyzed in this study. Comparison of Fig. 4C and D shows that the variability in the disorder predisposition between many variants of the ORF8 protein from SARS-CoV-2 isolates is noticeably greater than that between the reference ORF8 from SARS-CoV-2 and ORF8 proteins from Bat-RaTG13-CoV and Pangolin-CoV.

One could ask how these disorder predictions would correlate with the actual structure of the ORF8 protein from SARS-CoV-2. Recently, two crystal structures of this protein's dimeric form were reported, PDB ID: 7JTL and PDB ID: 7JX6 [41] Fig. S7 represents the results of multiple structure alignments of four ORF8 chains found in these structures, chains A and B of 7JTL and chains A and B of 7JX6. Despite high overall structural similarity (87 residues are aligned with the root mean square deviation, RMSD, of 0.57 Å), Fig.S7 shows that several loop regions are characterized by relatively high structural flexibility, with the highest flexibility being found in the 63–78 loop. Curiously, this long loop is characterized by high structural plasticity, being very differently present in different ORF8 chains. In fact, in the 7JX6 dimer, it is entirely missing in the chain B structure but exists as a loop with two short β-strands (residues 61–63 and 68–70) in the chain A structure. In the 7JTL dimer, residues 63–68 are missing. Note that residues 15–18 are missing in this structure as well. Fig. 4D provides an outlook of the correlation between the intrinsic disorder predisposition of ORF8 protein and its structure. It is seen that most stable secondary structure elements (8 β-strands shown by light cyan vertical bars) are preferentially located within regions with low intrinsic disorder propensity, whereas the structurally mobile 63–78 loop is predicted to be highly flexible.

The conformational flexibility of the ORF8 protein could be responsible for or mediating unique immune suppression and immune-evasion capabilities of SARS-CoV-2, which may contribute to the high transmissibility and vigorous pathogenesis of this virus [42]. Although at early stages of pandemics, the ORF8 protein was shown to be an immunogenic secreted protein that induces neutralizing antibodies and can be utilized for the accurate diagnosis of COVID-19 [43], the presence of multiple ORF8 mutations requires a systematic analysis of its peptide map to determine the effects of these mutations on the neutralization potential of the ant-ORF8 antibodies, which may have therapeutic and/or diagnostic values.

2.4. Shedding light on physicochemical properties of ORF8 across SARS-CoV-2, bat-RaTG13-CoV, and Pangolin-CoV

As the analyses mentioned above showed the similarity profile based on sequence and structure alignments, awareness of the physicochemical properties of the SARS-CoV-2 ORF8 protein is required to understand the composition of these viral proteins to develop subunit vaccines or for designing drugs targeting these specific proteins [44]. Physicochemical analysis revealed that the total number of hydrophilic residues in the SARS-CoV-2 ORF8 protein was higher than that of the hydrophobic residues [24]. However, the predicted secondary structure and solvent accessibility analysis (Fig. S8) indicated that the highest solubility score for this protein is four, indicating that although hydrophilic residues are higher in number, they are insufficient to ensure high protein solubility. Fig. S8 shows the predicted secondary structure and solvent accessibility of the ORF8 proteins of SARS-CoV-2, Bat-RaTG13-CoV, and Pangolin-CoV obtained using the ab-initio web server QUARK to perceive the differences. The frequencies of the hydrophobic, hydrophilic, and charged amino acids were compared among the four ORF8 proteins of SARS-CoV, SARS-CoV-2, Bat-RaTG13-CoV, and Pangolin-CoV. As seen in Table S3, SARS-CoV-2 ORF8, Bat-RaTG13-CoV ORF8, and Pangolin-CoV ORF8 are all similar in terms of hydrophobicity and hydrophilicity, and it is known that hydrophobicity and hydrophilicity play an essential role in protein folding, which determines the tertiary structure of the protein and thereby affects the functions of ordered proteins. The ORF8 sequences of SARS-CoV-2, Bat-RaTG13-CoV, and Pangolin-CoV have almost the exact content of positive and negative charged amino acids. Therefore, we can hypothesize that these proteins probably have similar electrostatic and

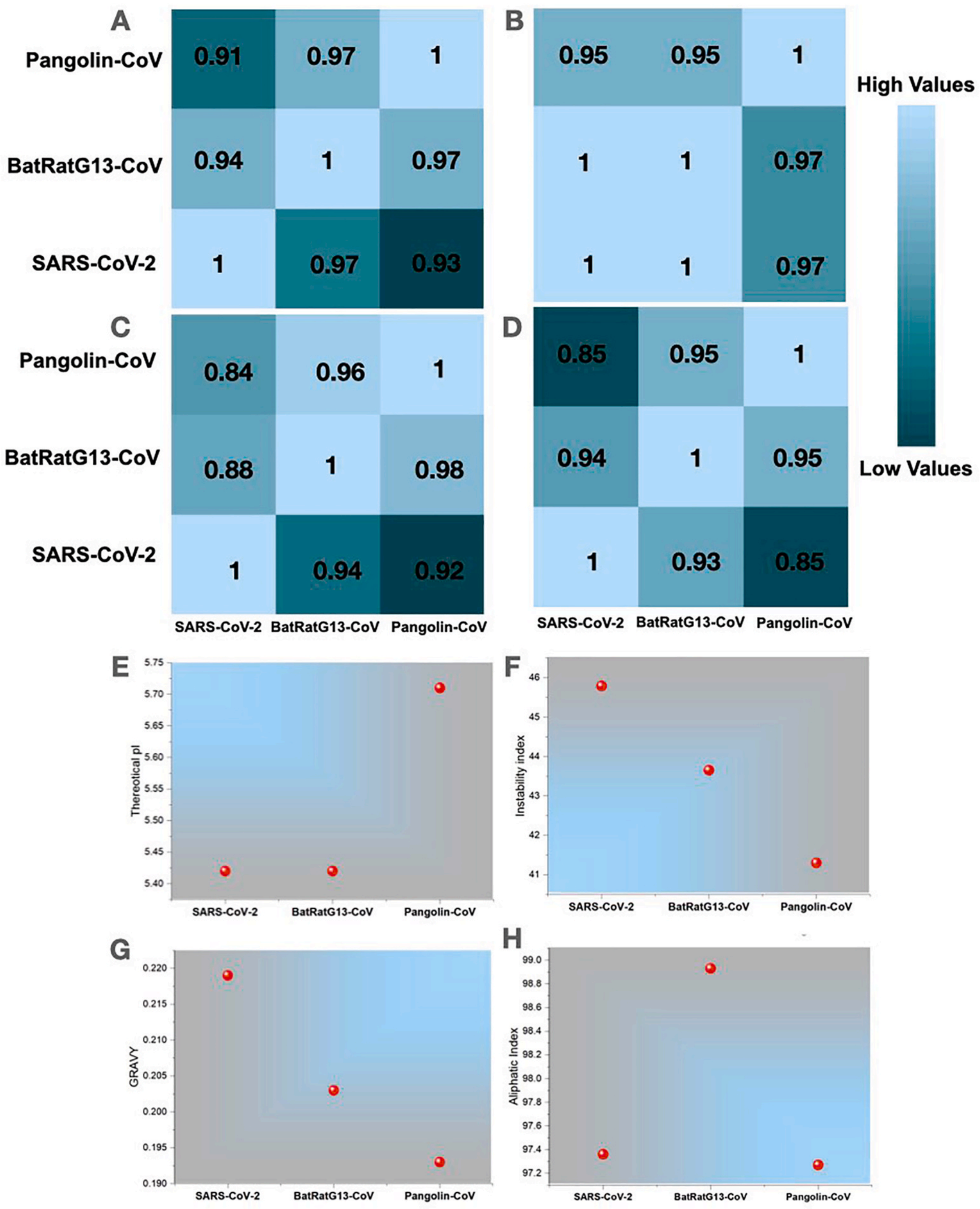

Fig. 5. Comparison of global biophysical properties of ORF8 proteins of SARS-CoV-2, Bat-RaTG13-CoV, and Pangolin-CoV. (A–D) A pairwise matrix of correlation coefficients between SARS-CoV-2 ORF8, Bat-RaTG13-CoV ORF8, and Pangolin-CoV ORF8 has been illustrated based on default parameters, the Pearson correlation coefficient R and the Spearman rank correlation coefficient using VOLPES in terms of the level of similarity between physicochemical properties i.e., hydrophobicity, pH7.0, beta propensity, and relative mutability, respectively. High values (cyan) and lower values (green) color represent correlation coefficient values. The individual panel across rows was compared against each other, and the coefficients were calculated. Similarly, each panel across rows was calculated. For instance, in Fig. 5A, R = 1 represents the hydrophobicity correlation of SARS-CoV-2 ORF8 with SARS-CoV-2 ORF8. Similarly, R = 0.97 represents the similarity between SARS-CoV-2 ORF8 with BatRatG13-CoV ORF8 across the same row. **(E–H)** Physicochemical properties, i.e., Theoretical pI, Instability Index, GRAVY, and Aliphatic index of ORF8, were calculated using ProtParam, respectively.

hydrophobic interactions, contributing to their functionality. Again, for the SARS-CoV ORF8ab, it was found that the number of positively and negatively charged amino acids are similar to those of the SARS-CoV-2 ORF8. Although the SARS-CoV sequence bears less similarity with the SARS-CoV-2, these proteins are likely similar in terms of electrostatic and hydrophobic interactions as well.

Moreover, a pairwise matrix of correlation coefficients between SARS-CoV-2 ORF8, Bat-RaTG13-CoV ORF8, and Pangolin-CoV ORF8 has been illustrated based on default parameters, the Pearson correlation coefficient R and the Spearman rank correlation coefficient using VOLPES [24] in terms of the level of similarity between physicochemical properties i.e., hydrophobicity [45] (Fig. 5A), pH7.0 (Fig. 5B), beta propensity (Fig. 5C) and relative mutability [46] (Fig. 5D). The analysis revealed that in most properties, SARS-CoV-2 ORF8 has a high correlation with BatRatG13-CoV ORF8 protein and shows less correlation with Pangolin-CoV ORF8. However, BatRatG13-CoV ORF8 and Pangolin-CoV ORF8 share an excellent correlation. Since we have previously explored the amino acid changes based on multiple sequence alignments, the relative mutability factor may shed light on the similarity correlation between SARS-CoV-2 and BatRatG13-CoV, proving that SARS-CoV-2 is closer to BatRatG13-CoV.

Furthermore, we analyzed three ORF8 sequences and checked their molecular weights, isoelectric points (pIs) (Fig. 5E), Instability index (Fig. 5F), hydropathy (GRAVY) (Fig. 5G), Aliphatic Index (Fig. 5H), net charge, and extinction coefficient using a peptide property calculator (https://pepcalc.com/) (Fig. S9) and ProtParam. We found that all properties are almost similar between the SARS-CoV-2 and the BatRatG13-CoV ORF8 protein. While inspecting the chemical aspects, e.g., molecular weight, the ORF8 protein of Bat-RaTG13-CoV, Pangolin-CoV, and SARS-CoV-2 are very closely related based on their chemical aspects of amino acid residues (Fig. S9). The isoelectric point (pI) and the protein's molecular weight tell us about the protein's biochemical and functional aspects. Since the ORF8 sequences of Bat-RaTG13-CoV and SARS-CoV-2 have the same pI and molecular weights, they can be grouped under a single functional header. The pI of the Pangolin-CoV ORF8 is higher than that of the SARS-CoV-2 ORF8, indicating that the ORF8 of the Pangolin-CoV is more negatively charged than the SARS-CoV-2 ORF8.

2.5. Natural variants of SARS-CoV-2 ORF8 protein

Each of the ORF8 amino acid sequences (96 variants) was aligned concerning the ORF8 protein (YP 009724396.1) from Wuhan, China, using the multiple sequence alignment tools (NCBI Blastp suite), and the corresponding results were used to identify mutations and their associated positions [47]. It is noted that a mutation from an amino acid A1 to A2 at a position p is denoted by A1pA2 or A1(p)A2. Fig. 6 and Table S4 describe various mutations with their respective locations. The missense mutations were found within the entire ORF8 sequence starting from the amino acid position 3 to 121, and some insertion mutations occurred at

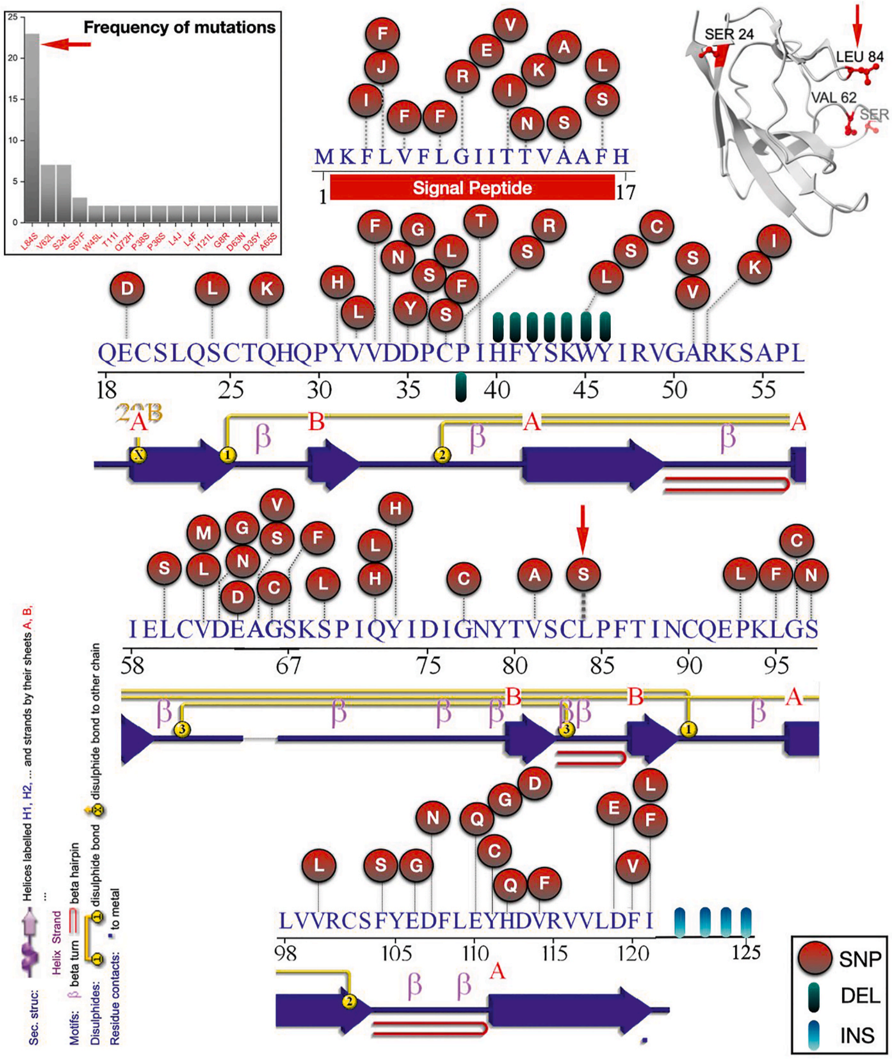
Frequency of mutations
25
20
15
10
5
0
L84S
V62L
S24L
S67F
W45L
T11I
Q72H
P38S
P36S
L4J
L4F
I121L
G8R
D63N
D35Y
A65S
SER 24
LEU 84
VAL 62
SER 67
MKFLVFLGIITTVAAFH
1
Signal Peptide
17
QECSLQSCTQHQPYVVDDPCPIHFYSKWYIRVGARKSAPL
18
25
30
35
40
45
50
55
IELCVDEAGSKSPIQYIDIGNYTVSCLPFTINCQEPKLGS
58
67
75
80
85
90
95
LVVRCSFYEDFLEYHDVRVVLDFI
98
105
110
115
120
125
SNP
DEL
INS
Sec. struc:
Helix Strand
Helices labelled H1, H2, ... and strands by their sheets A, B, ...
Motifs: β beta turn beta hairpin
Disulphides: disulphide bond disulphide bond to other chain
Residue contacts: to metal

Fig. 6. Mutational profiling and their amino acid positions in ORF8 proteins of SARS-CoV-2 (96 variants). (Upper left) Frequency distribution of various mutations in the ORF8 protein variants (96 variants) of SARS-CoV-2. A red circle marked the mutations (Single Nucleotide Polymorphism (SNP)), Deletions with green/black, and Insertions with cyan/blue. (Upper left) Structural representation of SARS-CoV-2 ORF8 monomer showing high-frequency mutations. Below each sequence, a 2D presentation of the secondary structure plot has been depicted using PDBsum. (β Strands are presented in blue with naming convention as A (long β strands) and B (short β strands)).

the end of the C-terminal region. Furthermore, it was discovered that several positions within the amino acid sequence of SARS-CoV-2 ORF8 protein might have more than one mutation [36]. For example, at position 11, which is a threonine (T) in a reference ORF8 protein, one might find isoleucine (I), alanine (A), or lysine (K) in some SARS-CoV-2 ORF8 natural variants. Based on the observed mutations in different ORF8 variants, it is noticed that threonine (T) and tryptophan (W) are the most vulnerable for mutations. It is noteworthy that the SARS-CoV-2 ORF8 is rapidly undergoing mutational changes, indicating that it is a highly evolving protein, whereas the Bat-RaTG13-CoV ORF8 (Fig. S5) and the Pangolin-CoV ORF8 are highly conserved (Fig. S6). A list of mutations and their frequency distributions is presented in Table S5, along with a histogram plot (Fig. 6 (upper left)). Structural representation of a single chain ORF8 protein depicting the high-frequency mutation has been shown in (Fig. 6 (upper right)).

The N-terminal signal peptide of ORF8 (D1) of SARS-CoV-2 is hydrophobic. We further analyzed mutations within this region and observed that hydrophobic to hydrophobic mutations were dominating, indicating that the domain's hydrophobicity is maintained (Table S6). Therefore, we can postulate that there are probably no functional changes in the hydrophobic N-terminal signal peptide associated with the evolutionary variability. Furthermore, it was found that there was a change from hydrophilic to hydrophobic residues in two positions of the D1 region, thereby further enhancing its hydrophobic nature. Although hydrophobic to hydrophilic and hydrophilic to hydrophilic mutations were also observed, they were not expected to have significant effects when compared to hydrophobicity changes, as hydrophobic mutations were observed at eight positions.

In contrast, hydrophilic mutations were only present at five positions. The ORF8 chain (D2) was demonstrated to be a region enriched in mutations affecting hydrophilic residues, with corresponding mutations being found in thirty-eight positions. Out of all mutations, only twenty-three mutations were affecting hydrophobic residues in D2.

2.6. Mutation profiling of SARS-CoV-2 ORF8 natural variants

Distinct, non-synonymous mutations and the associated frequency of mutations predicted effects (using Meta-SNP), as well as the predicted changes in the structural stability (using I-MUTANT) due to mutation(s), are presented in (supplementary table S1). The most frequent mutation in the ORF8 proteins turned out to be L84S (hydrophobic (L) to non-charged hydrophilic (S)), which is a clade (S) determining mutation with the frequency of 23 (Hosseini Rad Sm and McLellan 2020). The results suggest that the L84S mutation decreases structural stability and can change the ORF8 functions. Based on the predicted effects and changes of stability, we classified the mutations into four types Table S7: Disease-Decreasing: This class includes disease mutations that are decreasing the stability of the protein, with most of them occurring in D2; Neutral-Decreasing: Although the mutations are of a neutral type and supposedly are not harmful to the host, they cause protein structure stability to decrease; Disease-Increasing: These mutations lie within the D1; they increase the protein's stability, making the hydrophobic N-terminal more stable and thereby making the localization of ORF8 to ER more efficient; Neutral-Increasing: The frequency of mutations is very low in this class, although the mutations are neutral, they increase the protein's stability effectively, and they all occur in the D2 domain (supplementary table S2). Supplementary tables S3 and S4 list unique ORF8 protein IDs and their associated mutations with domain(s), and the predicted effects and changes of structural stability are presented.

Furthermore, based on the three different types of mutations viz. neutral, disease, and mix of neutral & disease, all ORF8 proteins are classified into three groups, which are presented in (supplementary table S5). It was concluded that most mutations examined in the distinct variants of the ORF8 proteins of SARS-CoV-2 turned up to be neutral, while 42% of the mutations become disease-causing as predicted. Furthermore, based on their abundance in several SARS-CoV-2 ORF8 variants, mutations S24L (which is present in 27 variants) and L84S (which was found in 23 variants) can be classified as strain-determining. Notably, one of them (L84S) was already defined in the literature [48]. Tables S6 and S7 represent the lists of ORF8 protein IDs and associated details on sequences with these two strain-determining mutations. Note that there are 64 ORF8 sequences, which do not possess strain-determining mutations. This high mutational variability suggests that the ORF8 protein is undoubtedly one of the essential proteins, which directs the pathogenicity of a variety of strains of SARS-CoV-2.

2.7. Remarks based on mutations over ORF8 proteins of SARS-CoV-2, bat-RaTG13-CoV, and Pangolin-CoV

Next, we compared the SARS-CoV-2 ORF8 with the Bat-RaTG13-CoV and the Pangolin-CoV ORF8 to study mutation evolution. The mutations in the ORF8 protein regarding the reference ORF8 sequence of Bat-RaTG13-CoV were found to be of the neutral type as predicted through the webserver Meta-SNP. All of them are expected to cause a decrease in ORF8 stability as determined using the server I-MUTANT (Fig. S5). The detailed analysis of all mutations is presented in Table S8. Based on these data, it can be suggested that several mutations in SARS-CoV-2 ORF8 can be considered as a reversal mutation. These are mutations at specific positions in the SARS-CoV-2 ORF8, which are substituted by the residues present in the Bat-RaTG13-CoV and the Pangolin-CoV ORF8 proteins. The results indicate that some positions in the SARS-CoV-2 ORF8 represent a kind of reverse genetic engineering compared with the Bat-RaTG13-CoV and the Pangolin-CoV ORF8 (Table S9).

2.8. Possible flow of mutations in ORF8 evolution

Here we present five different possible mutation flows according to data collection of the virus samples from patients [49,50]. Sequence homology and amino acid composition-based phylogenies have been drawn for the associated ORF8 proteins in each flow. Note that the ORF8 sequence QLJ93922.1 (USA) possesses consecutive (38–46 aa) deletion mutations. The other sequence, QKI36860.1 (Guangzhou, China) accommodates four insertion mutations at the end of the C-terminal region (122–125) along with two other mutations, S84L and D119E.

2.9. Flow-I

In this flow of mutations, we have described the occurrence of mutations in the US sequences based on chronological order, considering the Wuhan ORF8 sequence YP 009724396 as the reference sequence (Fig. 7). The protein sequence QMI92505.1 possesses a mutation L4F of neutral type with no change in hydropathy. However, it showed a decreasing effect on the stability of the protein. Following this sequence, another sequence, QMT48896.1, was identified following the time scale, in which a second mutation located at D63 N emerged. This mutation is of neutral type, and no change in hydropathy was observed. Therefore, this sequence accumulated two neutral mutations, which may affect the protein's function as both mutations cause a decrease in protein stability. The QMT96239.1 sequence harbors another mutation, G8R, which is of the disease-increasing type, and the hydropathy changed from hydrophobic to hydrophilic. Another mutation, D35Y, occurred as a second-order mutation in the QMU92030.1 sequence in addition to the G8R mutation. As D35Y is neutral and G8R is of the disease-increasing type, their combination may alter both the protein's structure and function. To support these mutation flows, we analyzed the protein sequence similarity based on phylogeny and amino acid composition. The reference ORF8 sequence YP 009724396 was found to be much more like the variants QMT48896.1 and QMI92505.1, which are more like each other as depicted in the sequence-based phylogeny (Fig. 7A). This sequence-based similarity of the QMT48896.1 and QMI92505.1 ORF8 proteins is illustrated in the chronology of mutations. Similarly,

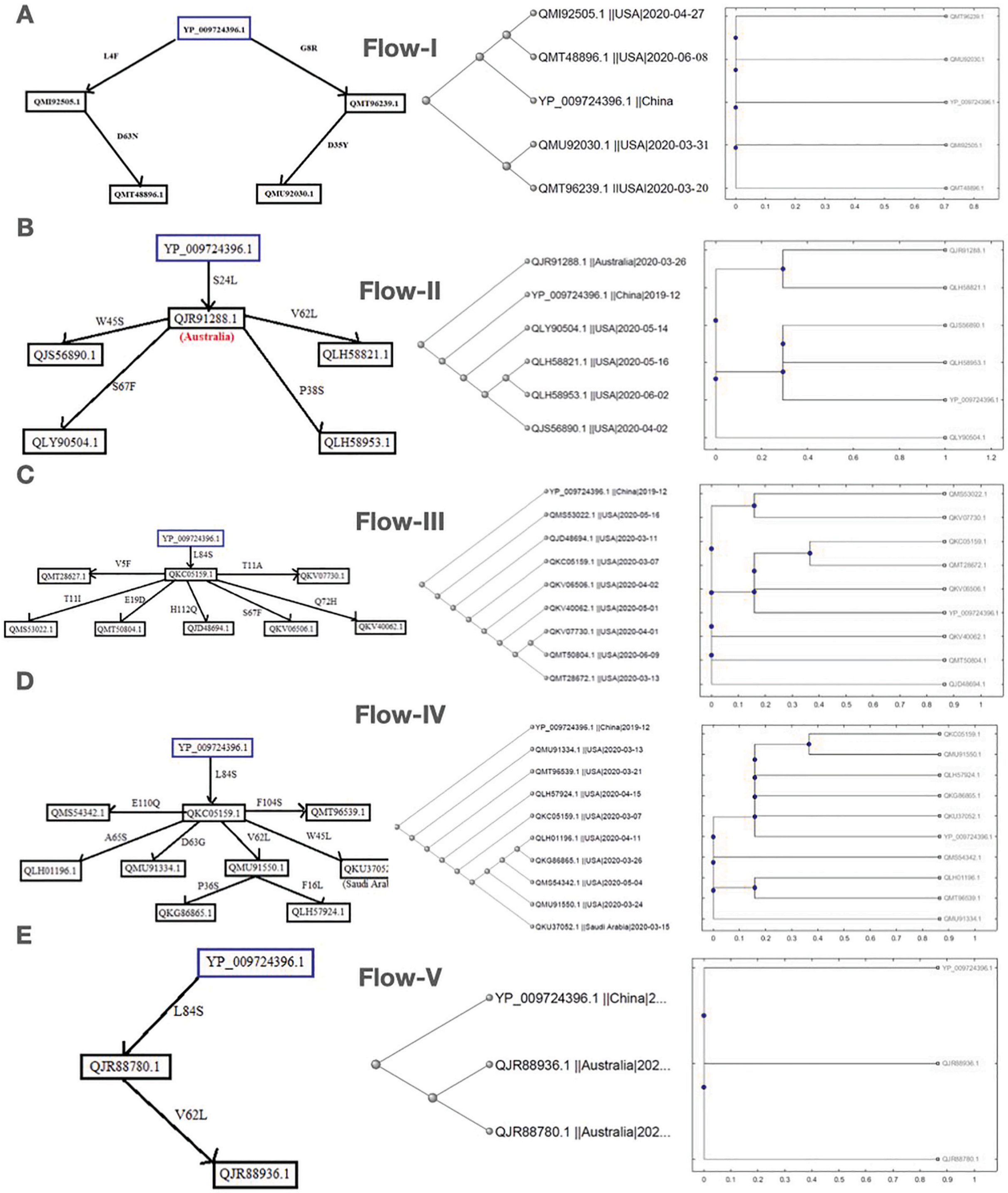

Fig. 7. Possible flow of mutations in ORF8 evolution. (A–E) The possible flow of mutations in the ORF8 (SARS-CoV-2) sequences isolated in the US, Australia, US, US and Saudi Arabia, Australia. (Right panel) Phylogenetic relationship based on amino acid sequence similarity and amino acid composition (right) of ORF8 proteins of SARS-CoV-2 in US, Australia, US, US and Saudi Arabia, Australia.

the mutation flow of the sequences QMT96239.1 and QMU92030.1 is supported by the respective sequence-based similarity. The network of five ORF8 protein variants from the US is justified based on the similar amino acid compositions/conservations across the five sequences.

2.10. Flow-II

We observed one sequence with first-order mutations in this flow of mutations, where only one mutation accumulated in the sequence (Fig. 7B). Additionally, four sequences (all are from the US) were identified with second-order mutations, suggesting that four sequences were found to have two mutations. The protein sequence QLY90504.1 possesses a second mutation at position 67, which changed the hydrophilic serine (S) to the hydrophobic phenylalanine (F). Therefore, it may account for disrupting the ionic interactions as it is a neutral mutation, the corresponding sequence accumulated two neutral mutations. The protein sequence QLH58953.1 acquired a second mutation, P38S, which was found to be of the disease-increasing type, and the hydropathy also changed from hydrophobic to hydrophilic, indicating that these mutations may be of some importance. The protein sequence QLH58821.1 possesses a second mutation, V62L, which was found to be of the disease-neutral type with no hydropathy change. Here, this sequence accumulated two neutral mutations, which may account for some functional changes. By comparing both the sequence-based phylogeny and amino acid conservation-based phylogeny, we found that according to sequence-based phylogeny, the Australian sequence is closely related to the ORF8 Wuhan sequence. However, according to the pathway, it should be closely related to both the Wuhan sequence and second-order mutations. This can be attributed to the presence of 119 amino acid residues instead of 121 amino acid residues. In this case, the sequence has two amino acid deletions. Therefore, it is present at the first node.

2.11. Flow-III

We analyzed the US sequences considering the Wuhan sequence (YP 009724396.1) as the reference and found one sequence, QKC05159.1, with a single mutation and seven sequences with two mutations each (Fig. 7C). The first sequence, QKC05159.1, contained the L84S mutation (strain-determining mutation), neutral. However, the hydropathy changed from hydrophobic to hydrophilic, which may account for some significant change of a function. The sequences that accumulated a second mutation along with L84S are as follows:

QMT28672.1: This sequence possesses a second mutation V5F, which was predicted to be of neutral type with no hydropathy change. Hence this sequence acquired two neutral mutations, and together these mutations may alter the protein's function.

QMS53022.1: This protein sequence acquired a second mutation at position 11, which changed the hydrophilic threonine (T) to the hydrophobic isoleucine (I), affecting the ionic interactions. This mutation was found to be a disease-increasing type, so it may affect the protein structure.

QMT50804.1: This sequence gained a second mutation, E19D, which was predicted to be of the disease-increasing type with no hydropathy change. The sequence first accumulated a neutral mutation then a disease-increasing mutation, signifying that these mutations may have some functional importance.

QJD48694.1: H112Q occurred as a second mutation in this sequence, which was found to be of the disease-increasing type with no hydropathy change. Consequently, these mutations may contribute to the immune evasion property of the virus.

QKV06506.1: This sequence possesses the S67F mutation, which was predicted to be of neutral type, and changed the hydrophilic serine (S) to the hydrophobic phenylalanine (F), thus interfering with the ionic interactions that potentially increase or decrease the affinity of the viral protein for a particular host cell protein.

QKV40062.1: This sequence acquired a second mutation at Q72H, which was found to be a neutral mutation, and no change in hydropathy was observed. As this sequence accumulated two neutral mutations, it can be assumed that neutral mutations also have significant importance.

QKV07730.1: The T11A mutation occurred as the second mutation in this sequence, which was predicted to be of the disease-increasing type, and the hydropathy was changed from hydrophilic to hydrophobic. Hence the structure and function of the protein are expected to differ.

From the sequence-based phylogeny, it was observed that the Wuhan sequence was the first to originate. Although QKC05159.1 is the first sequence in our flow considering the time, it was found that in the phylogenetic tree, it is present at the fourth node instead of the second node, which is probably due to the presence of ambiguous mutations in this sequence. It was also determined that QKV07730.1 is very similar to QMT50804.1, and QMT28672.1 was observed to be similar to QKV07730.1 and QMT50804.1. All the other sequences have second-order mutations and are closely related to each other and follow the chronology. From the amino acid-based analysis, the Wuhan sequence has a high sequence similarity to QKV06506.1, thus proving that this sequence was identified chronologically after the Wuhan sequence QKC05159.1 and QMT28672.1, which again are very similar to each other.

2.12. Flow-IV

In another possible flow of mutations (Fig. 7D), we have found one sequence with a single mutation, six sequences with two mutations, and another two sequences with three mutations. The US sequence QKC05159.1 was identified to have the L84S mutation, which is a strain determining mutation, and was predicted to be a neutral mutation where hydropathy was changed from hydrophobic to hydrophilic. The sequences that accumulated second mutations along with L84S are the following, and it should be noted that the mutational accumulation occurred in a single strain:

QMS54342.1: This US sequence acquired the E110Q mutation, which was predicted to be of the disease-increasing type, where no change in hydropathy was observed, and consequently, it may contribute to virulence properties of the virus.

QLH01196.1: The A65S mutation occurred as a second mutation in this US sequence, which was found to be of neutral type. However, the hydropathy changed from hydrophobic to hydrophilic, thus potentially influencing the function of the protein.

QMU91334.1: This US sequence possesses the D63G mutation, which was predicted to be of neutral type. However, the hydropathy changed from hydrophilic to hydrophobic so, this sequence accumulated two neutral mutations, which may allow the virus to evolve in terms of virulence.

QMU91550.1: This US sequence mutated at position 62, which changed valine (V) to leucine (L), thereby mostly keeping the protein's hydrophobicity unchanged. It was a neutral mutation even though it may influence the virulence properties of the protein. This sequence was followed by another sequence, QKG86865.1, with a third mutation at position 36, which changed the hydrophobic proline (P) to the hydrophilic serine (S). Thus, the mutation was neutral, thus accumulating two neutral and one disease-increasing mutation, being significant for the evolution of the virus. We identified one more sequence, QLH57924.1, which possesses a third mutation, F16L, which was predicted to be neutral, and no hydropathy change was observed. This sequence acquired three neutral mutations that may promote virus survival.

QKU37052.1: This sequence with the W45L mutation was reported in Saudi Arabia, which was found to be of the disease-increasing type with no hydropathy change. Therefore, this sequence also accumulated one neutral and one disease-increasing mutation, affecting both the protein's structure and function.

QMT96539.1: The F104S mutation was reported in the US sequence, which was found to be of a disease-increasing type, and the hydropathy changed from hydrophobic to hydrophilic. Altogether, the sequence possesses one neutral and one disease-increasing mutation that may allow the virus to acquire new properties for better survival strategies. Sequence-based phylogeny suggested that the Wuhan sequence originated first.

Due to ambiguous amino acids, the sequence QKC05159.1 was not observed close to be close to the Wuhan sequence. QKG86865.1 and QLH57924.1 were found to have third-order mutations, and they are assumed to be closely related by the flow, and the same has been supported by amino acid conservation-based phylogeny.

2.13. Flow-V

QJR88780.1 (Australia) possesses the mutation L84S compared to the Wuhan ORF8 sequence YP 009724396.1 (Fig. 7E). Another sequence, QJR88936.1, was reported, which possesses a second mutation, V62L. This mutation was predicted to be neutral with no change in hydropathy. However, the hydrophobicity increased. This sequence belongs to a particular strain and acquires two neutral mutations, indicating that these mutations may play some vital role in the function of ORF8a. As shown in both the sequence-based phylogeny and amino acid conservation-based phylogeny, the Wuhan sequence originated earlier, and the sequences QJR88780.1 and QJR88936.1 are more closely related to each other than to the Wuhan sequence as both sequences have one common mutation not present in the Wuhan sequence.

3. Discussion

Among SARS-CoV-2 proteins, the ORF8 accessory protein is unique because it plays a vital role in bypassing the host immune surveillance mechanism. This protein is found to have a wide variety of mutations, and among them, L84S (23) and S24L (7) have the highest frequency of occurrence, which bears distinct functional significance. It has been reported that L84S and S24L show antagonistic effects on protein folding stability of the SARS-CoV-2 ORF8 [48]. L84S destabilizes protein, thereby up-regulating host-immune activity and S24L positively favors folding stability, thus enhancing the functionality of the ORF8 protein. L84S is already established as a strain-determining mutation, and since, according to our studies, both L84S and S24L do not occur together in a single sequence of the SARS-CoV-2 ORF8 protein, it is proposed that virus with the S24L mutation is a new strain altogether. We also observed that hydrophobic to hydrophobic mutations are dominant in the D1 domain. Therefore, hydrophobicity is an essential property for the N-terminal signal peptide. However, in the D2 domain, hydrophobic to hydrophilic mutations are observed more frequently, consequently making the ionic interactions more favorable and allowing the protein to evolve, providing better pathogenicity efficacy.

The ORF8 sequence of SARS-CoV-2 shows 93% similarity with the Bat-RaTG13-CoV and 88% similarity with that of the Pangolin-CoV ORF8. Thus, the ORF8 protein of SARS-CoV-2 can be considered a valuable candidate for deterministic evolutionary studies and the determination of the origin of SARS-CoV-2. We also analyzed a wide variety of mutations in the SARS-CoV-2 ORF8, where we compared them with the ORF8 of Bat-RaTG13-CoV and the Pangolin-CoV in relation to charge and hydrophobicity. We found that the Bat-RaTG13-CoV ORF8 protein exhibits precisely the same properties as that of the SARS-CoV-2 ORF8 protein, whereas the properties of the Pangolin-CoV ORF8 are relatively less similar to the SARS-CoV-2 ORF8. Furthermore, to study the evolutionary nature of mutations in the ORF8, we aligned three bat sequences and found that two of them were the same, and there were only six amino acid differences in the third compared to the other two sequences. So, only two variants were identified for the Bat-RaTG13-CoV ORF8. Therefore, it shows that the mutation rate is slow in the Bat-RaTG13-CoV ORF8.

However, for pangolins, no differences were observed among four Pangolin-CoV ORF8 sequences, and therefore, only a single variant of ORF8 was identified. The Bat-RaTG13-CoV, the Pangolin-CoV, and the SARS-CoV-2 ORF8 displayed a high similarity index based on sequence alignment, biochemical characteristics, and secondary structure analysis [51]. Additionally, in the ORF8 of SARS-CoV-2, specific mutations were found to exhibit exact reversal regarding bats and pangolins and, therefore, point towards the genomic origin of SARS-CoV-2. However, unlike Bat-RaTG13-CoV and Pangolin-CoV, the mutational distribution of SARS-CoV-2 ORF8 is widespread, ranging from position 3 to 121, having no defined conserved region. This is a rather surprising observation. Furthermore, this property differentiates the SARS-CoV-2 ORF8 from Bat-RaTG13-CoV and Pangolin-CoV, raising the question of the natural path of evolution of mutations in SARS-CoV-2.

We further predicted the types and effects of mutations of 95 sequences and grouped them into four clusters, and found that the disease-decreasing type mutations with decreasing effect on stability are more prominent. Consequently, it is hypothesized that these mutations are promoting viral survival. Furthermore, we tracked the possible flow of mutations following time and geographic locations and validated our proposal concerning sequence-based and amino acid conservation-based phylogeny and therefore putting forward the order of accumulation of mutations. We are aware of the need to confirm our prediction by structural biology data and experimental observations from in vitro and in vivo studies. However, the powerful tools in bioinformatics have allowed us to generate a reasonable basis for the potential effect of SARS-CoV-2 ORF8 mutations.

4. Conclusions

This study represents the results of a comprehensive analysis of the uniqueness of the pandemic-causing SARS-CoV-2 by focusing on one of its accessory proteins, the ORF8 protein. ORF8 is involved in modulating the adaptive host immunity and innate immune responses by surpassing interferon-mediated antiviral host responses. Our study relies heavily on bioinformatics to analyze the 87 unique mutations identified in the SARS-CoV-2 ORF8 protein, the phylogenetic comparison to bat and pangolin CoV ORF8 proteins and evaluation of the potential effect of the identified mutations on SARS-CoV-2 virulence. We acknowledge that our findings lack support from direct structural analyses and cell-based confirmations but believe that the data presented on ORF8 can provide a suitable basis for further explorations to improve our understanding of the essential functions of providing all potential means to tackle the current pandemic. In future endeavors, more critical studies on the ORF8 protein of SARS-CoV-2 are necessary for a better understanding of the importance of high-frequency mutations and their role related to the host immune system and to validate the origin of SARS-CoV-2 more precisely.

5. Materials and methods

5.1. Dataset of the ORF8 of SARS-CoV-2

As of February 14, 2021, - 29,881 complete genomes of SARS-CoV-2 were available on the NCBI (National Center for Biotechnology Information) database. Each genome contains one gene for the accessory protein ORF8, and among them, only 127 sequences were found to be unique. The amino acid sequences of the ORF8 variants were exported in FASTA format using the file operations through MATLAB (version 9.3.0.713579 (R2020a)). Among these 127 unique ORF8 sequences, only 96 ORF8 protein sequences contain various mutations, and the remaining sequences either do not possess any mutations or only ambiguous mutations. The present study focused on 96 ORF8 proteins. An ORF8 protein sequence (YP 009724396.1) of the SARS-CoV-2 genome (NC 045512) from Wuhan, China, was used as the reference to identify mutations [52].

The ORF8 protein sequences of SARS-CoV, Pangolin-CoV, and Bat-RaTG13-CoV were retrieved from the NCBI as reference sequences for understanding the proximal evolutionary origin of the SARS-CoV-2 ORF8 protein. The unique ORF8 variant was obtained among the four available Pangolin-CoV sequences (QIA48620.1, QIA48638.1, QIA48647.1, and QIQ54055.1), were retrieved along with three available ORF8 sequences of Bat-RaTG13-CoV (AVP78048.1, AVP78037.1, and QHR63307.1), of which two variants (QHR63307.1 and AVP78048.1) showed 96% sequence similarity.

5.2. Structural modeling and visualization

The structures of the SARS-CoV-2 ORF8 and the SARS-CoV ORF7a have been retrieved from Protein Data Bank bearing with the accession number ID:7JTL (www.rcsb.org) and the accession number 6W37 (www.rcsb.org), respectively. All other ORF8 structures i.e., ORF8a and ORF8b of SARS-CoV, ORF8 of Bat-RaTG13-CoV, and Pangolin-CoV were modeled using the Swiss Model server [53] since ITASSER was not available due to server maintenance. The structures were analyzed and visualized using the UCSF ChimeraX tool [54]. The plotting was done by using ORIGIN8 software. PDBsum database has been used to retrieve the SARS-CoV-2 ORF8 protein structure information, including the 2D secondary structure plot and interface regions [55].

5.3. Mutation identification

Mutations are responsible for several genetic orders/disorders. Identifying these mutations requires novel detection methods, which have been reported in the literature [56]. In this study, each unique ORF8 sequence was aligned using NCBI protein p-BLAST and omega-blast suites to determine mismatches, and thereby missense mutations (amino acid changes) were identified [57,58]. For the effect of identified mutations, a web server Meta-SNP was used, and for the structural effects of mutations, another web server I-MUTANT was used [59,60]. The web server QUARK was used to predict the secondary structure of ORF8 proteins [61,62]. The mutation profiles have been presented using the Web-Logo3 server [63].

5.4. Amino acids compositions and phylogeny

The frequency of occurrence of each amino acid A_i was determined for each primary sequence of ORF8 proteins. For all 96 ORF8 proteins, a twenty-dimensional frequency vector of amino acids was obtained. A distance matrix (Euclidean distance) was formed by measuring the distance (pairwise) between the twenty-dimensional frequency vectors for each ORF8 protein [64,65]. Thereby, applying the nearest neighbor-joining method, a phylogeny was derived from the distance matrix formed for each ORF8 protein of interest [66–68]. The following method was used to compute the distances of the new nodes to all other nodes at every iteration; the equation to calculate the distances between the new node, n, after joining i and j and all nodes (k), was the following:

$$D(n, k) = a * D(i, k) + (1 - a) * D(j, k) - a * D(n, i) - (1 - a) * D(n, j)$$

This equation could help us find the correct tree with additive data (minimum variance reduction). Note that, typically, equal variance and independence of evolutionary distance estimates ($a = 1/2$) are assumed [69]. The fundamental physicochemical properties were obtained by VOLPES, allowing unprecedented insights into the amino acid sequences of ORF8 variants among all species and in-depth exploration.

5.5. The propensity of intrinsic disorder

Per-residue disorder distribution in sequences of query proteins was evaluated by PONDR® VSL2 [70], one of the most accurate standalone disorder predictors [71–74]. The per-residue disorder predisposition scores are on a scale from 0 to 1, where 0 and 1 indicate fully ordered and disordered residues, respectively. Values above the threshold of 0.5 are considered to correspond to disordered residues, whereas residues with disorder scores between 0.25 and 0.5 are considered highly flexible, and residues with disorder scores between 0.1 and 0.25 are regarded as moderately flexible.

Author contributions

SSH conceived the project. SG, DA, SSH, and VNU examined the mutations and performed Analysis. SSH, PPC, SG, DA, MMT and VNU analyzed the results. SH, MMT and AAAA, wrote the initial draft. SSH, SG, DA, PPC, MS, DP, PA, TMAEA, ASA, RK, KL, MT, AL, GKA, VNU, SPS, WBC, BDU, NR, and AMB reviewed and edited the manuscript. PKP made all images. All the authors checked, reviewed, and approved the final version of the manuscript.

Data and materials availability

The following are available online as supplementary information/material, 96 unique ORF8 protein IDs with associated information.

Mutations across ORF8 proteins of SARS-CoV-2, and their predicted effects, ORF8 protein IDs and corresponding mutations, predicted effects, list of the three types of mutations possessed by 93 distinct SARS-CoV-2 ORF8 proteins, and Observations on mutations in ORF8 concerning ORF8 of Pangolin-CoV, Bat-RaTG13-CoV. All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Additional data related to this paper may be requested from the authors.

Declaration of competing interest

All authors declare no conflict of interest.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2021.104380.

References

[1] M. Hussain, S. Shabbir, A. Amnaullah, F. Raza, M.J. Imdad, S.J.J.o.M.V. Zahid, Immunoinformatic Analysis of Structural and Epitope Variations in Spike and Orf8 Proteins of SARS-CoV-2/B. 1.1, vol. 7, 2021.

[2] H.A. Rothan, S.N. Byrareddy, The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak, J. Autoimmun. 109 (2020) 102433.

[3] M.U.G. Kraemer, C.H. Yang, B. Gutierrez, C.H. Wu, B. Klein, D.M. Pigott, C.-D.W. G. Open, L. du Plessis, N.R. Faria, R. Li, W.P. Hanage, J.S. Brownstein, M. Layan, A. Vespignani, H. Tian, C. Dye, O.G. Pybus, S.V. Scarpino, The effect of human mobility and control measures on the COVID-19 epidemic in China, Science 368 (2020) 493–497.

[4] A. Zumla, M.S. Niederman, Editorial, The explosive epidemic outbreak of novel coronavirus disease 2019 (COVID-19) and the persistent threat of respiratory tract infectious diseases to global health security, Curr. Opin. Pulm. Med. 26 (2020) 193–196.

[5] H. Amawi, G.a.I. Abu Deiab, A.A.A. Aljabali, K. Dua, M.M.J.T.d. Tambuwala, COVID-19 pandemic: an overview of epidemiology, pathogenesis, diagnostics and potential vaccines and therapeutics 11 (2020) 245–268.

[6] M. Ceccarelli, M. Berretta, E. Venanzi Rullo, G. Nunnari, B. Cacopardo, Differences and similarities between Severe Acute Respiratory Syndrome (SARS)-CoronaVirus (CoV) and SARS-CoV-2. Would a rose by another name smell as sweet? Eur. Rev. Med. Pharmacol. Sci. 24 (2020) 2781–2783.

[7] J. Xu, S. Zhao, T. Teng, A.E. Abdalla, W. Zhu, L. Xie, Y. Wang, X. Guo, Systematic comparison of two animal-to-human transmitted human coronaviruses: SARS-CoV-2 and SARS-CoV, Viruses (2020) 12.

[8] A.A. Aljabali, H.A. Bakshi, S. Satija, M. Metha, P. Prasher, R.M. Ennab, D. K. Chellappan, G. Gupta, P. Negi, R.J.P.n. Goyal, COVID-19: underpinning research for detection, Therapeutics, and Vaccines Development 8 (2020) 323–353.

[9] R. Giri, T. Bhardwaj, M. Shegane, B.R. Gehi, P. Kumar, K. Gadhave, C.J. Oldfield, V. N. Uversky, Understanding COVID-19 via comparative analysis of dark proteomes of SARS-CoV-2, human SARS and bat SARS-like coronaviruses, Cell. Mol. Life Sci. (2020).

[10] Y.Z. Zhang, E.C. Holmes, A genomic perspective on the origin and emergence of SARS-CoV-2, Cell 181 (2020) 223–227.

[11] Z. Shen, Y. Xiao, L. Kang, W. Ma, L. Shi, L. Zhang, Z. Zhou, J. Yang, J. Zhong, D. Yang, L. Guo, G. Zhang, H. Li, Y. Xu, M. Chen, Z. Gao, J. Wang, L. Ren, M. Li, Genomic diversity of severe acute respiratory syndrome-coronavirus 2 in patients with coronavirus disease 2019, Clin. Infect. Dis. 71 (2020) 713–720.

[12] J. Shang, N. Han, Z. Chen, Y. Peng, L. Li, H. Zhou, C. Ji, J. Meng, T. Jiang, A. Wu, Compositional Diversity and Evolutionary Pattern of Coronavirus Accessory Proteins, Brief Bioinform, 2020.

[13] J. Wu, W. Deng, S. Li, X. Yang, Advances in Research on ACE2 as a Receptor for 2019-nCoV, Cell Mol Life Sci, 2020.

[14] J.Y. Li, C.H. Liao, Q. Wang, Y.J. Tan, R. Luo, Y. Qiu, X.Y. Ge, The ORF6, ORF8 and nucleocapsid proteins of SARS-CoV-2 inhibit type I interferon signaling pathway, Virus Res. 286 (2020) 198074.

[15] D. Kim, J.Y. Lee, J.S. Yang, J.W. Kim, V.N. Kim, H. Chang, The architecture of SARS-CoV-2 transcriptome, Cell 181 (2020) 914–921 e910.

[16] D.X. Liu, Q. Yuan, Y. Liao, Coronavirus envelope protein: a small membrane protein with multiple functions, Cell. Mol. Life Sci. 64 (2007) 2043–2048.

[17] Y. Zhang, J. Zhang, Y. Chen, B. Luo, Y. Yuan, F. Huang, T. Yang, F. Yu, J. Liu, B. Liu, Z. Song, J. Chen, T. Pan, X. Zhang, Y. Li, R. Li, W. Huang, F. Xiao, H. Zhang, The ORF8 Protein of SARS-CoV-2 Mediates Immune Evasion through Potently Downregulating MHC-I, bioRxiv, 2020, 2020.2005.2024.111823.

[18] S. Mohammad, A. Bouchama, B. Mohammad Alharbi, M. Rashid, T. Saleem Khatlani, N.S. Gaber, S.S. Malik, SARS-CoV-2 ORF8 and SARS-CoV ORF8ab: genomic divergence and functional convergence, Pathogens 9 (2020).

[19] Y. Tan, T. Schneider, M. Leong, L. Aravind, D. Zhang, Novel immunoglobulin domain proteins provide insights into evolution and pathogenesis of SARS-CoV-2-related viruses, mBio (2020) 11.

[20] S. Chen, X. Zheng, J. Zhu, R. Ding, Y. Jin, W. Zhang, H. Yang, Y. Zheng, X. Li, G. Duan, Extended ORF8 gene region is valuable in the epidemiological investigation of severe acute respiratory syndrome-similar coronavirus, J. Infect. Dis. 222 (2020) 223–233.

[21] R.A. Khailany, M. Safdar, M. Ozaslan, Genomic characterization of a novel SARS-CoV-2, Gene Rep 19 (2020) 100682.

[22] I. Alam, A. Kamau, M. Kulmanov, S.T. Arold, A. Pain, T. Gojobori, C.M. Duarte, Functional Pangenome Analysis Provides Insights into the Origin, Function and Pathways to Therapy of SARS-CoV-2 Coronavirus, bioRxiv, 2020, p. 2020, 2002.2017.952895.

[23] I. Alam, A.A. Kamau, M. Kulmanov, L. Jaremko, S.T. Arold, A. Pain, T. Gojobori, C. M. Duarte, Functional pangenome analysis shows key features of E protein are preserved in SARS and SARS-CoV-2, Front Cell Infect Microbiol 10 (2020) 405.

[24] J. Wu, X. Yuan, B. Wang, R. Gu, W. Li, X. Xiang, L. Tang, H. Sun, Severe acute respiratory syndrome coronavirus 2: from gene structure to pathogenic mechanisms and potential therapy, Front. Microbiol. 11 (2020) 1576.

[25] S.R. Schaecher, A. Pekosz, SARS coronavirus accessory gene expression and function, in: S.K. Lal (Ed.), Molecular Biology of the SARS-Coronavirus, Springer Berlin Heidelberg, Berlin, Heidelberg, 2010, pp. 153–166.

[26] B.E. Young, S.W. Fong, Y.H. Chan, T.M. Mak, L.W. Ang, D.E. Anderson, C.Y. Lee, S. N. Amrun, B. Lee, Y.S. Goh, Y.C.F. Su, W.E. Wei, S. Kalimuddin, L.Y.A. Chai, S. Pada, S.Y. Tan, L. Sun, P. Parthasarathy, Y.Y.C. Chen, T. Barkham, R.T.P. Lin, S. Maurer-Stroh, Y.S. Leo, L.F. Wang, L. Renia, V.J. Lee, G.J.D. Smith, D.C. Lye, L.F. P. Ng, Effects of a major deletion in the SARS-CoV-2 genome on the severity of infection and the inflammatory response: an observational cohort study, Lancet 396 (2020) 603–611.

[27] A. Parlikar, K. Kalia, S. Sinha, S. Patnaik, N. Sharma, S.G. Vemuri, G. Sharma, Understanding genomic diversity, pan-genome, and evolution of SARS-CoV-2, PeerJ 8 (2020), e9576.

[28] M.D. Park, Immune evasion via SARS-CoV-2 ORF8 protein? Nat. Rev. Immunol. 20 (2020) 408.

[29] S. Kumar, R. Nyodu, V.K. Maurya, S.K. Saxena, Host immune response and immunobiology of human SARS-CoV-2 infection, in: S.K. Saxena (Ed.), Coronavirus Disease 2019 (COVID-19), Springer, 2020, pp. 43–53.

[30] S.C. Sung, C.Y. Chao, K.S. Jeng, J.Y. Yang, M.M. Lai, The 8ab protein of SARS-CoV is a luminal ER membrane-associated protein and induces the activation of ATF6, Virology 387 (2009) 402–413.

[31] M. Frieman, R. Baric, Mechanisms of severe acute respiratory syndrome pathogenesis and innate immunomodulation, Microbiol. Mol. Biol. Rev. 72 (2008) 672–685. Table of Contents.

[32] E. Krissinel, K.J.A.C.S.D.B.C. Henrick, Secondary-structure Matching (SSM), a New Tool for Fast Protein Structure Alignment in Three Dimensions, vol. 60, 2004, pp. 2256–2268.

[33] C.-S. Shi, N.R. Nabar, N.-N. Huang, J.H.J.C.d.d. Kehrl, SARS-Coronavirus Open Reading Frame-8b triggers intracellular stress pathways and activates, NLRP3 inflammasomes 5 (2019) 1–12.

[34] S. Mohammad, A. Bouchama, B. Mohammad Alharbi, M. Rashid, T. Saleem Khatlani, N.S. Gaber, S.S.J.P. Malik, SARS-CoV-2 ORF8 and SARS-CoV ORF8ab: Genomic Divergence and Functional Convergence, vol. 9, 2020, p. 677.

[35] F. Meng, V. Uversky, L. Kurgan, Computational prediction of intrinsic disorder in proteins, Curr. Protocols Protein Sci. 88 (2017), 2.16. 11-16.12.16. 14.

[36] A. Alkhansa, G. Lakkis, L.J.G.r. El Zein, Mutational analysis of SARS-CoV-2 ORF8 during six months of COVID-19, Pandemic 23 (2021) 101024.

[37] K. Peng, S. Vucetic, P. Radivojac, C.J. Brown, A.K. Dunker, Z.J.J.o.b. Obradovic, c. biology, Optimizing Long Intrinsic Disorder Predictors with Protein Evolutionary Information, vol. 3, 2005, pp. 35–60.

[38] X. Fan, L.J.J.o.B.S. Kurgan, Dynamics, Accurate Prediction of Disorder in Protein Chains with a Comprehensive and Empirically Designed Consensus, vol. 32, 2014, pp. 448–464.

[39] F. Meng, V.N. Uversky, L.J.C. Kurgan, M.L. Sciences, Comprehensive review of methods for prediction of intrinsic disorder and its molecular functions 74 (2017) 3069–3090.

[40] Z.-L. Peng, L.J.C.P. Kurgan, P. Science, Comprehensive comparative assessment of in-silico predictors of disordered regions 13 (2012) 6–18.

[41] T.G. Flower, C.Z. Buffalo, R.M. Hooy, M. Allaire, X. Ren, J.H. Hurley, Structure of SARS-CoV-2 ORF8, a Rapidly Evolving Coronavirus Protein Implicated in Immune Evasion, bioRxiv, 2020.

[42] T.G. Flower, C.Z. Buffalo, R.M. Hooy, M. Allaire, X. Ren, J.H.J.B. Hurley, Structure of SARS-CoV-2 ORF8, a Rapidly Evolving Coronavirus Protein Implicated in Immune Evasion, 2020.

[43] X. Wang, J.-Y. Lam, W.-M. Wong, C.-K. Yuen, J.-P. Cai, S.W.-N. Au, J.F.-W. Chan, K.K. To, K.-H. Kok, K.-Y.J.M. Yuen, Accurate Diagnosis of COVID-19 by a Novel Immunogenic Secreted SARS-CoV-2 Orf8 Protein, 2020, p. 11.

[44] S.H. Shahcheraghi, J. Ayatollahi, A.A. Aljabali, M.D. Shastri, S.D. Shukla, D. K. Chellappan, N.K. Jha, K. Anand, N.K. Katari, M.J.T.D. Mehta, An overview of vaccine development for COVID-19 12 (2021) 235–244.

[45] J. Kyte, R.F.J.J.o.m.b. Doolittle, A simple method for displaying the hydropathic character of a protein 157 (1982) 105–132.

[46] M.O.J.A.o.p.s. Dayhoff, structure, A model of evolutionary change in proteins 5 (1972) 89–99.

[47] A. Parlikar, K. Kalia, S. Sinha, S. Patnaik, N. Sharma, S.G. Vemuri, G.J.P. Sharma, Understanding genomic diversity, pan-genome, and evolution of SARS-CoV-2 (8) (2020), e9576.

[48] A. Hosseini Rad Sm, A.D. McLellan, Implications of SARS-CoV-2 mutations for genomic RNA structure and host microRNA targeting, Int. J. Mol. Sci. 21 (2020).

[49] S.S. Hassan, D. Attrish, S. Ghosh, P.P. Choudhury, B.J.b. Roy, Pathogenetic Perspective of Missense Mutations of Orf3a Protein of Sars-Cov2, 2020.

[50] M. Seyran, S. Hassan, V.N. Uversky, P. Pal Choudhury, B.D. Uhal, K. Lundstrom, D. Attrish, N. Rezaei, A.A. Aljabali, S. Ghosh, Urgent Need for Field Surveys of Coronaviruses in Southeast Asia to Understand the SARS-CoV-2 Phylogeny and Risk Assessment for Future Outbreaks, Multidisciplinary Digital Publishing Institute, 2021.

[51] M. Seyran, K. Takayama, V.N. Uversky, K. Lundstrom, G. Palù, S.P. Sherchan, D. Attrish, N. Rezaei, A.A. Aljabali, S.J.T.F.j. Ghosh, The Structural Basis of Accelerated Host Cell Entry by SARS-CoV-2, 2020.

[52] X. Wang, Q. Zhou, Y. He, L. Liu, X. Ma, X. Wei, N. Jiang, L. Liang, Y. Zheng, L. Ma, Y. Xu, D. Yang, J. Zhang, B. Yang, N. Jiang, T. Deng, B. Zhai, Y. Gao, W. Liu, X. Bai, T. Pan, G. Wang, Y. Chang, Z. Zhang, H. Shi, W.L. Ma, Z. Gao, Nosocomial outbreak of COVID-19 pneumonia in Wuhan, China, Eur. Respir. J. 55 (2020).

[53] A. Waterhouse, M. Bertoni, S. Bienert, G. Studer, G. Tauriello, R. Gumienny, F. T. Heer, T.A.P. de Beer, C. Rempfer, L.J.N.a.r. Bordoli, SWISS-MODEL: homology modelling of protein structures and complexes 46 (2018) W296–W303.

[54] E.F. Pettersen, T.D. Goddard, C.C. Huang, E.C. Meng, G.S. Couch, T.I. Croll, J. H. Morris, T.E.J.P.S. Ferrin, U.C.S.F. ChimeraX, Structure visualization for researchers, educators, and developers 30 (2021) 70–82.

[55] R.A. Laskowski, J. Jabłońska, L. Pravda, R.S. Vařeková, J.M.J.P.s. Thornton, PDBsum: Structural summaries of PDB entries 27 (2018) 129–134.

[56] R.G. Cotton, Current methods of mutation detection, Mutat. Res. 285 (1993) 125–144.

[57] G.M. Boratyn, J. Thierry-Mieg, D. Thierry-Mieg, B. Busby, T.L. Madden, Magic-BLAST, an accurate RNA-seq aligner for long and short reads, BMC Bioinf. 20 (2019) 405.

[58] F. Madeira, Y.M. Park, J. Lee, N. Buso, T. Gur, N. Madhusoodanan, P. Basutkar, A. R.N. Tivey, S.C. Potter, R.D. Finn, R. Lopez, The EMBL-EBI search and sequence analysis tools APIs in 2019, Nucleic Acids Res. 47 (2019) W636–W641.

[59] E. Capriotti, R.B. Altman, Y. Bromberg, Collective judgment predicts disease-associated single nucleotide variants, BMC Genom. 14 (Suppl 3) (2013) S2.

[60] E. Capriotti, P. Fariselli, R. Casadio, I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure, Nucleic Acids Res. 33 (2005) W306–W310.

[61] D. Xu, Y. Zhang, Toward optimal fragment generations for ab initio protein structure assembly, Proteins 81 (2013) 229–239.

[62] D. Xu, Y. Zhang, Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field, Proteins 80 (2012) 1715–1735.

[63] G.E. Crooks, G. Hon, J.-M. Chandonia, S.E.J.G.r. Brenner, WebLogo: a sequence logo generator 14 (2004) 1188–1190.

[64] J.A. Irving, R.N. Pike, A.M. Lesk, J.C. Whisstock, Phylogeny of the serpin superfamily: implications of patterns of amino acid conservation for structure and function, Genome Res. 10 (2000) 1845–1864.

[65] D. Peacock, D. Boulter, Use of amino acid sequence data in phylogeny and evaluation of methods using computer simulation, J. Mol. Biol. 95 (1975) 513–527.

[66] S.S. Hassan, P.P. Choudhury, B. Roy, S.S. Jana, Missense Mutations in SARS-CoV2 Genomes from Indian Patients, Genomics, 2020.

[67] S.S. Hassan, P.P. Choudhury, B. Roy, SARS-CoV2 Envelope Protein: Non-synonymous Mutations and its Consequences, Genomics, 2020.

[68] F. Austerlitz, O. David, B. Schaeffer, K. Bleakley, M. Olteanu, R. Leblois, M. Veuille, C. Laredo, DNA barcode analysis: a comparison of phylogenetic and statistical classification methods, BMC Bioinf. 10 (Suppl 14) (2009) S10.

[69] N. Saitou, M. Nei, The neighbor-joining method: a new method for reconstructing phylogenetic trees, Mol. Biol. Evol. 4 (1987) 406–425.

[70] Z. Obradovic, K. Peng, S. Vucetic, P. Radivojac, A.K. Dunker, Exploiting heterogeneous sequence properties improves prediction of protein disorder, Proteins 61 (Suppl 7) (2005) 176–182.

[71] X. Fan, L. Kurgan, Accurate prediction of disorder in protein chains with a comprehensive and empirically designed consensus, J. Biomol. Struct. Dyn. 32 (2014) 448–464.

[72] F. Meng, V.N. Uversky, L. Kurgan, Comprehensive review of methods for prediction of intrinsic disorder and its molecular functions, Cell. Mol. Life Sci. 74 (2017) 3069–3090.

[73] Z.L. Peng, L. Kurgan, Comprehensive comparative assessment of in-silico predictors of disordered regions, Curr. Protein Pept. Sci. 13 (2012) 6–18.

[74] F. Meng, V. Uversky, L. Kurgan, Computational prediction of intrinsic disorder in proteins, Curr. Protein Pept. Sci. 88 (2017) 2 16 11–12 16 14.