Contents lists available at ScienceDirect

Virus Research

journal homepage: www.elsevier.com/locate/virusres

Pathogenic perspective of missense mutations of ORF3a protein of SARS-CoV-2

Sk. Sarif Hassan [a],*, Diksha Attrish [b], Shinjini Ghosh [c], Pabitra Pal Choudhury [d], Bidyut Roy [e]

[a] *Department of Mathematics, Pingla Thana Mahavidyalaya, Maligram 721140, India*
[b] *Dr. B. R. Ambedkar Centre For Biomedical Research (ACBR), University of Delhi (North Campus), Delhi 110007, India*
[c] *Department of Biophysics, Molecular Biology and Bioinformatics, University of Calcutta, Kolkata 700009, West Bengal, India*
[d] *Applied Statistics Unit, Indian Statistical Institute, Kolkata 700108, West Bengal, India*
[e] *Human Genetics Unit, Indian Statistical Institute, Kolkata 700108, West Bengal, India*

ARTICLE INFO

Keywords:
SARS-CoV-2
ORF3a
COVID-19
Missense mutations
Shannon entropy
Genetic variations

ABSTRACT

One of the most important proteins for COVID-19 pathogenesis in SARS-CoV-2 is the ORF3a which is the largest accessory protein among others coded by the SARS-CoV-2 genome. The major roles of the protein include virulence, infectivity, ion channel activity, morphogenesis, and virus release. The coronavirus, SARS-CoV-2 is mutating rapidly, therefore, critical study of mutations in ORF3a is certainly important from the pathogenic perspective. Here, a sum of 175 non-synonymous mutations in the ORF3a of SARS-CoV-2 were identified from 7194 complete genomes of SARS-CoV-2 available from NCBI database. Effects of these mutations on structural stability, and functions of ORF3a were also studied. Broadly, three different classes of mutations, such as neutral, disease, and mixed (neutral and disease) types of mutations were observed. Consecutive phenomena of mutations in ORF3a protein were studied based on the timeline of detection of the mutations. Considering the amino acid compositions of the ORF3a protein, twenty clusters were detected using the K-means clustering method. The present findings on 175 novel mutations of ORF3a proteins will extend our knowledge on ORF3a, a vital accessory protein in SARS-CoV-2, to enlighten the pathogenicity of this life-threatening virus.

1. Introduction

Severe Acute Respiratory Syndrome (SARS-CoV) emerged in 2002 infecting about 8000 people with a 10% mortality rate (Guarner, 2020; Huang et al., 2004). Similarly, Middle East Respiratory Syndrome Coronavirus (MERS-CoV) emerged in 2012 with 2300 cases, and a 35% mortality rate (Al-Osail and Al-Wazzah, 2017). However, since the December 2019, another outbreak caused by a novel severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) rapidly became a pandemic with high mortality rate within just 7 months; urging the World Health Organization to declare it as a Public Health Emergency of International Concern (Perrella et al., 2020; Hintze et al., 2020; Fiorino et al., 2020; Harapan et al., 2020). It was found that SARS-CoV and SARS-CoV-2 bear 79% of sequence identity (Van Doremalen et al., 2020; Andersen et al., 2020). Similar to the SARS-CoV, the ORF3a gene in SARS-CoV-2 lies between the spike and envelope gene in virus genome (Law et al., 2005). The ORF3a protein of SARS-CoV and SARS-CoV-2 contain a conserved cysteine residue which helps in protein-protein interaction (Meitzler et al., 2013; To and Torres, 2018). The RNA genome of SARS-CoV-2 is about 30 kb in length and codes for four structural proteins, 16 non-structural proteins, and six accessory proteins (Tang et al., 2020; Shen et al., 2020; Phan, 2020; Zhang and Holmes, 2020). The structural proteins are known as Spike protein (S), Nucleocapsid protein (N), Membrane protein (M), and Envelope protein (E) (Buchholz et al., 2004).

Here, we studied reported mutations in ORF3a, the largest accessory protein, and a unique membrane protein consisting of three transmembrane domains (Gao et al., 2020; Lu et al., 2010). SARS-CoV-2 ORF3a is a 275 amino acid transmembrane protein that holds an N-terminal, three transmembrane helices followed by a cytosolic domain with multiple β-strands (Lu et al., 2006). Functionally ORF3a proteins is divided into six domains (Siu et al., 2019). Domain-I contains N terminus signal peptide involved in subcellular localization of ORF3a protein (Lu et al., 2010). Domain-II contains a TNF receptor-associated

* Corresponding author.
E-mail addresses: sarimif@gmail.com (Sk.S. Hassan), dikshaattrish@gmail.com (D. Attrish), shinjinighosh2014@gmail.com (S. Ghosh), pabitrapalchoudhury@gmail.com (P.P. Choudhury), broy@isical.ac.in (B. Roy).

https://doi.org/10.1016/j.virusres.2021.198441
Received 13 January 2021; Received in revised form 26 April 2021; Accepted 27 April 2021
Available online 30 April 2021

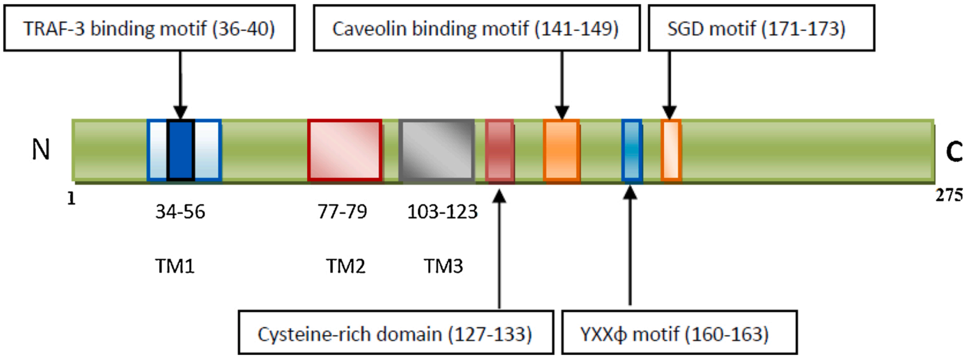

TRAF-3 binding motif (36-40)
Caveolin binding motif (141-149)
SGD motif (171-173)
N
C
1
275
34-56
77-79
103-123
TM1
TM2
TM3
Cysteine-rich domain (127-133)
YXXϕ motif (160-163)

Fig. 1. Schematic view of the domains in primary sequence of ORF3a protein.

factor 3 (TRAF-3) binding motif (amino acid (aa) 36-40) through which it activates the NF-kB, and NLRP3 inflammasome by promoting TRAF3-mediated ubiquitination of apoptosis-associated speck-like protein containing a caspase recruitment domain (ASC) (Siu et al., 2019). Domain-III (aa 93-133) is important for ion channel activity, and has a Cysteine-rich domain which is associated with homodimerization of ORF3a protein similar to SARS-CoV cysteine rich domain responsible for tetramerization (aa 81-160) (Wang et al., 2011; Issa et al., 2020). Domain-IV has a caveolin binding motif (aa 141-149) which regulates viral uptake and trafficking of protein to the plasma membrane or intracellular membranes (Padhan et al., 2007). Domain-V contains a tyrosine-based sorting motif *YXXϕ* (aa 160-163) responsible for Golgi to plasma membrane transport (Minakshi and Padhan, 2014). Domain-VI has an SGD motif (aa 171-173) (Issa et al., 2020). ORF3a has pro-apoptotic activity, and membrane association is required for this activity. SARS-CoV-2 ORF3a has relatively weaker proapoptotic activity, and this property is probably contributing to asymptomatic infection, and thus causing rapid transmission of the virus (Ren et al., 2020). Therefore, ORF3a may become an important therapeutic target, and thus studying mutations in the ORF3a protein sequence becomes an important area in control of virus infection (Fig. 1).

In our present study, we found 175 non-synonymous mutations in the ORF3a protein sequence from 7194 complete SARS-CoV-2 genomes. Among them, 32 were already reported previously (Hassan et al., 2020; Issa et al., 2020). So, we analyzed 143 new mutations along with the already existing ones. Mutations in the domain-III alters the NF-kB activation, and NLRP3 inflammasome. Mutations in domain-V were linked to the aggregation of the ORF3a protein in the Golgi apparatus (Smirnova et al., 2015). Apart from these residues, mutations at the position 230 (insertion of F), W131C, R134L, T151I, N152S, and D155Y regions may contribute to a greater significance as they are poised to form a network of hydrophobic, polar, and electrostatic interactions mediating dimerization, and tetramerization respectively (Kern et al., 2020). For this study, we collected the SARS-CoV-2 genome data from NCBI virus database, identified the mutations, predicted the effect of mutations based on chemical and structural properties. In addition, using the Meta-SNP and I-MUTANT web-servers, effects of the mutations in functions and structures were predicted (Capriotti et al., 2013, 2005; Hall et al., 2009). We also performed K-means clustering of the distinct variants ORF3a proteins in order to form twenty disjoint clusters based on the amino acid compositions embedded in the proteins (Likas et al., 2003; Zhong et al., 2005). In addition, Shannon entropy was calculated to determine the amount of disorderliness of the amino acids over the ORF3a proteins which had the wide distinct variations of ORF3a in the USA (Strait and Dewey, 1996).

2. Data and methods

This present study is based on available genome data of SARS-CoV-2 from the NCBI virus database. Here we discuss about data followed by methods which are employed in this study.

2.1. Data

Among 7194 complete genomes of SARS-CoV-2 taken from the NCBI database, only 296 sequences are found to be distinct from each other. The amino acid sequences of ORF3a were exported in fasta format using file operations through Matlab (The Mathworks Inc, 2020). In this present study, we only concentrate on 296 ORF3a proteins which are listed in Tables 1 and 2 . Note that, among these 296 sequences, three ORF3a proteins QKO00487 (India: Ahmedabad), QLA10225 (India: Vadodara) and QLA10069 (India: Surat) had the length 241, 253, and 257 respectively, and were found to be truncated due to nonsense mutation at 242, 254, and 258 amino acid positions respectively. It is also note worthy that some (13.51%) of 296 ORF3a amino acid sequences contain ambiguous amino acids such as *X*, *B*, and *Z*, and so on. In order to find mutations, we hereby consider the reference ORF3a protein as the ORF3a sequence (YP_009724391.1) of the SARS-CoV-2 genome (NC_045512) from China: Wuhan (Wang et al., 2020).

2.2. Methods

Here in a nutshell, we present the methods used in this study.

2.2.1. Frequency probability of amino acids

A protein sequence of ORF3a is composed of twenty different amino acids with various frequencies (starting from zero). The probability of occurrence of each amino acid A_i is determined by the formula $\frac{f(A_i)}{l}$ where $f(A_i)$ denotes the frequency of occurrence of the amino acid A_i in the primary sequence ORF3a, and l stands as the length of ORF3a protein (Brooks et al., 2002). Hence for each of the 296 ORF3a proteins, a twenty dimensional vector considering the frequency probability of twenty amino acids can be obtained. Based on these frequency probability vectors, a classification is performed using clustering technique.

2.2.2. K-means clustering algorithm

Clustering is one of the most widely used methods in vector-data analysis to develop an intuitive idea about closeness of data based on the structured feature vectors. By clustering we find homogeneous subclasses within the data such that data points in each cluster are as similar as possible according to a similarity measure such as euclidean-based distance. One of the most commonly used simple clustering techniques is the *K-means clustering* (Likas et al., 2003; Zhong et al., 2005).

Algorithm: K-means algorithm is an iterative algorithm that tries to form equivalence classes from the feature vectors into K (pre-defined) clusters where each data point belongs to only one cluster (Likas et al., 2003).

Table 1
List of accessions of the ORF3a protein, geo-location and respective data collection date.

Accession	Geo_Location	Collection_Date	Accession	Geo_Location	Collection_Date	Accession	Geo_Location	Collection_Date
YP_009724391	***China: Wuhan***	***2019-12***	QLH00578	USA: CA	2020-04-23	QLA10225	India: Vadodara	2020-06-02
QLI49698	India: Himatnagar	2020-06-14	QLH01238	USA: CA	2020-04-21	QKY77929	USA: CA	2020-03-16
QLI50222	USA: New York, Rockland county	2020-06-26	QLH01250	USA: CA	2020-04-22	QKY59990	India: Surat	2020-06-11
QLI50282	USA: Wisconsin, Dane county	2020-06-26	QLH01298	USA: CA	2020-04-22	QKX46204	USA	2020-05-11
QLI50414	USA: Wisconsin, Dane county	2020-06-28	QLH01334	USA: CA	2020-04-24	QKX47995	Bangladesh: Rangpur	2020-06-07
QLI50570	USA: Wisconsin, Dane county	2020-06-27	QLH01382	USA: CA	2020-05-04	QKX49024	Bangladesh	2020-05-23
QLI51038	USA: Wisconsin, Dane county	2020-06-30	QLH01502	USA: CA	2020-05-04	QKW88844	USA	2020-03-14
QLI51614	USA: Wisconsin, Dane county	2020-05-09	QLF97736	Bangladesh	2020-06-17	QKW89480	USA	2020-03-25
QLI51746	USA: Wisconsin, Ozaukee county	2020-03-19	QLF97772	Bangladesh	2020-06-18	QKV35400	USA: Washington, Yakima County	2020-04-15
QLI51782	USA: Wisconsin, Fond du Lac county	2020-03-31	QLF97844	Bangladesh	2020-06-18	QKV35688	USA: Washington, Yakima County	2020-04-13
QLI46290	USA: Arkansas, Little Rock	2020-04-01	QLF97952	India: Vadodara	2020-06-08	QKV36900	USA: Washington, Yakima County	2020-04-11
QLH64816	India: Modasa	2020-06-14	QLF98036	Bangladesh	2020-06-17	QKV37633	Australia: Victoria	2020-03-24
QLH93202	India: Surat	2020-06-13	QLF98048	Bangladesh	2020-06-17	QKV38005	Australia: Northern Territory	2020
QLH93429	Bangladesh: Jashore	2020-07-07	QLF98084	India: Talod	2020-06-19	QKV38209	Australia: Victoria	2020-04-10
QLH93441	Bangladesh: Jashore	2020-07-07	QLF98201	India: Rajkot	2020-06-12	QKV38257	Australia: Victoria	2020-04-10
QLH93453	Bangladesh: Jashore	2020-07-07	QLF98261	India: Surat	2020-06-11	QKV38281	Australia: Victoria	2020-04-11
QLH55720	Bangladesh: Barishal	2020-07-06	QLF99991	USA: MD	2020-04-01	QKV38401	Australia: Victoria	2020-04-13
QLH55768	Bangladesh: Barishal	2020-07-06	QLF78310	Poland	2020-06-01	QKV38810	USA: Washington, Snohomish County	2020-04-18
QLH55816	Bangladesh: Barishal	2020-07-06	QLF80217	Brazil	2020-03-13	QKV38894	USA: Washington, Yakima County	2020-05-03
QLH55840	Bangladesh: Barishal	2020-07-06	QLF95245	USA: Virginia	2020-03	QKV39324	USA: Washington, King County	2020-04-27
QLH56099	Saudi Arabia	2020-02-10	QLF95641	USA: Virginia	2020-03	QKV39588	USA: Washington, Snohomish County	2020-04-27
QLH56231	Saudi Arabia	2020-02-26	QLF95737	USA: Virginia	2020-03	QKV39840	USA: Washington, Yakima County	2020-05-06
QLH56255	Saudi Arabia	2020-03-01	QLF95773	USA: Virginia	2020-03	QKV40164	USA: Washington, Yakima County	2020-05-03
QLH56279	Bangladesh: Barishal	2020-07-06	QLE11150	Bangladesh	2020-06-18	QKV40440	USA: Washington, Yakima County	2020-05-06
QLH57751	USA: FL	2020-04-14	QLC91545	USA: Wisconsin, Dane County	2020-03-20	QKV40716	USA: Washington, Yakima County	2020-05-05
QLH57846	USA: FL	2020-04-14	QLC91617	USA: Wisconsin, Dane County	2020-03-19	QKV41592	USA: Washington, Cowlitz County	2020-04-22
QLH58037	USA: FL	2020-04-16	QLC91905	USA: Wisconsin, Dane County	2020-03-24	QKV41616	USA: Washington, Benton County	2020-04-28
QLH58085	USA: FL	2020-04-16	QLC92097	USA: Wisconsin, Dane County	2020-03-31	QKV42204	USA: Washington	2020-04-26
QLH58601	USA: FL	2020-05-14	QLC92421	USA: Wisconsin	2020-04-02	QKV42875	USA: Washington, Cowlitz County	2020-04-27
QLH58947	USA: FL	2020-06-02	QLC92553	USA: Wisconsin, Richland county	2020-04-08	QKV42947	USA: Washington, Yakima County	2020-04-29
QLH59007	USA: FL	2020-06-03	QLC92601	USA: Wisconsin, Dane County	2020-04-09	QKV26659	USA: Virginia	2020-05
QLG75126	Bahrain	2020-06-22	QLC93129	USA: Wisconsin, Milwaukee county	2020-03-21	QKS89844	USA: Washington, King County	2020-03-04
QLG75678	Australia: Victoria	2020-06-01	QLC93357	USA: Wisconsin, Waukesha county	2020-03-24	QKS90192	USA: Washington, King County	2020-02-29
QLG75822	Australia: Victoria	2020-06-06	QLC94305	USA: Wisconsin, Milwaukee county	2020-04-13	QKU28463	USA: Washington, King County	2020-03-03
QLG75930	Australia: Victoria	2020-06-11	QLC94473	USA: Wisconsin, Milwaukee county	2020-04-14	QKU28847	USA: Washington, King County	2020-04-29
QLG75942	Australia: Victoria	2020-06-11	QLC94737	USA: Wisconsin, Milwaukee county	2020-03-24	QKU29039	USA: Washington, King County	2020-04-19
QLG76026	Australia: Northern Territory	2020	QLC46314	USA: FL	2020-04-03	QKU30570	USA: Washington	2020-04-16
QLG76386	Australia: Victoria	2020-06-19	QLC46986	USA: FL	2020-04-21	QKU31182	USA: CA	2020-04-02
QLG76542	Australia: Victoria	2020-06-20	QLC47346	USA: FL	2020-05-03	QKU31266	USA: CA	2020-04-11
QLG97055	Italy	2020-04-04	QLB39261	USA	2020-04-06	QKU31638	USA: CA	2020-03-20
QLG97460	USA: Wisconsin, Dane county	2020-06-15	QLB39321	USA	2020-04-11	QKU31746	USA: CA	2020-03-25
QLG97484	USA: Wisconsin, Dane county	2020-06-14	QLA47500	USA: Virginia	2020-05	QKU31806	USA: CA	2020-03-30

(*continued on next page*)

Table 1 (*continued*)

Accession	Geo_Location	Collection_Date	Accession	Geo_Location	Collection_Date	Accession	Geo_Location	Collection_Date
QLG98012	USA: Wisconsin, Jackson county	2020-06-01	QLA47776	USA: Virginia	2020-05	QKU31818	USA: CA	2020-03-30
QLG99677	USA: CA	2020-06-03	QKR84274	Egypt	2020-06-02	QKU32046	USA: CA	2020-05-01
QLG99737	USA: CA	2020-04-16	QKR84421	Egypt	2020-06-02	QKU32202	USA: CA	2020-03-30
QLG99773	USA: CA	2020-04-16	QKS66941	Egypt	2020-06-02	QKU32934	USA: CA	2020-03-24
QLH00026	USA: CA	2020-04-27	QLA09656	USA: Ak	2020-03-23	QKU32982	USA: CA	2020-03-26
QLH00290	USA: CA	2020-04-28	QLA10069	India: Surat	2020-06-11	QKU37034	Saudi Arabia: Jeddah	2020-03-15
QLH00362	USA: CA	2020-04-17	QLA10165	India: Kapadvanj	2020-06-08	QKU37202	USA: CA	2020-04-18

- Assign the number of desired clusters (K) (in the present study, $K = 20$).
- Finding centroids by first shuffling the dataset, and then randomly selecting K data points for the centroids without replacement.
- Keep iterating until there is no change to the centroids.
- Find the sum of the squared distance between data points and all centroids.
- Assign each data point to the closest cluster (centroid).
- Compute the centroids for the clusters by taking the average of the all data points that belong to each cluster.

In this study we did clustering using Matlab by customizing the value of K and inputting the frequency of amino acid compositions over the ORF3a proteins.

2.2.3. Amino acid conservation Shannon entropy

How conserved/disordered the amino acids are, over ORF3a protein is addressed by the information theoretic measure known as '*Shannon entropy*(SE)' which we deploy here to find out conservation entropy of each ORF3a protein. For each ORF3a protein, Shannon entropy of amino acid conservation over the amino acid sequence of ORF3a protein is computed using the following formula (Johansson and Toh, 2010):

For a given amino acid sequence of ORF3a protein of length l, the conservation of amino acids is calculated as follows:

$$\mathrm{SE} = -\sum_{i=1}^{20} p_{s_i} \log_{20}(p_{s_i})$$

where $p_{s_i} = \frac{k_i}{l}$; k_i represents the number of occurrences of an amino acid s_i in the given sequence.

In this study, SE describes the wide variety of 296 distinct ORF3a proteins collected from various countries across the world.

3. Results

All mutations, compared to Chinese-Wuhan sequence, over the set of distinct ORF3a proteins were detected, and consequently they have been classified based on their predicted effect as disease/neutral in important functions of ORF3a protein (Table 9). Also, some important known domains are identified for the observed mutations, and accordingly the predicted effect of mutations in protein functions have been discussed. Further, consecutive mutations observed in ORF3a proteins according to the timelines of detection of various mutations for a subgroup of ORF3a proteins located in Australia, Bangladesh, India, USA are derived (Fig. 7–11). Using a web-server ($i - MUTANT$: http://gpcr2.biocomp.unibo.it/cgi/predictors/I-Mutant3.0/I-Mutant3.0.cgi) stability of ORF3a protein structures were predicted due to various mutations. At last, twenty clusters are formed using the K-means clustering method based on frequency probability of amino acids of 296 ORF3a proteins. The wide variations of 296 ORF3a proteins are finally supported by the Shannon entropy (SE), and remarkably we found that the most variations in ORF3a proteins was detected in the viruses reported in the USA.

3.1. Mutations over the ORF3a protein of SARS-CoV-2

Each of the ORF3a amino acid sequences (fasta formatted) are aligned with respect to the ORF3a protein (YP_009724391.1) from China-Wuhan using multiple sequence alignment tool (NCBI Blastp suite), and found the mutations, and their associated positions (Madeira et al., 2019). It is noted that a mutation from an amino acid A_1 to A_2 at a position p is denoted by A_1pA_2 or $A_1(p)A_2$. Fig. 2 describes various mutations with their respective locations. The mutations are found in the entire ORF3a sequence starting from the amino acid position 7 to 271. It is found that an amino acid at a fixed position mutated to two different amino acids. For examples, at 9th position of the reference ORF3a protein, the amino acid Threonine (T) maps to Isoleucine (I), and Lysine (K) in two different ORF3a proteins. At the 18th position Glycine maps to three amino acids Valine, Serine, and Cysteine in three ORF3a proteins. The amino acid Alanine (A) maps to Valine, Serine, Threonine, and Aspartic acid at the 99th position in four ORF3a proteins.

Based on observed mutations, it is noticed that amino acids Alanine (A), and Tryptophan (W) are found to be most vulnerable to mutate to various amino acids. It is noted that the mutation of Tryptophan (W) at 131 position are found in the Cysteine-rich domain (127–133).

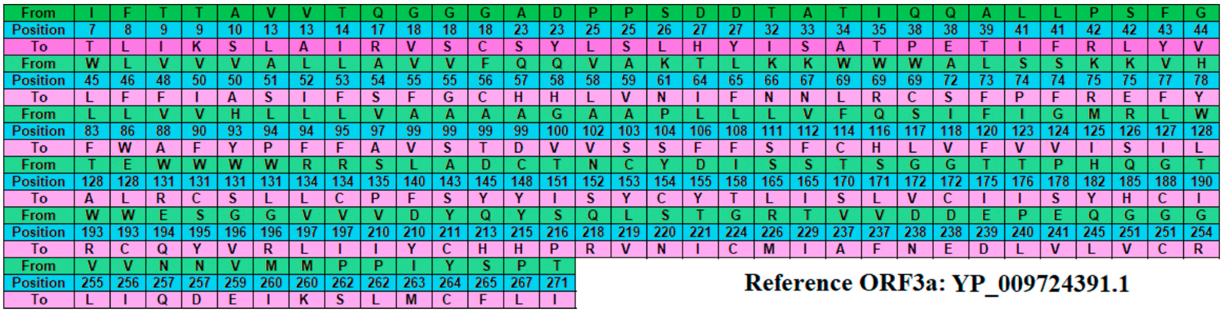
Distinct mutations and its respective mutation of frequency are presented in Table 3. The most frequent mutation over the ORF3a is to be Q57H (Acidity: Neutral (Q) to Basic (weakly)(H)) with frequency 142. A pie chart accounting the frequency distribution of various mutations is shown in Fig. 3. In addition to the list of mutations (Fig. 2), two deletion, and two insertion mutations were found in five different ORF3a proteins at various positions.

Among 296 ORF3a proteins, 40 sequences possess few ambiguous mutations which we have not considered for analysis. The details of mutations, in the 256 ORF3a unique proteins from viruses of 256 patients, in specific domain(s), and predicted effects of mutations viz. disease, and neutral effects through the web-server Meta-SNP (https://snps.biofold.org/meta-snp/) are presented in Table 4 (in multiple). Note that the META-SNP web-server integrates other SNP prediction tools such as SNAP, SIFT, PANTHER, PHD-SNP. It is worth mentioning that the effect of SNPs are also checked using PredictSNP (https://loschmidt.chemi.muni.cz/predictsnp1/), which also integrates various other tools such as Polyphen-1,2, PHD-SNP, SIFT, SNAP, MAPP, etc.) and found same results as we obtained using Meta-SNP. As an example, the mutation Q57H turned out to be 'Disease' (0.637) in the META-SNP server and in the PredictSNP server, Q57H shows as 'Deleterious' with 76% confidence. Note that among 296 ORF3a proteins, 40 sequences possess only ambiguous mutations which we have neglected. A snapshot of predicted result (disease causing variant with reliability score 3) of the most frequent mutation Q57H is shown in Fig. 4.

Based on the predicted type of mutations, all the 256 ORF3a proteins are classified into three classes which are presented in Table 5. The three

Table 2
List of accessions of the ORF3a protein, geo-location and respective data collection date.

Accession	Geo_Location	Collection_Date	Accession	Geo_Location	Collection_Date	Accession	Geo_Location	Collection_Date
QKU37646	USA: CA	2020-04-02	QKG86518	USA	2020-04	QJR88822	Australia: Victoria	2020-03-20
QKU52834	USA: Washington,King County	2020-03-18	QKE61733	India: Rajkot	2020-04-28	QJR89110	Australia: Victoria	2020-03-22
QKU52870	USA: Washington,Snohomish County	2020-03-16	QKE44990	USA	2020-04	QJR89278	Australia: Victoria	2020-03-23
QKU53050	USA: Washington	2020-03-20	QKE45765	USA: CA	2020-04-26	QJR89362	Australia: Victoria	2020-03-23
QKU53650	USA: Washington,King County	2020-03-17	QKE45861	USA: CA	2020-04-30	QJR89446	Australia: Victoria	2020-03-24
QKU53854	USA: Washington,King County	2020-03-07	QKE45885	USA: CA	2020-04-30	QJR91282	Australia: Victoria	2020-03-26
QKV06224	USA: Washington,Yakima County	2020-04-02	QKE45933	USA: CA	2020-04-29	QJR91354	Australia: Victoria	2020-03-29
QKV06236	USA: Washington,Pierce County	2020-03-31	QKE10935	Czech Republic	2020-03-31	QJR95110	Australia: Victoria	2020-04-08
QKV06920	USA: Washington,Pierce County	2020-03-31	QJY78272	USA	2020-03-20	QJQ84173	USA: NEW ORLEANS, LA	2020-04-04
QKV07184	USA: Washington,King County	2020-03-31	QKC05357	USA	2020-03-11	QJQ38625	USA: CA	2020-04-22
QKV07340	USA: Washington,Yakima County	2020-04-02	QJY40110	USA	2020-03-17	QJQ39045	USA: MI	2020-03-13
QKV07400	USA: Washington,Yakima County	2020-03-26	QJY40506	India: Junagadh	2020-05-09	QJQ39081	USA: MI	2020-03-16
QKV08048	USA: Washington,King County	2020-03-31	QJX68859	USA: Michigan	2020-03-16	QJQ39297	USA: MI	2020-03-18
QKS65597	USA: CA	2020-03-15	QJX70192	USA: Michigan	2020-03-30	QJQ39741	USA: MI	2020-03-25
QKS65621	USA: CA	2020-03-15	QJX70592	USA: Illinois	2020-04-14	QJI07211	USA: VA	2020-04
QKS65777	USA: CA	2020-03-16	QJX45032	USA: CA	2020-03-23	QJI54123	USA: CA	2020-03-05
QKS65849	USA: MA	2020-03-15	QJX45308	Poland	2020-04-11	QJI54254	USA: CA	2020-03-03
QKS66041	USA: NJ	2020-03-14	QJW00412	India: Gandhinagar	2020-05-02	QJF75396	USA: Michigan	2020-03-20
QKS66053	USA: NJ	2020-03-14	QJX44383	India: Ahmedabad	2020-04-29	QJF77147	USA: WA	2020-04-02
QKS66305	USA: UT	2020-03-12	QJX44407	India: Ahmedabad	2020-04-29	QJE38451	USA: CA	2020-03-28
QKS66737	USA: NY	2020-03-15	QJW69308	Germany: Bavaria	2020-02	QJD47203	USA: WA	2020-03-26
QKS67001	USA	2020-04-09	QJU70306	USA: AK	2020-03-23	QJD47299	USA: WA	2020-03-28
QKS67456	China	2020-01-23	QJV21807	USA: CA	2020-04-01	QJD47419	USA: WA	2020-04-05
QJY78153	Egypt	2020-05-02	QJW28449	USA: VA	2020-04	QJD47539	USA: CT	2020-04-07
QKQ63773	USA: Virginia	2020-04	QJW28665	USA: VA	2020-04	QJD47551	USA: CT	2020-04-06
QKO25735	Bangladesh: Dhaka	2020-06-01	QJU11458	USA: FL	2020-03-06	QJD47849	Taiwan	2020-03-16
QKO25747	Bangladesh: Dhaka	2020-06-01	QJT72327	France	2020-03	QJD47873	Taiwan	2020-03-18
QKO00487	India: Ahmedabad	2020-05-27	QJT72387	France	2020-03	QJD47956	USA: WA	2020-03-10
QKN19672	USA: Michigan	2020-04-26	QJT72471	France	2020-03	QJD48484	USA: WA	2020-03-13
QKN20740	USA	2020-04-04	QJT72507	France	2020-03	QJD20838	Sri Lanka	2020-03-16
QKN20812	USA	2020-04-03	QJT72951	France	2020-03	QJD23478	USA: NY	2020-03-18
QKN20824	USA	2020-04-04	QJS53735	Greece: Athens	2020-03-12	QJD23730	USA: NY	2020-03-18
QKM76547	Germany: Dusseldorf	2020-03-15	QJS53831	Greece: Athens	2020-03-13	QJD25758	USA: NY	2020-03-19
QKM76907	Germany: Heinsberg	2020-02-28	QJS54023	Greece: Athens	2020-03-12	QJC19648	USA: WA	2020-03-31
QKK12852	Bangladesh	2020-05-23	QJS54155	Greece: Athens	2020-03-08	QJC20380	USA: WA	2020-03-27
QKK14612	USA	2020-05-11	QJS54191	Greece: Athens	2020-03-23	QJC20500	USA: WA	2020-03-30
QKG87087	USA: Massachusetts	2020-04-01	QJS54383	Greece: Athens	2020-03-10	QJA17681	USA: PA	2020-03-07
QKG87159	USA: Massachusetts	2020-04-02	QJS54923	USA: CA	2020-04-30	QIZ13336	USA	2020-03-23
QKG87195	USA: Massachusetts	2020-03-27	QJS57052	USA: WA	2020-04-03	QIZ13838	USA	2020-03-22
QKG87267	USA: Massachusetts	2020-04-01	QJS39520	Netherlands	2020-04-29	QIZ14498	USA	2020-03-21
QKG88539	USA: Massachusetts	2020-04-02	QJS39568	Netherlands	2020-04-29	QIZ16438	USA: MA	2020-03-06
QKG88935	USA: Massachusetts	2020-04-01	QJS39616	Netherlands	2020-05-06	QIZ16548	Greece	2020-03-18
QKG90147	USA: Massachusetts	2020-03-21	QJR84550	USA: CA	2020-04-01	QIU78768	Spain	2020-03-02
QKG90399	USA: Massachusetts	2020-03-26	QJR84790	USA: CA	2020-04-13	QIU81286	USA: WA	2020-03-17
QKG90495	USA: Massachusetts	2020-03-26	QJR86050	Australia: Victoria	2020-03-15	QIS61075	USA: IL	2020-03-13
QKG90867	USA: Massachusetts	2020-03-25	QJR87574	Australia: Victoria	2020-03-20	QIS61315	USA: WA	2020-03-16
QKG91107	USA: Massachusetts	2020-03-27	QJR87598	Australia: Victoria	2020-03-21	QIS30116	USA: San Francisco, CA	2020-03-18
QKG64052	USA	2020-04	QJR87730	Australia: Victoria	2020-03-21	QII57239	USA	2020-02-25
QKG81824	USA: Virginia	2020-04	QJR88306	Australia: Victoria	2020-03-23	QHZ00380	South Korea	2020-01
QKG81932	USA: Virginia	2020-04	QJR88390	Australia: Victoria	2020-03-23			



From	I	F	T	T	A	V	V	T	Q	G	G	G	A	D	P	P	S	D	D	T	A	T	I	Q	Q	A	L	L	P	S	F	G
Position	7	8	9	9	10	13	13	14	17	18	18	18	23	23	25	25	26	27	27	32	33	34	35	38	38	39	41	41	42	42	43	44
To	T	L	I	K	S	L	A	I	R	V	S	C	S	Y	L	S	L	H	Y	I	S	A	T	P	E	T	I	F	R	L	Y	V
From	W	L	V	V	V	A	L	L	A	V	V	F	Q	Q	V	A	K	T	L	K	K	W	W	W	A	L	S	S	K	K	V	H
Position	45	46	48	50	50	51	52	53	54	55	55	56	57	58	58	59	61	64	65	66	67	69	69	69	72	73	74	74	75	75	77	78
To	L	F	F	I	A	S	I	F	S	F	G	C	H	H	L	V	N	I	F	N	N	L	R	C	S	F	P	F	R	E	F	Y
From	L	L	V	V	H	L	L	L	V	A	A	A	A	G	A	A	P	L	L	L	V	F	Q	S	I	F	I	G	M	R	L	W
Position	83	86	88	90	93	94	94	95	97	99	99	99	99	100	102	103	104	106	108	111	112	114	116	117	118	120	123	124	125	126	127	128
To	F	W	A	F	Y	P	F	F	A	V	S	T	D	V	V	S	S	F	F	S	F	C	H	L	V	F	V	V	I	S	I	L
From	T	E	W	W	W	W	R	R	S	L	A	D	C	T	N	C	Y	D	I	S	S	T	S	G	G	T	T	P	H	Q	G	T
Position	128	128	131	131	131	131	134	134	135	140	143	145	148	151	152	153	154	155	158	165	165	170	171	172	172	175	176	178	182	185	188	190
To	A	L	R	C	S	L	L	C	P	F	S	Y	Y	I	S	Y	C	Y	T	L	I	S	L	V	C	I	I	S	Y	H	C	I
From	W	W	E	S	G	G	V	V	V	D	Y	Q	Y	S	Q	L	S	T	G	R	T	V	V	D	D	E	P	E	Q	G	G	G
Position	193	193	194	195	196	196	197	197	210	210	211	213	215	216	218	219	220	221	224	226	229	237	237	238	238	239	240	241	245	251	251	254
To	R	C	Q	Y	V	R	L	I	I	Y	C	H	H	P	R	V	N	I	C	M	I	A	F	N	E	D	L	V	L	V	C	R
From	V	V	N	N	V	M	M	P	P	I	Y	S	P	T																		
Position	255	256	257	257	259	260	260	262	262	263	264	265	267	271																		
To	L	I	Q	D	E	I	K	S	L	M	C	F	L	I																		

Reference ORF3a: YP_009724391.1

Fig. 2. Mutations in the respective position in ORF3a protein sequence compared with reference Wuhan sequence YP_009724391.1. **Note:** From: existing amino acid in reference sequence; position: amino acid position in the sequence; To: mutated amino acid in studied sequence.

classes representing disease, neutral, and mixture of disease as well as neutral mutations are constituted of protein IDs with respective geo-locations.

Almost 72% of the ORF3a proteins possess disease type of mutations whereas 14% (of which two mutations: 12%, three mutations: 1.5%, and four mutations: 0.5%), and 14% of ORF3a proteins possess mixture type (i.e. both disease as well as neutral), and neutral types of mutations respectively (Fig. 5).

For each of the three types of mutations, we put the frequency and percentage of ORF3a proteins corresponding to each geo-locations as presented in Table 6.

Except the countries where the total number of occurred mutation is one, in the USA, the amount of disease (deleterious) mutations over the ORF3a proteins was found to be the highest (74.36%) among other countries. Accordingly, it is suggested that the mortality rate is expected to be high which is supported by the real-time data. On the other hand, the least amount (8.97%) of neutral mutations were also observed in the USA which is expected to be contributing to the weaker apoptotic activity of ORF3a, and this weaker activity may be responsible for asymptomatic or relatively mildly symptomatic cases thus causing rapid transmission of the virus.

In Fig. 6, the world maps are marked as per occurrence of different types of mutations in ORF3a variants.

3.2. Possible consecutive mutations in ORF3a proteins during its journey from China to other countries

Several ORF3a proteins (Table 4) contain more than one mutation, and maximally up to four mutations. It takes time for multiple mutations in a given ORF3a protein, and relying on time-line, and order occurrence of mutations several flows of consecutive mutations were derived. The predicted effects of these mutations on stability of the tertiary structure of the ORF3a proteins was determined in the flow of consecutive mutations (Table 7).

Flow of consecutive mutation-I: In the Australian region, it can be observed that the first mutation may have occurred in sequence QJR87730.1 with respect to the Wuhan sequence (YP_009724391.1) from Q to H at 57th position which is a disease type mutation, and also this mutation is having the highest frequency which may indicate that it has an important role to play in infectivity part of the virus. As we move along the flow, six ORF3a sequences were considered based on the consecutive time scale of detection that was found to have 2nd mutation on the background of initial Q57H mutation with reference to Wuhan sequence (YP 009724391.1) (Fig. 7).

In this flow of mutation, six ORF3a proteins possess various mutations as follows:

- In QKV38005.1, there is a mutation K75R which was found to be a diseased type. We have to consider disease type mutation which may change the function of the protein.
- In QLG75822.1, there is a mutation A23S which was found to be a neutral type with no polarity change. So this is a synonymous mutation from the functionality perspective.
- In QLG76542.1, there is a mutation V55G which was found to be a diseased type, and hydrophobicity changed to hydrophilicity. This indicates that there may be a functional importance of this mutation.
- In QJR95110.1, there is a mutation L140F which was found to be a diseased type with no polarity change. Since no polarity change is observed the type of amino acid remains same but the mutation effect becomes harmful for the host.
- In QLG75942.1, there is a mutation at M260I that was found to be a diseased type with no polarity change. This mutation may increase the virus virulence.
- In QKV37633.1, there is a mutation at P262S which was found to be a diseased type, and polarity changed from hydrophobic to hydrophilic. Consequently, it may account for change in structure of the protein.

Flow of consecutive mutation-II: The most frequent mutation Q57H occurred in the ORF3a protein QKU53050.1. In this network flow (Fig. 8) there are other nine sequences which are considered based on the succeeding time scale that was found to have 2nd level mutations along with Q57H.

- The ORF3a protein QKU30570.1 contains a mutation W131C which was found to be a diseased type, and polarity changed from hydrophobic to hydrophilic. This mutation might affect the function of the ORF3a protein.
- QIZ13838.1 possesses a mutation L95F which was found to be a diseased type with no polarity change.
- There is a mutation a V55F in QKU29039.1, which was found to be a diseased type with no polarity change. But the mutation may cause an increase in pathogenesis.
- In the protein QKU28847.1, a mutation M260I occurred which was found to be a diseased type with no polarity change, and hence functional change of ORF3a can be expected.
- In QLH58085.1, there is a mutation Q185H which was found to be a diseased type with no polarity change, and so the structure of ORF3a protein may vary.
- In QKG88539.1, there is a mutation at L108F which was found to be a neutral type with no polarity change. This mutation needs further investigation in order to confirm about its neutrality.
- In QLI50282.1, there is a mutation G18S which was found to be a neutral type, and polarity changed from hydrophobic to hydrophilic.

Table 3
Distinct mutations across the ORF3a proteins and their respective frequency.

Mutations	Q(57)H	G(251)V	A(23)S	H(78)Y	V(13)L	V(88)A	A(99)V	D(27)H	G(196)V	M(260)I	S(171)L	S(26)L
Frequency of mutations	**124**[a]	9	4	4	4	4	3	3	3	3	3	3
Mutations	T(175)I	V(237)A	V(55)F	A(54)S	A(99)S	D(155)Y	D(22)Y	Deletion (256)	G(172)C	G(18)S	G(18)V	G(224)C
Frequency of mutations	3	3	3	2	2	2	2	2	2	2	2	2
Mutations	G(254)R	H(182)Y	k(75)R	L(108)F	L(140)F	L(52)I	L(53)F	L(65)F	L(86)W	L(95)F	P(25)L	P(42)R
Frequency of mutations	2	2	2	2	2	2	2	2	2	2	2	2
Mutations	Q(185)H	Q(38)E	Q(38)P	R(134)L	T(151)I	T(271)I	T(32)I	T(9)I	V(197)L	W(128)L	W(131)C	W(131)R
Frequency of mutations	2	2	2	2	2	2	2	2	2	2	2	2
Mutations	D(238)N	G(172)V	I(123)V	L(106)F	L(111)S	S(117)L	A(103)S	A(103)V	A(10S)S	A(143)S	A(33)S	A(39)T
Frequency of mutations	1	1	1	1	1	1	1	1	1	1	1	1
Mutations	A(51)S	A(59)V	A(72)S	A(99)D	A(99)T	C(148)Y	C(153)Y	D(210)Y	D(238)E	D(27)Y	Deletion IGT (10-12)	E(194)Q
Frequency of mutations	1	1	1	1	1	1	1	1	1	1	1	1
Mutations	E(239)D	E(241)V	F (114)C	F(120)L	F(43)Y	F(56)C	G(100)V	G(18)C	G(188)C	G(196)R	G(224)V	G(251)C
Frequency of mutations	1	1	1	1	1	1	1	1	1	1	1	1
Mutations	G(44)V	H(93)Y	I(118)V	I(158)T	I(263)M	I(35)T	I(7)T	Insertion D (101)	Insertion F (230)	K(61)N	K(66)N	K(67)N
Frequency of mutations	1	1	1	1	1	1	1	1	1	1	1	1
Mutations	K(75)E	L(127)I	L(219)V	L(41)F	L(41)I	L(46)F	L(73)F	L(83)F	L(94)F	L(94)P	M(125)I	M(260)K
Frequency of mutations	1	1	1	1	1	1	1	1	1	1	1	1
Mutations	N(152)S	N(257)D	N(257)Q	P(104)S	P(178)S	P(240)L	P(25)S	P(262)L	P(262)S	P(267)L	Q(116)H	Q(17)R
Frequency of mutations	1	1	1	1	1	1	1	1	1	1	1	1
Mutations	Q(213)H	Q(218)R	Q(245)L	R(126)M	R(126)S	R(134)C	S(135)P	S(165)F	S(165)I	S(165)L	S(195)Y	S(216)P
Frequency of mutations	1	1	1	1	1	1	1	1	1	1	1	1
Mutations	S(220)N	S(40)L	S(74)F	S(74)P	T(128)A	T(14)I	T(170)S	T(176)I	T(190)I	T(221)I	T(229)I	T(34)A
Frequency of mutations	1	1	1	1	1	1	1	1	1	1	1	1
Mutations	T(64)I	T(9)K	V(112)F	V(13)A	V(197)I	V(201)I	V(237)F	V(255)L	V(256)I	V(259)E	V(48)F	V(50)A
Frequency of mutations	1	1	1	1	1	1	1	1	1	1	1	1
Mutations	V(50)I	V(55)G	V(77)F	V(88)L	V(90)F	V(97)A	W (131)L	W(131)S	W(193)C	W(193)R	W(45)L	W(69)C
Frequency of mutations	1	1	1	1	1	1	1	1	1	1	1	1
Mutations	W(69)L	W(69)R	Y(154)C	Y(211)C	Y(215)H	Y(264)C						
Frequency of mutations	1	1	1	1	1	1						

[a] 124 is the frequency of the mutation Q to H occurred at the 57th position.

Although this is a neutral mutation but the change in polarity may bear some significance in structural properties.

- In QJU70306.1, there is a mutation at G224C which was found to be a diseased type polarity changed from hydrophobic to hydrophilic. This mutation may change the structure and functions of the protein.
- The ORF3a protein QLI51614.1 contains a mutation V197L which was found to be a diseased type with no polarity change.

In this network flow of mutations, it was also found sequences possessing 3rd level mutations which are described below:

- QLA47776.1: this sequence contains three mutations (Q57H, V55S, A23S), 3rd mutation is the neutral type, and polarity changed from hydrophobic to hydrophilic. Such mutations altogether may affect both structure and function of the protein.

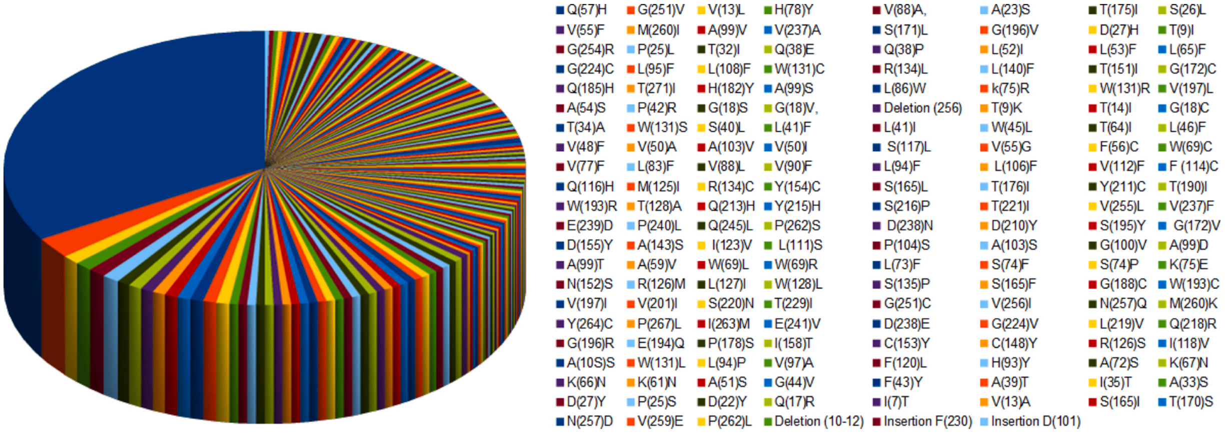


Fig. 3. Pie chart of the frequency of distinct mutations.

- QJW28665.1: this sequence contains three mutations (Q57H, G224C, L65F), 3rd mutation is the neutral type with no polarity change. The mutation L65F might not affect the virulence property of the SARS-CoV-2.

Flow of consecutive mutation-III: In this case, network flow (Fig. 9) of mutations is designed based on the ORF3a proteins of Indian origin. The sequence QLA10225.1 contains a mutation Q57H as usual. Further five ORF3a proteins are turned up in the network flow in the succeeding time scale of collection of samples. It was found that, all of them possess second mutation along with Q57H.

- The mutation R134L in the ORF3a protein QLF98201.1, which was found to be a disease type, and there was a polarity change from hydrophilic to hydrophobic. Here the change in mutations may lead to changes in tetramerization properties of the protein.
- The protein QLF98084.1 possesses a mutation at A54S, which was found to be a disease type, and the polarity changed from hydrophobic to hydrophilic, and hence the structure of the protein is expected to be differed, and accordingly the functions of the ORF3a protein would be affected.
- QLH64816.1, there is a mutation at P42R which was found to be a disease type, and there was a change in polarity from hydrophobic to hydrophilic, and consequently the mutation may contribute to structural changes of the ORF3a protein.
- The protein QLI49698.1 contains the mutation T271I which was found to be a neutral type, and there was a change in polarity from hydrophilic to hydrophobic. Although the mutation is predicted to be neutral but the hydrophobicity is changed, and hence alternation of functions of the proteins is expected.
- In ORF3a protein QLA10165.1, there is a mutation G18V which was found to be a neutral type of mutation, and there is no change in polarity, and consequently functions of the proteins would remain the same.

Flow of consecutive mutation-IV: The sequence QLC46986.1 contains a mutation Q38P which is a disease mutation with the change in polarity from hydrophilic to hydrophobic which might cause a change in functions of the protein. The network flow of mutations is presented in Fig. 10.

A second level mutation along with Q38P occurred in QKG81932.1 sequence from W131S which is also a disease type mutation, and polarity changed from hydrophobic to hydrophilic, and so it may change the structure of the protein. Also the ORF3a protein QKV07184.1 possesses G254R which changed the polarity from hydrophobic to hydrophilic and caused disease type mutation. On further analysis, the QJC19648.1 sequence was identified to have G254R along with T9K which is a disease mutation with no change in polarity. This is a mutation at the C-terminal region of protein so this mutation may effect the protein-protein interaction.

There is another sequence QKU53050.1(from USA) present in the work-flow, which contains the usual mutation Q57H, and a France based ORF3a sequence QJT72471.1 possessing a Q57H mutation along with A99V mutation which is a disease type mutation with no change in polarity. QJT72507.1 is another sequence of France origin, in which there is a mutation at Y154C along with Q57H mutation. Also in the QKG87159.1 sequence, another mutation apart from Q57H, and A99V at position V237A which is a disease type with no change in polarity.

Another possible traffic of mutation was observed in which an Australian sequence QLG75678.1 had a mutation at 78th position from H to Y, a neutral mutation with no change in polarity which may be a virulence promoting factor. Another Australian sequence QJR88822.1 was identified in which H78Y mutation was observed with V13L which is a disease mutation with no change in polarity. So here we observed that along with a neutral mutation a disease mutation has occurred, and it can be assumed that the virus first evolved in terms of virulence then enhanced its functional activity. Although there is no change in polarity, but it may affect the chemical properties. The sequence QLF98036.1 was another sequence from Bangladesh found to have H78Y mutation in addition to Q38E which is a disease mutation with no change in polarity. here also a disease mutation is also observed along with neutral mutation again signifying the evolutionary importance of these mutations.

Flow of consecutive mutation-V: The network flow of mutations (Fig. 11) with reference sequence of Wuhan's (ID YP_009724391.1) is formed.

The ORF3a protein QJR89362.1 possess a mutation G251V. It was found to be a disease type mutation, and here no change in polarity is observed but it may have some significance as it is a disease causing mutation. From this originates another sequence in the flow whose sample collection date is ensuing to the previous one. This sequence (ID QKV38209.1) bears a mutation in W69L which is a disease mutation without any change of polarity that is both W and L are neutral. As this sequence has both the disease mutations, it indicates their functional importance.

In the second case, when the sequence (ID QJS54023.1) of geo-location Greece, is compared with the Wuhan sequence it bore the

Table 4
protein IDs and respective mutations, geo-locations, total number of mutations in the protein, domains, and predicted effect of the mutations.

Protein ID	Country	Mutations	Total mutations	Domain	Effect of mutation(s)	RI
QJD47419.1	USA	T(9)I	1	FD I	Disease (0.649)	3
QLH01250.1	USA	V(13)L	1	FD I	Neutral (0.119)	8
QLB39261.1	USA	T(14)I	1	FD I	Disease (0.650)	3
QJW69308.1	GERMANY	P(25)L	1		Neutral (0.125)	8
QKV38281.1	AUSTRALIA	S(26)L	1		Neutral (0.157)	7
QKS67456.1	CHINA	T(32)I	1		Disease (0.652)	3
QJS39568.1	Netherlands	T(34)A	1		Neutral (0.297)	4
QLH93429.1	Bangladesh	Q(38)E	1	FD II (TRAF3 binding domain)	Disease (0.631)	3
QLC46986.1	USA	Q(38)P	1	FD II (TRAF3 binding domain)	Disease (0.638)	3
QKE61733.1	India	L(41)F	1	FD II	Neutral (0.114)	8
QKV41616.1	USA	L(41)I	1	FD II	Neutral (0.266)	5
QJR88306.1	Australia	L(46)F	1	TransmembraneDomain I (FD II)	Neutral (0.114)	8
QLF97772.1	Bangladesh	V(48)F	1	TransmembraneDomain I (FD II)	Disease (0.717)	4
QLF95641.1	USA	Q(57)H	1	TransmembraneDomain I	Disease (0.637)	3
QJD23478.1	USA	V(50)A	1	TransmembraneDomain I	Disease (0.599)	2
QJR89110.1	AUSTRALIA	L(52)I	1	TransmembraneDomain I	Neutral (0.454)	1
QKG64052.1	USA	F(56)C,	1	TransmembraneDomain I	Disease (0.673)	3
QLH58601.1	USA	Q(57)H,	1	TransmembraneDomain I	Disease (0.637)	3
QKU53050.1	USA	Q(57)H	1	TransmembraneDomain I	[a]Disease (0.637)	3
QLA10225.1	India	Q(57)H,	1	TransmembraneDomain I	Disease (0.637)	3
QJC20380.1	USA	Q(57)H	1	TransmembraneDomain I	Disease (0.637)	3
QKO25747.1	Bangladesh	W(69)L	1		Disease (0.625)	3
QKX47995.1	Bangladesh	W(69)R	1		Disease (0.650)	3
QJT72387.1	France	L(73)F	1		Disease (0.623)	2
QLG75930.1	Australia	S(74)F	1		Neutral (0.478)	0
QKV38257.1	Australia	S(74)P	1		Disease (0.657)	3
QQKX49024.1	Bangladesh	K(75)E	1		Disease (0.649)	3
QKU37034.1	Saudi Arabia	V(88)A	1	FD III	Disease (0.636)	3
QKQ63773.1	USA	L(106)F	1	FD III	Disease (0.631)	3
QKU32202.1	USA	L(106)F	1	FD III	Disease (0.631)	3
QKV40716.1	USA	R(126)M	1	FD III	Disease (0.696)	4
QIZ16548.1	Greece	L(127)I	1	FD III (cysteine rich domain)	Neutral(0.447)	1
QKE45861.1	USA	W(128)L	1	FD III (cysteine rich domain)	Disease (0.675)	4
QJD47873.1	Taiwan	W(131)C	1	FD III (cysteine rich domain)	Disease(0.666)	3
QKV35688.1	USA	W(131)R	1	FD III (cysteine rich domain)	Disease (0.717)	4
QLC93357.1	USA	R(134)L	1	FD III	Disease(0.712)	4
QII57239.2	USA	S(135)P	1	FD III	Disease(0.688)	3
QKU53854.1	USA	L(140)F	1	FD III	Disease(0.595)	2
QLF98261.1	India	T(151)I	1	FD III	Disease(0.624)	2
QKV07340.1	USA	S(165)F	1		Disease (0.614)	2
QLF80217.1	Brazil	S(171)L	1	FD VI (SGD motif)	Disease (0.602)	2
QLI50570.1	USA	G(172)C	1	FD VI (SGD motif)	Disease(0.646)	3
QLH59007.1	USA	T(175)I	1		Disease(0.728)	5
QLH55816.1	Bangladesh	G(188)C	1		Disease (0.668)	3
QKE10935.1	Czech Republic	W(193)C	1		Disease (0.600)	2
QLC92601.1	USA	V(197)I	1		Neutral (0.330)	3
QKK14612.1	USA	V(197)L	1		Disease (0.509)	0
QKU28463.1	USA	V(201)I	1		Neutral(0.255)	6
QLF97844.1	Bangladesh	S(220)N	1		Neutral (0.422)	1
QKX46204.1	USA	T(229)I	1		Disease(0.648)	3
QLH01382.1	USA	V(237)A	1		DiseasE(0.583)	2
QJY78272.1	USA	P(240)L	1		Disease(0.583)	2
QKU52834.1	USA	G(251)C	1		Disease(0.713)	4
QKU31182.1	USA	M(260)K	1		Disease(0.632)	3
QJX70592.1	USA	Y(264)C	1		Disease(0.651)	3
QLC92421.1	USA	P(267)L	1		Disease(0.525)	1
QKC05357.1	USA	T(271)I	1		Neutral (0.255)	5
QJD20838.1	Shri Lanka	I(263)M	1		Disease(0.510)	0
QKV07184.1	USA	G(254)R	1		Disease(0.728)	5
QLH57846.1	USA	G(251)V	1		Disease (0.770)	5
QJR89362.1	Australia	G(251)V	1		Disease (0.770)	5
QJS54191.1	Greece	E(241)V	1		Neutral(0.061)	9
QKV06236.1	USA	D(238)E	1		Neutral(0.244)	5
QJD47956.1	USA	G(224)V	1		Disease (0.686)	4
QJS39520.1	Netherlands	L(219)V	1		Neutral(0.137)	7
QKV42204.1	USA	Q(218)R	1		Disease(0.584)	2
QKG90867.1	USA	G(196)R	1		Disease (0.664)	3
QLG76026.1	Australia	G(196)V	1		Disease (0.687)	4
QLH56279.1	Bangladesh	E(194)Q	1		Neutral(0.140)	7
QJS39616.1	Netherlands	H(182)Y	1		Neutral(0.139)	7
QLG76386.1	Australia	P(178)S	1		Disease(0.565)	1
QLF97736.1	Bangladesh	G(172)V	1	FD VI (SGD motif)	Disease(0.646)	3

(*continued on next page*)

Table 4 (*continued*)

Protein ID	Country	Mutations	Total mutations	Domain	Effect of mutation(s)	RI
QLI51782.1	USA	I(158)T	1		Disease (0.734)	5
QLC92097.1	USA	D(155)Y	1	FD VI	Disease (0.829)	7
QIS61315.1	USA	C(153)Y	1	FD VI	Disease (0.692)	4
QJF77147.1	USA	C(148)Y	1	FD IV (caveolin binding domain)	Disease (0.785)	6
QJE38451.1	USA	R(126)S	1	FD III	Disease (0.671)	3
QKS67001.1	USA	I(118)V	1	FD III	Neutral (0.063)	9
QLF78310.1	Poland	A(99)S	1	FD III	Disease (0.577)	2
QKO25735.1	Bangladesh	A(99)V	1	FD III	Disease (0.602)	2
QLF95773.1	USA	H(93)Y	1	FD III	Disease (0.649)	3
QLG75678.1	Australia	H(78)Y	1	TD2	Neutral (0.349)	3
QIZ14498.1	USA	A(72)S	1		Disease (0.580)	2
QKK12852.1	Bangladesh	K(67)N	1		Disease (0.551)	1
QKY59990.1	India	K(66)N	1		Neutral (0.031)	9
QJD23730.1	USA	K(61)N	1	TD1	Disease (0.622)	2
QLF98048.1	Bangladesh	A(54)S	1	TDI	Disease (0.613)	2
QLC94305.1	USA	A(39)T	1	FD II	Disease (0.648)	3
QLF95737.1	USA	Q(57)H	1	TD I	Disease (0.637)	3
QJT72951.1	France	A(33)S	1	TD I	Disease (0.578)	2
QKG81824.1	USA	D(27)H	1		Neutral (0.139)	7
QKU53650.1	USA	D(27)Y	1		Neutral (0.220)	6
QLH93202.1	India	A(23)S	1		Neutral (0,494)	0
QLH55768.1	Bangladesh	D(22)Y	1		Neutral(0.187)	6
QLH55720.1	Bangladesh	G(18)V	1		Neutral (0.036)	9
QKW88844.1	USA	Q(17)R	1		Neutral (0.139)	7
QKV07400.1	USA	I(7)T	1	FD I	Neutral (0.213)	6
QKW89480.1	USA	V(13)A	1	FD I	Neutral (0.175)	7
QLH01334.1	USA	V(13)L	1	FD I	Neutral (0.119)	8
QLH00290.1	USA	S(26)L	1		Neutral (0.157)	7
QKS66305.1	USA	Q(57)H	1	TD I	Disease (0.637)	3
QKS65597.1	USA	Q(57)H	1	TD I	Disease (0.637)	3
QJS54923.1	USA	Q(57)H	1	TD I	Disease (0.637)	3
QJY78153.1	Egypt	Q(57)H	1	TD I	Disease (0.637)	3
QJQ39081.1	USA	Q(57)H	1	TD I	Disease (0.637)	3
QKV08048.1	USA	Q(57)H	1	TD I	Disease (0.637)	3
QKE45933.1	USA	Q(57)H	1	TD I	Disease (0.637)	3
QLI51746.1	USA	Q(57)H	1	TD I	Disease (0.637)	3
QLG99773.1	USA	Q(57)H	1	TD I	Disease (0.637)	3
QLH00362.1	USA	Q(57)H	1	TD I	Disease (0.637)	3
QKU37646.1	USA	Q(57)H	1	TD I	Disease (0.637)	3
QKS65849.1	USA	Q(57)H	1	TD I	Disease (0.637)	3
QJC20500.1	USA	Q(57)H	1	TD I	Disease (0.637)	3
QKU32046.1	USA	Q(57)H	1	TD I	Disease (0.637)	3
QJU11458.1	USA	Q(57)H	1	TD I	Disease (0.637)	3
QJR87730.1	Australia	Q(57)H	1	TD I	Disease (0.637)	3
QKU31638.1	USA	Q(57)H	1	TD I	Disease (0.637)	3
QKU31746.1	USA	Q(57)H	1	TD I	Disease (0.637)	3
QJR89278.1	Australia	Q(57)H	1	TD I	Disease (0.637)	3
QJR89446.1	Australia	Q(57)H	1	TD I	Disease (0.637)	3
QLH01238.1	USA	Q(57)H	1	TD I	Disease (0.637)	3
QJQ39297.1	USA	Q(57)H	1	TD I	Disease (0.637)	3
QLB39321.1	USA	Q(57)H	1	TD I	Disease (0.637)	3
QLH01298.1	USA	V(237)A	1		DiseasE(0.583)	2
QLG99737.1	USA	V(259)E	1		DiseasE(0.595)	2
QKV38401.1	Australia	G(196)V	1		Disease (0.687)	4
QIU78768.1	Spain	G(196)V	1		Disease (0.687)	4
QKS89844.1	USA	P(262)L	1		Disease (0.601)	2
QJD47539.1	USA	K(75)R	1		Disease (0.595)	2
QIZ14498.1	USA	A(72)S	1		Disease (0.580)	2
QJS54023.1	Greece	G(251)V	1		Disease (0.770)	5
QKV35400.1	USA	W(131)R	1	FD III (cysteine rich domain)	Disease (0.717)	4
QKU37202.1	USA	Q(57)H	1	TD I	Disease (0.637)	3
QIS30116.1	USA	Q(57)H	1	TD I	Disease (0.637)	3
QJR91354.1	Australia	Q(57)H	1	TD I	Disease (0.637)	3
QKN20740.1	USA	Q(57)H	1	TD I	Disease (0.637)	3
QIU81286.1	USA	F(8)L, deletion mutation (10-12)	1	FD I	Disease (0.642)	3
QLH55840.1	Bangladesh	Deletion (256)	1			
QJR84790.1	USA	Insertion F(230)	1			
QLG97055.1	Italy	Insertion D(101)	1	FD III		
QLI50282.1	USA	G(18)S, Q(57)H	2	TD I	Neutral(0.055),Disease(0.637)	9,3
QKS66053.1	USA	T(9)I, Q(57)H	2	FD1, TD I	Disease(0.649), Disease(0.637)	3,3
QJC19648.1	USA	T(9)K, G(254)R	2	FD1	Disease(0.747),Disease(0.728)	5,5
QJR88390.1	AUSTRALIA	V(13)L T(175)I	2	FD I	Neutral (0.119), Neutral (0.142)	8, 7
QJR88822.1	AUSTRALIA	V(13)L, H(78)Y	2	FD I, TD II	Neutral (0.119), Neutral (0.349)	8,3
QLI46290.1	USA	Q(57)H, G(18)C	2	TD I,	Disease (0.637),Neutral(0.134)	3,7

(continued on next page)

Table 4 (*continued*)

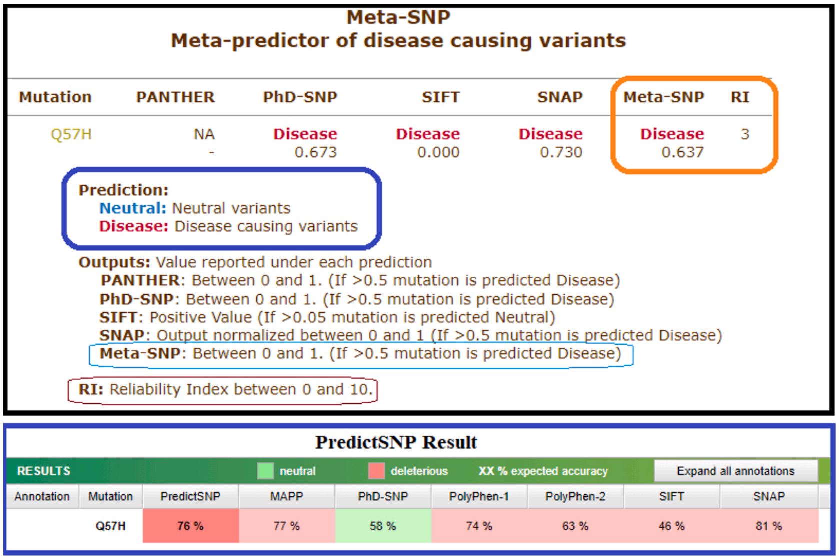
Protein ID	Country	Mutations	Total mutations	Domain	Effect of mutation(s)	RI
QKV40164.1	USA	Q(57)H, P(25)L	2	TD I	Disease(0.637), Neutral(0.125)	3,8
QLG97460.1	USA	Q(57)H, S(26)L	2	TD I	Disease (0.637),Neutral(0.157)	3,7
QJV21807.1	USA	Q(57)H, T(32)I	2	TD I	Disease (0.637),Disease(0.652)	3,3
QLF98036.1	Bangladesh	Q(38)E, H(78)Y	2	FD II (TRAF 3 binding domain), TDII	Disease(0.631),Neutral(0.349)	3,3
QKG81932.1	USA	Q(38)P, W(131)S	2	FD II (TRAF 3 binding domain), FDIII	Disease(0.638),Disease(0.674)	3,3
QLI50414.1	USA	Q(57)H, S(40)L	2	TD II, FDII	Disease(0.637),Disease (0.628)	3,3
QLH93441.1	Bangladesh	W(45)L, T(64)I	2	FDII, TD I	Disease(0.664),Neutral(0.166)	3,7
QLF97952.1	India	V(50)I, A(103)V	2	TDI, FDIII	Disease(0.588),Neutral (0.139)	2,7
QKE45885.1	USA	Q(57)H, L(52)I	2	TDI, TDI	Disease (0.637),Neutral(0.454)	3,1
QKS65621.1	USA	Q(57)H, L(53)F	2	TDI, TDI	Disease (0.637),Disease(0.601)	3,2
QKU29039.1	USA	Q(57)H, V(55)F	2	TDI, TDI	Disease (0.637),Disease(0.702)	3,4
QKS66941.1	Egypt	V(55)F, S(117)L	2	TDI, FD III	Disease(0.702),Disease(0.623)	4,2
QLG76542.1	AUSTRALIA	Q(57)H, V(55)G	2	TDI, TDI	Disease(0.637),Disease(0.649)	3,3
QLA47500.1	USA	Q(57)H, L(65)F	2	TDI	Disease (0.637),Neutral(0.233)	3,5
QKN20812.1	USA	Q(57)H, W(69)C	2	TDI	Disease(0.637), Disease (0.642)	3,3
QJX44383.1	India	Q(57)H, V(77)F	2	TDI, TD II	Disease (0.637), Neutral (0.079)	3, 8
QLF95245.1	USA	Q(57)H, L(83)F	2	TDI, FD III	Disease(0.637), Disease(0.636)	3,3
QLI50222.1	USA	Q(57)H, V(88)L	2	TDI, FD III	Disease(0.637), Disease90.665)	3,3
QLC94737.1	USA	Q(57)H, V(90)F	2	TDI, FD III	Disease(0.637),Disease(0.615)	3,2
QKG87087.1	USA	Q(57)H, L(94)F	2	TDI, FD III	Disease(0.637),Neutral(0.146)	3,7
QIZ13838.1	USA	Q(57)H, L(95)F	2	TDI, FD III	Disease(0.637),Disease(0.601)	3,2
QJQ84173.1	USA	Q(57)H, L(106)F	2	TDI, FD III	Disease (0.637),Disease(0.631)	3,3
QKG88539.1	USA	Q(57)H, L(108)F	2	TDI, FD III	Disease (0.637),Neutral(0.367)	3,3
QJY40110.1	USA	Q(57)H, V(112)F	2	TDI, FD III	Disease (0.637),Disease(0.621)	3,2
QJD47551.1	USA	Q(57)H, F (114)C	2	TDI, FD III	Disease (0.637),Disease(0.624)	3,2
QJD25758.1	USA	Q(57)H, Q(116)H	2	TDI, FD III	Disease (0.637),Disease(0.714)	3,4
QJD47849.1	Taiwan	Q(57)H, M(125)I	2	TDI, FD III	Disease(0.637),Disease(0.680)	3,4
QKU30570.1	USA	Q(57)H, W(131)C	2	TDI, FD III (cysteine rich domain)	Disease (0.637),Disease(0.666)	3,3
QKG90399.1	USA	Q(57)H, R(134)C	2	TDI, FD III	Disease(0.637),Disease(0.717)	3,4
QLF98201.1	India	Q(57)H, R(134)L	2	TDI, FD III	Disease (0.637),Disease(0.712)	3,4
QJR95110.1	AUSTRALIA	Q(57)H, L(140)F	2	TDI, FD III	Disease (0.637),Disease(0.595)	3,2
QIZ13336.1	USA	Q(57)H, T(151)I	2	TDI, FD III	Disease(0.637),Disease(0.624)	3,2
QJT72507.1	France	Q(57)H, Y(154)C	2	TDI, FD III	Disease(0.637),Disease(0.752)	3,5
QKV06224.1	USA	Q(57)H, S(165)L	2	TDI, TDI	Disease (0.637),Disease(0.592)	3,2
QLH58947.1	USA	Q(57)H, G(172)C	2	TDI, FD VI (SGD motif)	Disease(0.637),Disease(0.646)	3,3
QJI07211.1	USA	Q(57)H, T(176)I	2	TDI	Disease (0.637),Neutral(0.184)	3,6
QLH58085.1	USA	Q(57)H, Q(185)H	2	TDI	Disease (0.637),Disease(0.636)	3,3
QKO00487.1	India	Q(57)H, T(190)I	2	TDI	Disease (0.637),Neutral(0.118)	3,7
QKV39588.1	USA	Q(57)H, W(193)R	2	TDI	Disease (0.637),Neutral(0.067)	3,9
QKV38810.1	USA	Q(57)H, T(128)A	2	TDI, FD III	Disease (0.637),Disease(0.641)	3,3
QLC47346.1	USA	Q(57)H, Q(213)H	2	TDI	Disease (0.637),Disease(0.641)	3,3
QKG91107.1	USA	Q(57)H, Y(215)H	2	TDI	Disease (0.637),Neutral(0.139)	3,7
QLH56255.1	Saudi Arabia	Q(57)H, S(216)P	2	TDI	Disease (0.637),Disease(0.661)	3,3
QJQ39045.1	USA	Q(57)H, T(221)I	2	TDI	Disease(0.637),Disease(0.656)	3,3
QJU70306.1	USA	Q(57)H, G(224)C	2	TDI	Disease(0.637), Disease(0.693)	3, 4
QLG75126.1	Baharain	Q(57)H, V(255)L	2	TDI	Disease (0.637),Disease(0.588)	3,2
QJT72327.1	France	Q(57)H, V(237)F	2	TDI	Disease (0.637),Disease(0.648)	3,3
QIZ16438.1	USA	Q(57)H, E(239)D	2	TDI	Disease (0.637),Neutral (0.051)	3,9
QLI51038.1	USA	Q(57)H, P(240)L	2	TDI	Disease(0.637),Disease(0.583)	3,2
QKG86518.1	USA	Q(57)H, Q(245)L	2	TDI	Disease (0.637),Disease(0.625)	3,3
QLG75942.1	Australia	Q(57)H, M(260)I	2	TDI	Disease(0.637), Disease (0.563)	3, 1
QKU28847.1	USA	Q(57)H, M(260)I	2	TDI	Disease (0.637),Disease (0.563)	3,1
QLI49698.1	India	Q(57)H, T(271)I	2	TDI	Disease(0.637), Neutral (0.255)	3, 5
QKV37633.1	Australia	Q(57)H, P(262)S	2	TDI	Disease (0.637),Disease (0.601)	3,2
QKG90495.1	USA	Q(57)H, D(238)N	2	TDI	Disease(0.637), Neutral(0.144)	3, 7
QLH58037.1	USA	Q(57)H, D(210)Y	2	TDI	Disease(0.637),Disease (0.610)	3,2
QJX68859.1	USA	Q(57)H, S(195)Y	2	TDI	Disease (0.637),Disease (0.653)	3,3
QKR84274.1	USA	Q(57)H, H(182)Y	2	TDI	Disease (0.637), Neutral(0.139)	3, 7
QKV38894.1	Egypt	Q(57)H, G(172)V	2	TDI, FD VI (SGD motif)	Disease (0.637),Disease (0.646)	3,3
QJS54155.1	Greece	Q(57)H, D(155)Y	2	TDI	Disease(0.637),Disease (0.829)	3,7
QJX44407.1	India	Q(57)H, A(143)S	2	TDI, FD IV (Caveolin binding motif)	Disease (0.637),Disease (0.604)	3,2
QKG87267.1	USA	Q(57)H, I(123)V	2	TDI, FD III	Disease (0.637),Neutral (0.139)	3,7
QJS57052.1	USA	Q(57)H, L(111)S	2	TDI, FD III	Disease (0.637),Disease (0.636)	3,3
QKG87195.1	USA	Q(57)H, P(104)S	2	TDI, FD III	Disease (0.637),Neutral (0.143)	3,7
QLH93453.1	Bangladesh	Q(57)H, A(103)S	2	TDI, FD III	Disease (0.637),Neutral (0.448)	3,1
QIS61075.1	USA	Q(57)H, G(100)V	2	TDI, FD III	Disease (0.637),Disease (0.711)	3,7
QJW28449.1	USA	Q(57)H, A(99)D	2	TDI, FD III	Disease (0.637),Disease (0.723)	3,4
QLH56231.1	Saudi Arabia	Q(57)H, A(99)S	2	TDI, FD III	Disease (0.637), Disease (0.577)	3, 2
QLC91905.1	USA	Q(57)H, A(99)T	2	TDI, FD III	Disease (0.637)Disease (0.602)	3,2
QJT72471.1	France	Q(57)H, A(99)V	2	TDI, FD III	Disease (0.637), Disease (0.602)	3,2
QJW00412.1	India	Q(57)H, L(86)W	2	TDI, FD III	Disease (0.637), Disease(0.664)	3,3
QKG88935.1	USA	Q(57)H, L(86)W	2	TDI, FD III	Disease (0.637), Disease(0.664)	3,4

(*continued on next page*)

Table 4 (*continued*)

Protein ID	Country	Mutations	Total mutations	Domain	Effect of mutation(s)	RI
QLC91545.1	USA	Q(57)H, H(78)Y	2	TDI, TD II	Disease (0.637),Neutral (0.349)	3,3
QKV38005.1	Australia	Q(57)H, k(75)R	2	TDI, TD II	Disease (0.637),Disease (0.595)	3,2
QKN20824.1	USA	Q(57)H, A(59)V	2	TDI, TDI	Disease (0.637),Disease (0.622)	3,2
QKV38209.1	Australia	W(69)L, G(251)V	2		Disease (0.625),Disease (0.770)	3,5
QLA09656.1	USA	V(88)A, G(251)V	2	FD III	Disease (0.636),Disease (0.770)	3,5
QJD47203.1	USA	L(95)F, N(152)S	2	FD III	Disease(0.601),Neutral(0.189)	2,6
QHZ00380.1	South Korea	W(128)L, G(251)V	2	FD III	Disease (0.675),Disease (0.770)	4,5
QLA10069.1	India	V(256)I, N(257)Q	2		Disease (0.563),Disease (0.576)	1,2
QJS53735.1	Greece	G(251)V, M(260)I	2		Disease (0.770), Disease (0.563)	5, 1
QLG98012.1	USA	A(103)S, W(131)L	2	FD III, FDIII (Cysteine rich domain)	Neutral (0.448),Disease (0.661)	1,3
QLF98084.1	India	A(54)S, Q(57)H	2	TD I, TD I	Disease (0.613),Disease (0.637)	2,3
QLH56099.1	Saudi Arabia	A(51)S, Q(57)H	2	TD I, TD I	Disease (0.600),Disease (0.637)	2,3
QKV39324.1	USA	G(44)V, Q(57)H	2	FD II, TD I	Disease (0.628),Disease (0.637)	3,3
QKU32982.1	USA	F(43)Y, Q(57)H	2	FD II, TD I	Disease (0.625),Disease (0.637)	3,3
QLH64816.1	India	P(42)R, Q(57)H	2	FD II, TD I	Disease(0.615),Disease (0.637)	2,3
QJA17681.1	USA	P(42)R, Q(57)H	2	FD II, TD I	Disease(0.615),Disease (0.637)	2,3
QJY40506.1	India	I(35)T, L(53)F	2	TD I	Disease(0.628),Disease(0.601)	3,2
QLH57751.1	USA	D(27)H, Q(57)H	2	TD I	Neutral (0.139),Disease (0.637)	7,3
QLC46314.1	USA	D(27)H, Q(57)H	2	TD I	Neutral (0.139),Disease (0.637)	7,3
QKN19672.1	USA	P(25)S, T(175)I	2		Neutral(0.162),Disease(0.728)	7,5
QLG75822.1	Australia	A(23)S, Q(57)H	2	TD I	Neutral (0,494),Disease (0.637)	0, 3
QLG97484.1	USA	D(22)Y, Q(57)H	2	TD I	Neutral(0.187),Disease (0.637)	6,3
QLI50282.1	USA	G(18)S, Q(57)H	2	TD I	Neutral(0.055),Disease (0.637)	9,3
QLA10165.1	India	G(18)V, Q(57)H	2	TD I	Neutral (0.036),Disease (0.637)	9,3
QLI51614.1	USA	Q(57)H, V(197)L	2	TD I	Disease (0.637),Disease (0.509)	3, 0
QJD47299.1	USA	Q(57)H, S(165)I	2	TD I	Disease (0.637),Disease (0.605)	3,2
QLF99991.1	USA	Q(57)H, T(170)S	2	TD I	Disease (0.637),Neutral(0.174)	3,7
QJX70192.1	USA	Q(57)H, S(195)Y	2	TD I	Disease (0.637),Disease (0.653)	3,3
QLE11150.1	Bangladesh	N(257)D, deletion(256)	2		Disease (0.590)	2
QJW28665.1	USA	Q(57)H, L(65)F, G(224)C	3	TD I	Disease (0.637),Neutral(0.233),Disease (0.693)	3, 5, 4
QKV26659.1	USA	Q(57)H, Q(185)H, Y(211) C	3	TD I	Disease (0.637), Disease(0.636),Disease (0.733)	3, 3,5
QKG87159.1	USA	Q(57)H, A(99)V, V(237)A	3	TD I, FD III	Disease (0.637),Disease (0.602),Disease (0.583)	3, 2, 2
QKV42875.1	USA	V(88)A, S(171)L, G(251)V	3	FD III, FD VI (SGD motif)	Disease (0.636), Disease (0.602), Disease (0.770)	3, 2, 5
QKE44990.1	USA	L(94)P, V(97)A, F(120)L	3	FD III, FD III, FDIII	Disease(0.691),Neutral(0.157),Disease(0.641)	4,7,3
QLA47776.1	USA	Q(57)H, V(55), A(23)S	3	TD I, TD I	Disease (0.637),Disease(0.702)(3), Neutral (0,494)	3,4, 0
QKV41592.1	USA	V(88)A, L(108)F, S(171)L, G(251)V	4	FD III, FDIII, FD VI (SGD motif)	Disease (0.636),Neutral(0.367),Disease (0.602),Disease (0.770)	3, 3,2, 5

[a] Disease(0.637) denotes the effect of the mutation Q(57)H as 'disease' with the degree 0.637.



Meta-SNP
Meta-predictor of disease causing variants

Mutation	PANTHER	PhD-SNP	SIFT	SNAP	Meta-SNP	RI
Q57H	NA -	Disease 0.673	Disease 0.000	Disease 0.730	Disease 0.637	3

Prediction:
Neutral: Neutral variants
Disease: Disease causing variants

Outputs: Value reported under each prediction
PANTHER: Between 0 and 1. (If >0.5 mutation is predicted Disease)
PhD-SNP: Between 0 and 1. (If >0.5 mutation is predicted Disease)
SIFT: Positive Value (If >0.05 mutation is predicted Neutral)
SNAP: Output normalized between 0 and 1 (If >0.5 mutation is predicted Disease)
Meta-SNP: Between 0 and 1. (If >0.5 mutation is predicted Disease)

RI: Reliability Index between 0 and 10.

PredictSNP Result

RESULTS 🟩 neutral 🟥 deleterious XX % expected accuracy Expand all annotations

Annotation	Mutation	PredictSNP	MAPP	PhD-SNP	PolyPhen-1	PolyPhen-2	SIFT	SNAP
	Q57H	76 %	77 %	58 %	74 %	63 %	46 %	81 %

Fig. 4. A snapshot of the predicted effect of the frequently occurred mutation Q57H in ORF3a using Meta-SNP web-server.

Table 5
ORF3a proteins possessed disease, neutral type of predicted mutations.

Disease		Disease		Disease		Neutral	
Protein ID	Geo-location	Protein ID	Geo-location	Protein ID	Geo-location	Protein ID	Geo-location
QLG76542.1	Australia	QJD47849.1	Taiwan	QLC93357.1	USA	QJR88390.1	Australia
QJR95110.1	Australia	QJD47873.1	Taiwan	QII57239.2	USA	QJR88822.1	Australia
QLG75942.1	Australia	QKS66053.1	USA	QKU53854.1	USA	QKV38281.1	Australia
QKV37633.1	Australia	QJC19648.1	USA	QKV07340.1	USA	QJR88306.1	Australia
QKV38005.1	Australia	QJV21807.1	USA	QLI50570.1	USA	QJR89110.1	Australia
QKV38209.1	Australia	QKG81932.1	USA	QLH59007.1	USA	QLG75930.1	Australia
QKV38257.1	Australia	QLI50414.1	USA	QKK14612.1	USA	QLG75678.1	Australia
QJR89362.1	Australia	QKS65621.1	USA	QKX46204.1	USA	QLF97844.1	Bangladesh
QLG76026.1	Australia	QKU29039.1	USA	QLH01382.1	USA	QLH56279.1	Bangladesh
QLG76386.1	Australia	QKN20812.1	USA	QJY78272.1	USA	QLH55768.1	Bangladesh
QJR87730.1	Australia	QLF95245.1	USA	QKU52834.1	USA	QLH55720.1	Bangladesh
QJR89278.1	Australia	QLI50222.1	USA	QKU31182.1	USA	QJW69308.1	Germany
QJR89446.1	Australia	QLC94737.1	USA	QJX70592.1	USA	QIZ16548.1	Greece
QKV38401.1	Australia	QIZ13838.1	USA	QLC92421.1	USA	QJS54191.1	Greece
QJR91354.1	Australia	QJQ84173.1	USA	QKV07184.1	USA	QKE61733.1	India
QLG75126.1	Baharain	QJY40110.1	USA	QLH57846.1	USA	QKY59990.1	India
QLH93429.1	Bangladesh	QJD47551.1	USA	QJD47956.1	USA	QLH93202.1	India
QLF97772.1	Bangladesh	QJD25758.1	USA	QKV42204.1	USA	QJS39568.1	Netherlands
QKO25747.1	Bangladesh	QKU30570.1	USA	QKG90867.1	USA	QJS39520.1	Netherlands
QKX47995.1	Bangladesh	QKG90399.1	USA	QLI51782.1	USA	QJS39616.1	Netherlands
QQKX49024.1	Bangladesh	QIZ13336.1	USA	QLC92097.1	USA	QLH01250.1	USA
QLH55816.1	Bangladesh	QKV06224.1	USA	QIS61315.1	USA	QKV41616.1	USA
QLF97736.1	Bangladesh	QLH58947.1	USA	QJF77147.1	USA	QLC92601.1	USA
QKO25735.1	Bangladesh	QLH58085.1	USA	QJE38451.1	USA	QKU28463.1	USA
QKK12852.1	Bangladesh	QKV38810.1	USA	QLF95773.1	USA	QKC05357.1	USA
QLF98048.1	Bangladesh	QLC47346.1	USA	QIZ14498.1	USA	QKV06236.1	USA
QLF80217.1	Brazil	QJQ39045.1	USA	QJD23730.1	USA	QKS67001.1	USA
QKS67456.1	CHINA	QJU70306.1	USA	QLC94305.1	USA	QKG81824.1	USA
QKE10935.1	Czech Republic	QLI51038.1	USA	QLF95737.1	USA	QKU53650.1	USA
QKS66941.1	Egypt	QKG86518.1	USA	QKS66305.1	USA	QKW88844.1	USA
QKV38894.1	Egypt	QKU28847.1	USA	QKS65597.1	USA	QKV07400.1	USA
QJY78153.1	Egypt	QLH58037.1	USA	QJS54923.1	USA	QKW89480.1	USA
QJT72507.1	France	QJX68859.1	USA	QJQ39081.1	USA	QLH01334.1	USA
QJT72327.1	France	QJS57052.1	USA	QKV08048.1	USA	QLH00290.1	USA
QJT72471.1	France	QIS61075.1	USA	QKE45933.1	USA		
QJT72387.1	France	QJW28449.1	USA	QLI51746.1	USA		
QJT72951.1	France	QLC91905.1	USA	QLG99773.1	USA		
QJS54155.1	Greece	QKG88935.1	USA	QLH00362.1	USA		
QJS53735.1	Greece	QKN20824.1	USA	QKU37646.1	USA		
QJS54023.1	Greece	QLA09656.1	USA	QKS65849.1	USA		
QLF98201.1	India	QKV39324.1	USA	QJC20500.1	USA		
QJX44407.1	India	QKU32982.1	USA	QKU32046.1	USA		
QJW00412.1	India	QJA17681.1	USA	QJU11458.1	USA		
QLA10069.1	India	QJD47419.1	USA	QKU31638.1	USA		
QLF98084.1	India	QLB39261.1	USA	QKU31746.1	USA		
QLH64816.1	India	QLC46986.1	USA	QLH01238.1	USA		
QJY40506.1	India	QLF95641.1	USA	QJQ39297.1	USA		
QLA10225.1	India	QJD23478.1	USA	QLB39321.1	USA		
QLF98261.1	India	QKG64052.1	USA	QLH01298.1	USA		
QLF78310.1	Poland	QLH58601.1	USA	QLG99737.1	USA		
QLH56255.1	Saudi Arabia	QKU53050.1	USA	QKS89844.1	USA		
QLH56231.1	Saudi Arabia	QJC20380.1	USA	QJD47539.1	USA		
QLH56099.1	Saudi Arabia	QKQ63773.1	USA	QIZ14498.1	USA		
QKU37034.1	Saudi Arabia	QKU32202.1	USA	QKV35400.1	USA		
QJD20838.1	Shri Lanka	QKV40716.1	USA	QKU37202.1	USA		
QHZ00380.1	South Korea	QKE45861.1	USA	QIS30116.1	USA		
QIU78768.1	Spain	QKV35688.1	USA	QKN20740.1	USA		
				QIU81286.1	USA		
				QKV26659.1	USA		
				QKG87159.1	USA		
				QKV42875.1	USA		

same mutation G251V. From here it is further divided into bi-flow according to geo-locations, and all of them have the G251V mutation along with certain new:

1. The left one bears a sequence (ID QJS53735.1) of geo-location Greece which has a mutation M260I which is a disease type of mutation, and has no change in polarity. Here, both the mutations are in the cytosolic domain indicating that these mutations are somehow important for the virus.
2. The right one is for the geo-location USA, which starts with the sequence (ID QLA09656.1) which has a mutation V88A. It is a disease type mutation with no change in polarity. So, it may be advantageous for the virus in terms of functionality. Following there is another sequence (ID QKV42875.1) with respect to the time scale, bearing a mutation at S171L. This is a disease type mutation, and there is a change in polarity from hydrophilic to hydrophobic. Since the polarity is changing which indicates that there is some effect on ionic, and electrostatic interactions that may cause structural changes. Lastly, the sequence QKV41592.1 which bears a mutation

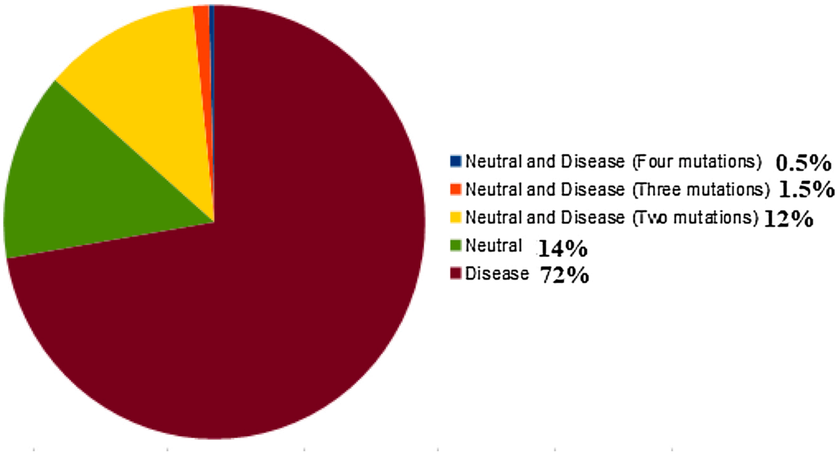


Fig. 5. Percentage of disease, neutral, and mixed (neutral & disease) type of mutations over the ORF3a proteins.

Table 6
Frequency and percentage of ORF3a proteins located at various countries, having type of mutations.

Disease				Neutral			
Geo-location	*Frequency*	*Total Frequency of Mutation*	*Percentage*	*Geo-location*	*Frequency*	*Total Frequency of Mutation*	*Percentage*
USA	116	156	74.36%	USA	14	156	8.97%
AUSTRALIA	15	22	68.19%	AUSTRALIA	7	22	31.82%
BANGLADESH	10	17	58.90%	BANGLADESH	4	17	23.53%
INDIA	9	16	56.25%	NETHERLANDS	3	3	100%
FRANCE	5	5	100%	INDIA	3	16	100%
SAUDI ARABIA	4	4	100%	GREECE	2	5	100%
EGYPT	3	3	100%	GERMANY	1	1	100%
GREECE	3	5	60%				
TAIWAN	2	2	100%				
BAHARAIN	1	1	100%				
SOUTH KOREA	1	1	100%				
CHINA	1	1	100%				
BRAZIL	1	1	100%				
CZECH REPUBLIC	1	1	100%				
SHRI LANKA	1	1	100%				
POLAND	1	1	100%				
SPAIN	1	1	100%				

at L108F which is a neutral mutation, and so has no change in polarity. This sequence has all disease mutations although no change in polarity is observed except for one mutation, so it signifies the order of occurrence of mutations allowing the virus to acquire new characteristics important for its survival.

In this study of mutation among many, we recognized five important mutations in the ORF3a proteins. While W131C, T151I, R134L, and D155Y forms a network of hydrophobic, polar, and electrostatic interactions which are important for the tetramerization process of ORF3a, F230 insertion is responsible for dimerization of ORF3a. We could see that all of the mutations have an effect of decrease in the stability apart from T151I which increases the stability of the protein. To get a better insight, we analyzed for these mutations from a structural point of view:

Case-I: We collected the available structure of ORF3a (Protein ID: 6XDC) from Protein Data Bank (PDB), (leftmost figure shown in colour grey) in Fig. 12

Then we took the mutated sequence which contains the mutation W131C, and performed homology modelling with the help of a web server called Swiss-model, and built the corresponding structure of W131C (middle picture shown in blue), and finally we superimposed the structure of Wuhan (reference structure) with that of the modelled (right most picture), and checked for the corresponding differences with respect to structural change; labelling the mutated portions with colour green (Q57H), and red (W131C).

Case-II: In this case, we consider the mutated sequence which possesses the mutation T151Y, and performed homology modelling, and built the corresponding structure of T151Y (middle picture shown in blue) as shown in Fig. 13.

Finally we overlaid the structure of Wuhan (reference structure) with that of the modelled (right most picture), and checked for the corresponding differences with respect to structural change; labelling the mutated portions with colour green (Q57H), and red (T151Y).

Case-III: With the available structure of ORF3a (Protein ID: 6XDC) from Protein Data Bank (PDB), (leftmost picture shown in colour grey) we took the mutated sequence of R134L, and performed homology modelling, and built the corresponding structure of R134L (middle picture shown in blue in Fig. 14)

Then we overlaid the structure of Wuhan (reference structure) with that of the modelled (right most picture), and checked for the corresponding differences with respect to structural change; labelling the mutated portions with colour green (Q57H), and red (R134L).

Case-IV: With the available structure of ORF3a (Protein ID: 6XDC)

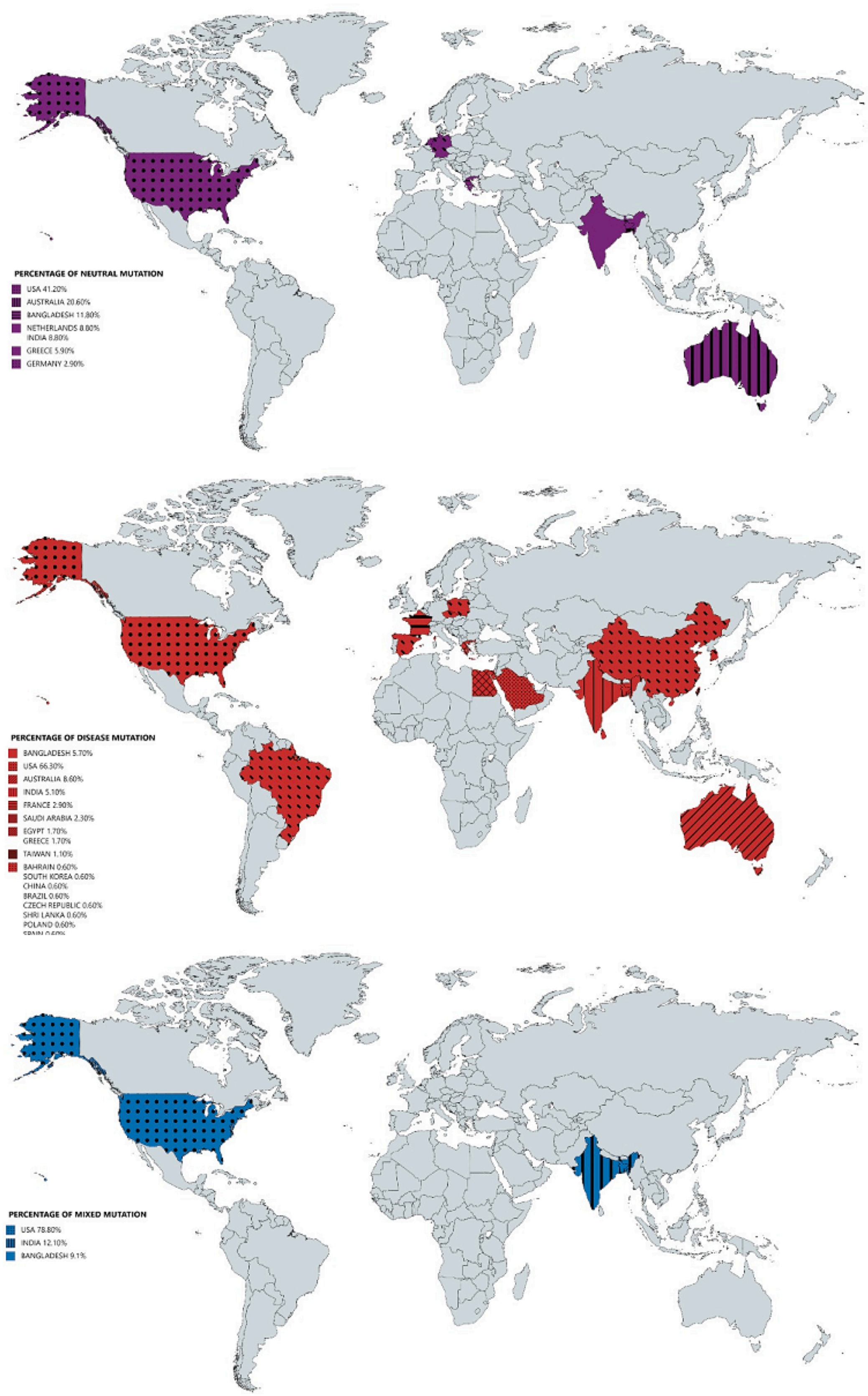


Fig. 6. World maps of percentage of occurrence of neutral, disease, and mixed type of mutations over the ORF3a proteins.

Table 7
ORF3a proteins with associated mutations, and predicted effect in stability of the structures.

Protein ID	Location	Mutation	Type of mutation	Effect on stability	[a]RI
QJR87730.1	Australia	Q(57)H	[a]P to P	Decrease	6
QKV38005.1	Australia	Q(57)H, K(75)R	P to P, P to P	Decrease, Increase	6, 3
QLG75822.1	Australia	Q(57)H, A(23)S	P to P, [a]NP to P	Decrease, Decrease	6, 8
QLG76542.1	Australia	Q(57)H, V(55)G	P to P, NP to NP	Decrease, Decrease	6,
QJR95110.1	Australia	Q(57)H, L (140)F	P to P, NP to NP	Decrease, Decrease	6, 9
QKU53050.1	USA	Q(57)H	P to P	Decrease	6
QKU30570.1	USA	Q(57)H, W(131)C	P to P, NP to P	Decrease, Decrease	6,7
QIZ13838.1	USA	Q(57)H, L (95)F	P to P, NP to NP	Decrease, Decrease	6, 7
QKU29039.1	USA	Q(57)H, V(55)F	P to P, NP to NP	Decrease, Decrease	6, 9
QKU28847.1	USA	Q(57)H, M(260)I	P to P, NP to NP	Decrease, Decrease	6, 6
QKG88539.1	USA	Q(57)H, L (108)F	P to P, NP to NP	Decrease, Decrease	6, 7
QLI50282.1	USA	Q(57)H, G(18)S	P to P, NP to P	Decrease, Decrease	6, 8
QJU70306.1	USA	Q(57)H, G(224)C	P to P, NP to P	Decrease, Decrease	6, 3
QLA47776.1	USA	Q(57)H, V(55)F, A (23)S	P to P, NP to NP,	Decrease, Decrease, Decrease	6, 9, 8
QLH58085.1	USA	Q(57)H, Q(185)H	P to P, P to P	Decrease, Decrease	6, 3
QJW28665.1	USA	Q(57)H, G(224)C, L(65)F	P to P, NP to P, NP to NP	Decrease, Decrease, Decrease	6, 8, 7
QLA10225.1	India	Q(57)H	P to P	Decrease	6
QLF98201.1	India	Q(57)H, R(134)L	P to P, P to NP	Decrease, Decrease	6, 9
QLF98084.1	India	Q(57)H, A(54)S	P to P, NP to P	Decrease, Decrease	6, 8
QLH64816.1	India	Q(57)H, P (42)R	P to P, NP to P	Decrease, Decrease	6, 9
QLI49698.1	India	Q(57)H, T (271)I	P to P, P to NP	Decrease, Increase	6, 3
QLA10165.1	India	Q(57)H, G(18)V	P to P, NP to NP	Decrease, Decrease	6, 4
QLC46986.1	USA	Q38P	P to NP	Decreases	6
QKG81932.1	USA	Q38P, W131S	P to NP, NP to P	Decreases, Decreases	6, 6
QKV07184.1	USA	G254R	NP to P	Decreases	7
QJC19648.1	USA	G254R, T9K	NP to P, P to P	Decreases, Decreases	7, 7
QKU53050.1	USA	Q57H	P to P	Decreases	6
QJT72471.1	FRANCE	Q57H, A99V	P to P, NP to NP	Decreases, Increases	6, 7
QKG87159.1	USA	Q57H, A99V, V237A	P to P, NP to NP, NP to NP	Decreases, Increases, Decreases	6, 7, 9
QJT72507.1	FRANCE	Q57H, Y154C	P to P, P to NP	Decreases, Decreases	6, 5
QLG75678.1	AUSTRALIA	H78Y	P to NP	Increases	6
QJR88822.1	AUSTRALIA	H78Y, V13L	P to NP, NP to NP	Increases, Increases	6, 0
QLF98036.1	BANGLADESH	H78Y, Q38E	P to NP, P to P	Increases, Increases	6, 1
QJR89362.1	AUSTRALIA	G251V	NP to NP	Decreases	4
QKV38209.1	AUSTRALIA	G251V, W69L	NP to NP, NP to NP	Decreases, Decreases	4, 5
QJS54023.1	GREECE	G251V	NP to NP	Decreases	4
QJS53735.1	GREECE	G251V, M260I	NP to NP, NP to NP	Decreases, Decreases	4, 6
QLA09656.1	USA	G251V, V88A	NP to NP, NP to NP	Decreases, Decreases	4, 9
QKV42875.1	USA	G251V, V88A, S171L	NP to NP, NP to NP, P to NP	Decreases, Decreases, Increases	4, 9, 1
QKV41592.1	USA	G251V, V88A, S171L, L108F	NP to NP, NP to NP, P to NP, NP to NP	Decreases, Decreases, Increases, Decreases	4, 9, 1, 7

[a] Here P and NP stands for Polar, Non-Polar, and RI: Reliability index.

(leftmost picture shown in colour grey), and then we took the mutated sequence ORF3a considering the mutation D155Y, and performed homology modelling, and obtained the corresponding structure of D155Y (middle picture shown in blue in Fig. 15).

We then overlayd the structure of Wuhan (reference structure) with that of the modelled (right most picture), and checked for the corresponding differences with respect to structural change; labelling the mutated portions with colour green (D155Y), and red (D155Y).

Case-V: Using the structure of the ORF3a (Protein ID: 6XDC) (leftmost picture shown in colour grey in Fig. 16) by homology modelling the structure of the ORF3a protein which contains the insertion mutation F230 (middle picture shown in blue), is constructed.

Then we overlaid the structure of ORF3a based in Wuhan (reference structure) with that of the modelled (right most picture), and checked for the corresponding differences with respect to structural change; labelling the mutated portions with colour green (difference in structure), and red (inserted amino acid).

In this study no significant change in protein structure was observed, we need a better soft-ware to find the difference between Wuhan sequence, and mutated sequences.

3.3. Phylogeny and clustering

We attempted to cluster to cluster each of the 296 ORF3a proteins into twenty disjoint clusters based on the probability distribution of amino acids using K-means clustering technique (Table 8). Note that, the number of clusters (twenty) is chosen optimally by heuristic method in such a manner that the clusters are separated from each other significantly. The frequency probability of each amino acid across all the 296 ORF3a proteins is available as a supplementary file-I. The three truncated ORF3a proteins (detected in Indian patients) are clustered in the cluster 11 as shown in Table 9.

The largest cluster 5 contains 53 ORF3a proteins of the USA patients including other 33 from various geo-locations as shown in Table 9. It is found that the ORF3a variants of the USA belong to each of the clusters except the cluster 11 which contains only three truncated proteins belong. This observation confirms the diversity of ORF3a isolates from the USA. It has been seen that the clusters 4, 6, 9, and 10 contain only the ORF3a proteins which are isolated from USA patients.

Based on the hierarchical clustering method, a single linkage dendrogram was obtained using the distance matrix of the clusters formed by the K-means clustering method over the 296 ORF3a proteins. This dendogram (Fig. 17) depicts the nearness of the clusters which are formed.

The most nearest pair of clusters are (2, 3), (4, 6), (1, 18), (5, 12), (9, 15), (7, 13), and (16, 17) as observed from the dendogram (Fig. 17).

3.4. Variability of ORF3a isolates

The variations among the ORF3a proteins based on the disorderly character of the amino acids over the proteins were determined using Shannon entropy (SE). For each sequence, SE is determined according to the formula stated in the method 2.2.3, and shown in Table 10.

The SE of all the ORF3a proteins is bounded by the global minima

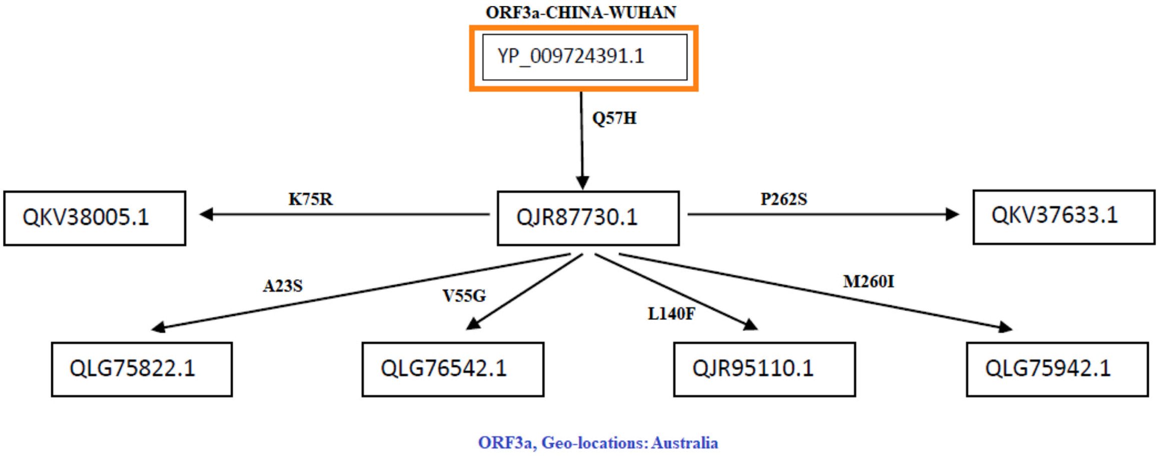


Fig. 7. Flow of mutations in Australian ORF3a proteins.

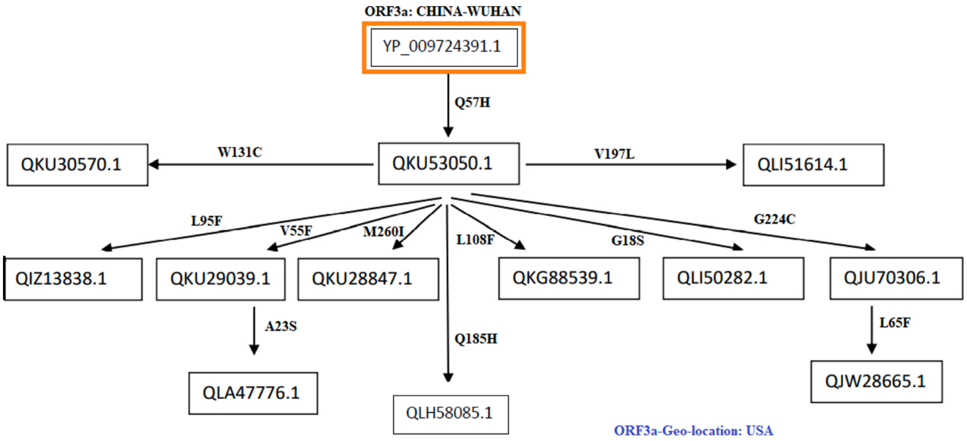


Fig. 8. Flow of mutations in ORF3a proteins from the USA.

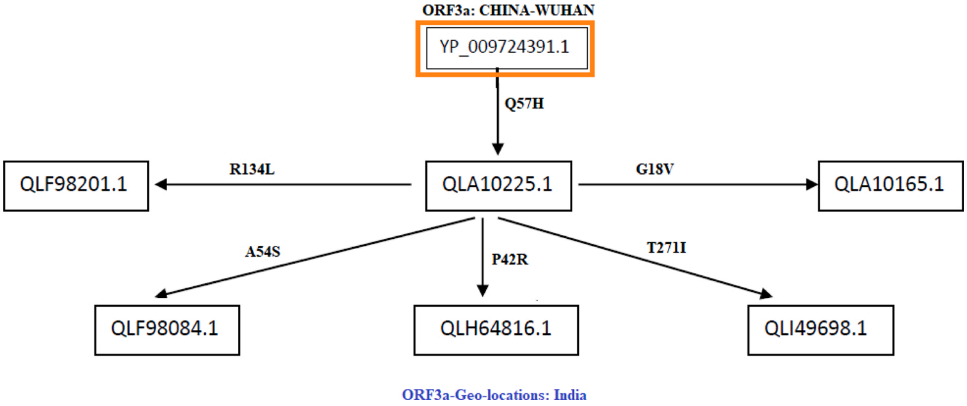


Fig. 9. Flow of mutations in ORF3a proteins of Indian origin.

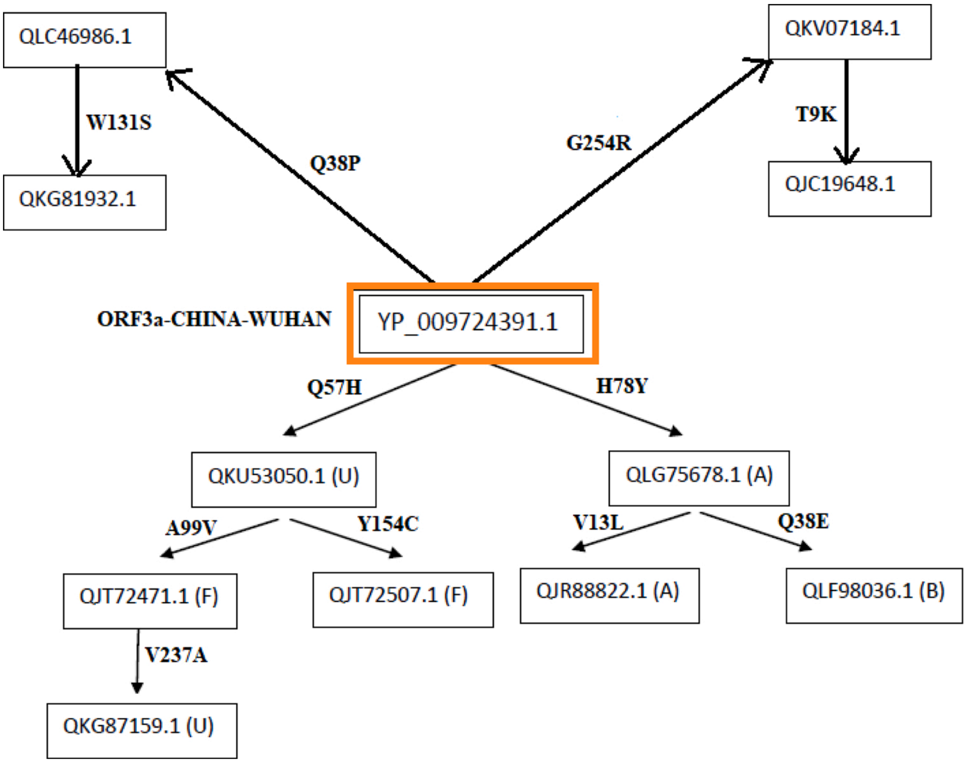


Fig. 10. Network flow of mutations of ORF3a proteins considering from various geo-locations.

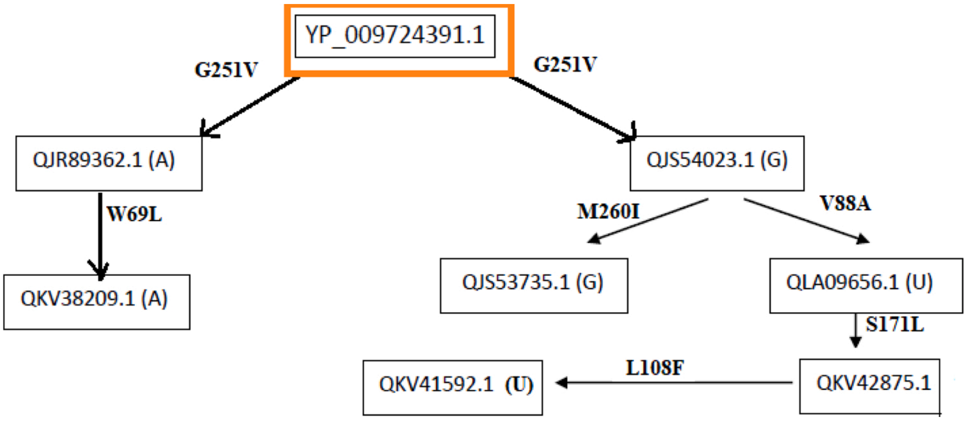


Fig. 11. Network flow of mutations of ORF3a proteins considering from various geo-locations. Note: A: Australia, B: Bangaldesh, F: France, G: Greece, and U: USA.

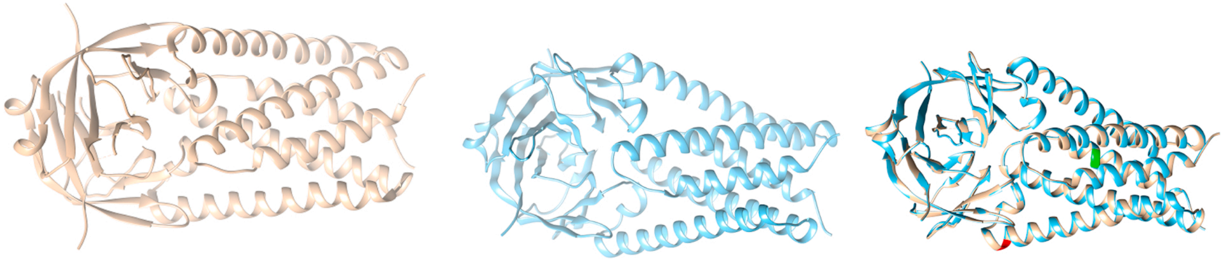


Fig. 12. Structures of ORF3a (Reference coloured as grey in left), Structure of mutated ORF3a (coloured with blue in the middle), and Overlaid ORF3a (right-most image).

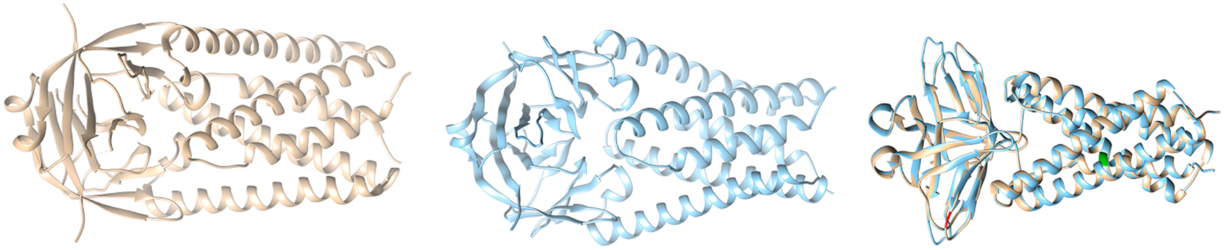


Fig. 13. Structures of ORF3a (Reference coloured as grey in left), Structure of mutated ORF3a (coloured with blue in the middle), and Overlaid ORF3a (rightmost image).

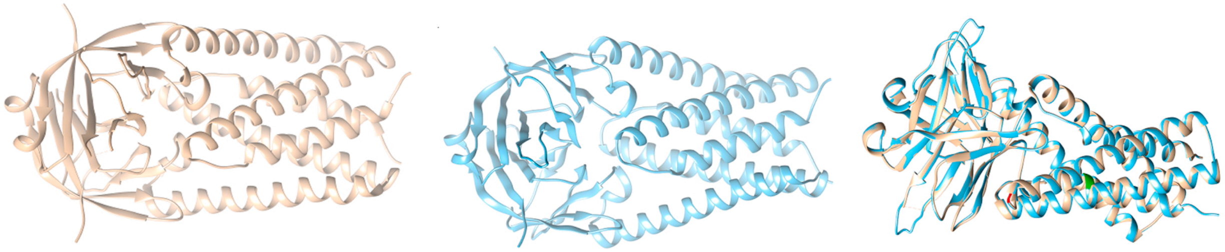


Fig. 14. Structures of ORF3a (Reference coloured as grey in left), Structure of mutated ORF3a (coloured with blue in the middle), and Overlaid ORF3a (rightmost image).

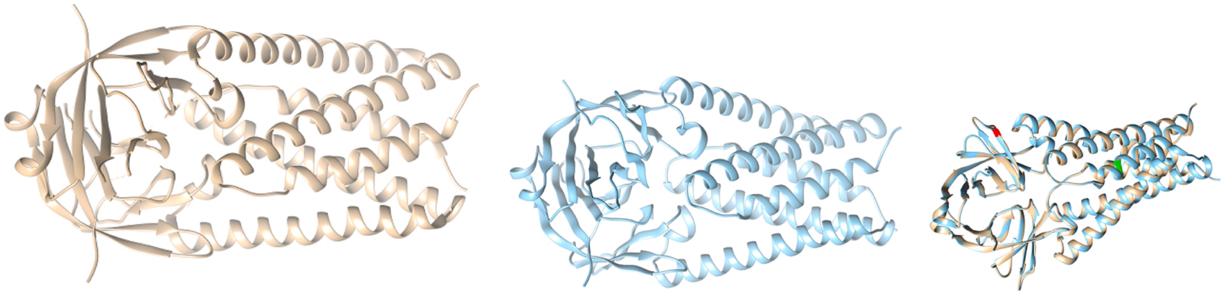


Fig. 15. Structures of ORF3a (Reference coloured as grey in left), Structure of mutated ORF3a (coloured with blue in the middle), and Overlaid ORF3a (rightmost image).

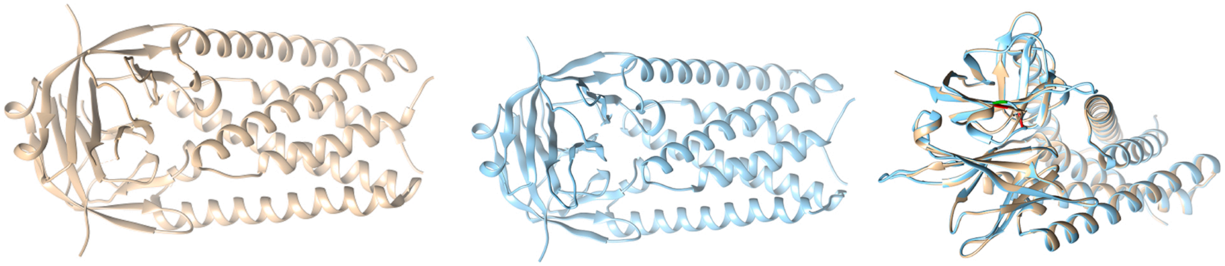


Fig. 16. Structures of ORF3a (Reference coloured as grey in left), Structure of mutated ORF3a (coloured with blue in the middle), and Overlaid ORF3a (rightmost image).

0.943, and global maxima 0.968 which are indeed the same as the minima and maxima of the ORF3a proteins which belong to the USA (Table 11). Clearly, the amount of disorderliness of the amino acids over the ORF3a proteins is extremely high.

The range of SE of the ORF3a proteins detected in the USA is comparatively more than others, and it ensures the wide variety of distinct ORF3a in USA patients. The SEs of 296 ORF3a proteins are plotted (Blue line) in Fig. 18. We found various non-smooth peaks, and those are clearly the SEs of the ORF3a proteins of the USA patients, and that is reconfirmed in the SE plot (Red line) of the ORF3a proteins of the USA.

4. Discussions

A total of 175 distinct mutations across the distinct 296 ORF3a proteins of SARS-CoV-2 are detected, and further analyzed. Among all the mutations, 32 mutations were already reported (Hassan et al., 2020; Issa et al., 2020). It was reported that in SARS-CoV, there exists an interchain disulphide bonds with that of the spike protein. SARS-CoV-2 ORF3a, contains a similar functional region (Domain III: C133) which is found to be conserved, as we did not find any mutation in this region. So, it can be assumed that this cysteine domain will perform a similar function in SARS-CoV-2 as in SARS-CoV, and is functionally important

Table 8
ORF3a proteins and corresponding cluster number based on amino acid distributions.

Protein ID	Cluster No	Protein ID	Cluster No	Protein ID	Cluster No	Protein ID	Cluster No
QLI46290.1 USA	1	QLH93441.1 Bangladesh	5	QKS66305.1 USA	6	QKX49024.1 Bangladesh	15
QKG81932.1 USA	1	QLF97772.1 Bangladesh	5	QKS65597.1 USA	6	QKV06236.1 USA	15
QKN20812.1 USA	1	QLF97952.1 India	5	QKG88935.1 USA	6	QLG99737.1 USA	15
QJD47551.1 USA	1	QKS66941.1 Egypt	5	QKS66041.1 USA	6	QJX45308.1 Poland	15
QJD25758.1 USA	1	QKG64052.1 USA	5	QJI54254.1 USA	6	QKS66053.1 USA	16
QKU30570.1 USA	1	QKO25747.1 Bangladesh	5	QKE61733.1 India	7	QJV21807.1 USA	16
QKG90399.1 USA	1	QKX47995.1 Bangladesh	5	QKV41616.1 USA	7	QLF95641.1 USA	16
QJT72507.1 France	1	QLG75930.1 Australia	5	QJR88306.1 Australia	7	QKE45885.1 USA	16
QLH58947.1 USA	1	QKV38257.1 Australia	5	QJR89110.1 Australia	7	QKS65621.1 USA	16
QKV26659.1 USA	1	QKV41592.1 USA	5	QJT72387.1 France	7	QKU29039.1 USA	16
QLH58085.1 USA	1	QKV42875.1 USA	5	QJD47203.1 USA	7	QLG76542.1 Australia	16
QLC47346.1 USA	1	QLA09656.1 USA	5	QKQ63773.1 USA	7	QJW28665.1 USA	16
QKG91107.1 USA	1	QLG97055.1 Italy	5	QKU32202.1 USA	7	QLA47500.1 USA	16
QJU70306.1 USA	1	QKV40716.1 USA	5	QIZ16548.1 Greece	7	QJX44383.1 India	16
QIZ16438.1 USA	1	QKE45861.1 USA	5	QKU53854.1 USA	7	QLF95245.1 USA	16
QLI49698.1 India	1	QJD47873.1 Taiwan	5	QKU31806.1 USA	7	QLC94737.1 USA	16
QJY78153.1 Egypt	1	QKV35688.1 USA	5	QJX45032.1 USA	7	QKG87087.1 USA	16
QKV08048.1 USA	1	QLC93357.1 USA	5	QJS39520.1 Netherlands	7	QIZ13838.1 USA	16
QKE45933.1 USA	1	QII57239.2 USA	5	QJR87598.1 Australia	7	QJQ84173.1 USA	16
QLG99773.1 USA	1	QLF98261.1 India	5	QJE38451.1 USA	7	QKG88539.1 USA	16
QKV38894.1 USA	1	QKV07340.1 USA	5	QJQ39741.1 USA	7	QJY40110.1 USA	16
QJX44407.1 India	1	QLF80217.1 Brazil	5	QLF78310.1 Poland	7	QJD47849.1 Taiwan	16
QJC20500.1 USA	1	QLI50570.1 USA	5	QLC93129.1 USA	7	QJR95110.1 Australia	16
QKG87267.1 USA	1	QLH59007.1 USA	5	QJR91282.1 Australia	7	QIZ13336.1 USA	16
QJS57052.1 USA	1	QLH55816.1 Bangladesh	5	QJD47539.1 USA	7	QKU53050.1 USA	16
QLH93453.1 Bangladesh	1	QKE10935.1 Czech Republic	5	QIZ14498.1 USA	7	QJI07211.1 USA	16
QJU11458.1 USA	1	QLC92601.1 USA	5	QJS53831.1 Greece	7	QKV38810.1 USA	16
QIS61075.1 USA	1	QKK14612.1 USA	5	QJS54023.1 Greece	7	QJQ39045.1 USA	16
QJR87730.1 Australia	1	QKU28463.1 USA	5	QLF98048.1 Bangladesh	7	QJT72327.1 France	16
QJW28449.1 USA	1	QLF97844.1 Bangladesh	5	QJD48484.1 USA	7	QJC20380.1 USA	16
QLH56231.1 Saudi Arabia	1	QKX46204.1 USA	5	QJY40506.1 India	7	QLG75942.1 Australia	16
QLC91905.1 USA	1	QJR84790.1 USA	5	QLH93202.1 India	7	QKU28847.1 USA	16
QKG87159.1 USA	1	QJY78272.1 USA	5	QKV39840.1 USA	7	QLI51746.1 USA	16
QJT72471.1 France	1	QKU52834.1 USA	5	QJR88822.1 Australia	8	QLH00362.1 USA	16
QKU31638.1 USA	1	QLE11150.1 Bangladesh	5	QLF98036.1 Bangladesh	8	QKU31746.1 USA	16
QLH01238.1 USA	1	QLH55840.1 Bangladesh	5	QJS39616.1 Netherlands	8	QJW00412.1 India	16
QKN20824.1 USA	1	QKU31182.1 USA	5	QIS61315.1 USA	8	QJR89278.1 Australia	16
QLF98084.1 India	1	QJX70592.1 USA	5	QJF77147.1 USA	8	QJR89446.1 Australia	16
QLH56099.1 Saudi Arabia	1	QLC92421.1 USA	5	QLF95773.1 USA	8	QLH57751.1 USA	16
QKU37202.1 USA	1	QKC05357.1 USA	5	QLG75678.1 Australia	8	QLA47776.1 USA	16
QKV39324.1 USA	1	YP_009724391.1 China	5	QJS54923.1 USA	9	QLG97460.1 USA	17
QIS30116.1 USA	1	QJD20838.1 Sri Lanka	5	QJD47299.1 USA	9	QLI50414.1 USA	17
QJT72951.1 France	1	QKV36900.1 USA	5	QJQ38625.1 USA	9	QLI50222.1 USA	17
QLC46314.1 USA	1	QKV07184.1 USA	5	QKG90147.1 USA	9	QLF98201.1 India	17
QLG75822.1 Australia	1	QKM76547.1 Germany	5	QKV42947.1 USA	10	QKV06224.1 USA	17
QLI50282.1 USA	1	QLH01502.1 USA	5	QKO00487.1 |truncated India	11	QLH58601.1 USA	17
QLA10165.1 India	1	QJF75396.1 USA	5	QLA10225.1 |truncated India	11	QLH56255.1 Saudi Arabia	17
QKG90495.1 USA	2	QKG90867.1 USA	5	QLA10069.1 |truncated India	11	QLG75126.1 Bahrain	17
QLH58037.1 USA	2	QKM76907.1 Germany	5	QKW89480.1 USA	12	QKG86518.1 USA	17
QKR84274.1 Egypt	2	QLH56279.1 Bangladesh	5	QJD23478.1 USA	12	QJX68859.1 USA	17
QJS54155.1 Greece	2	QKY77929.1 USA	5	QKU32046.1 USA	12	QKU37646.1 USA	17
QLC91545.1 USA	2	QLG76386.1 Australia	5	QKU37034.1 Saudi Arabia	12	QLI51614.1 USA	17
QLC92097.1 USA	2	QLG99677.1 USA	5	QLH01382.1 USA	12	QJQ39297.1 USA	17
QKU32982.1 USA	2	QKU31266.1 USA	5	QKV06920.1 USA	12	QLB39321.1 USA	17
QKG81824.1 USA	2	QLI51782.1 USA	5	QLH01298.1 USA	12	QKR84421.1 Egypt	17
QKU53650.1 USA	2	QLC91617.1 USA	5	QJI54123.1 USA	12	QJR91354.1 Australia	17
QLG97484.1 USA	2	QLG98012.1 USA	5	QKE44990.1 USA	12	QKS65849.1 USA	18
QLH55768.1 Bangladesh	2	QLC94473.1 USA	5	QKS89844.1 USA	12	QJS54383.1 Greece	18
QJX70192.1 USA	3	QLH00578.1 USA	5	QKV38209.1 Australia	13	QKV38401.1 Australia	18
QKS65777.1 USA	3	QKK12852.1 Bangladesh	5	QHZ00380.1 South Korea	13	QKU32934.1 USA	19
QKV35400.1 USA	3	QKE45765.1 USA	5	QJS53735.1 Greece	13	QKS66737.1 USA	19
QKV40440.1 USA	4	QLH00026.1 USA	5	QLH57846.1 USA	13	QKV40164.1 USA	20
QJD47419.1 USA	5	QKY59990.1 India	5	QJR89362.1 Australia	13	QKV39588.1 USA	20
QJC19648.1 USA	5	QJD23730.1 USA	5	QJS54191.1 Greece	13	QLI51038.1 USA	20
QJR88390.1 Australia	5	QKU52870.1 USA	5	QJD47956.1 USA	13	QKV37633.1 Australia	20
QLH01250.1 USA	5	QLC92553.1 USA	5	QLG76026.1 Australia	13	QJQ39081.1 USA	20
QLH01334.1 USA	5	QJR87574.1 Australia	5	QLF97736.1 Bangladesh	13	QKG87195.1 USA	20
QLB39261.1 USA	5	QKU31818.1 USA	5	QIU78768.1 Spain	13	QKV38005.1 Australia	20
QJW69308.1 Germany	5	QJR86050.1 Australia	5	QKS67001.1 USA	13	QKV42204.1 USA	20
QKV38281.1 Australia	5	QLC94305.1 USA	5	QKO25735.1 Bangladesh	13	QLH64816.1 India	20
QLH00290.1 USA	5	QKN19672.1 USA	5	QLH55720.1 Bangladesh	13	QJA17681.1 USA	20
QKS67456.1 China	5	QKS90192.1 USA	5	QLF99991.1 USA	14	QLF95737.1 USA	20
QJS39568.1 Netherlands	5	QIU81286.1 USA	5	QJR84550.1 USA	14	QKN20740.1 USA	20
QLC46986.1 USA	5	QKV07400.1 USA	5	QLH93429.1 Bangladesh	15	QKW88844.1 USA	20

Table 9
Clusters and its frequencies.

Cluster	Frequency	Cluster	Frequency
1	47	11	3
2	11	12	10
3	3	13	13
4	1	14	2
5	86	15	5
6	5	16	36
7	28	17	16
8	7	18	3
9	4	19	2
10	1	20	13

for virulence. In SARS-CoV, it was reported that tetramerization of the ORF3a protein is an important step for the ion channel formation which further increases the infectivity of the virus. From this study we found mutations W131C, T151I, R134L, and D155Y which may facilitate the tetramerization process in SARS-CoV-2, and thereby assisting the ion channel formation and favouring the virus with its infectivity. Similar to that of SARS-CoV, it is also responsible for apoptosis mediated by TRAF-3 (Domain III). We found two mutations in this region Q38E and Q38P which may enhance the effect of apoptosis but further studies are required. Caveolin-binding domain is responsible for viral uptake of the host cell, and its translocation to various endomembrane organelle. We have also detected mutations in this region (C148Y and A143S) which may enhance the viral uptake by the host, thereby increasing the infectivity rate. However, it is noteworthy that in *YXXϕ* motif domain, no mutation is observed so far, and consequently this domain is conserved. In seven ORF3a variants from the USA, two mutations are found in SGD domain (S171L & G172C), however the function of this SGD domain is unknown.

We characterized the mutations as three types: Neutral, Disease, and Mixed. Among these three types of mutations, we found that disease mutations are highly prevalent (66%) in the geo-loaction of the USA, indicating disease-causing character of the virus getting intensified, and thus posing a threat to mankind. Simultaneously, we have a mixed type of mutation occurring with a rate of 79% in the geo-location of the USA. Mixed type had both disease, and neutral types occurring together in the same protein. Although, neutral mutations are there in mixed type but frequency of disease mutation is high, again pointing towards the viral advantage over host. In France although the infectivity rate was very high, but disease (2.9%) mutation rate was low compared to the USA; where we find the maximum variety of mutation as shown with Shannon entropy in this study. So, we can suggest that the possible wide variety of mutations in the USA is due to the high rate of travel within the USA, and from outside USA, while in France there might be within-country transmission which resulted in less frequent mutations. We also checked the mortality rate of the USA (3.3%), France (13.4%), and India (2.1%), and from the results we found that France has the highest mortality rate than the USA followed by India. So, consequently we can draw a conclusion that France has only disease type mutation unlike that of the USA, and India in which all three types of mutations are present. This may prove that the presence of only disease types of mutation in a sequence may pose more danger to mankind than a sequence containing either mixed type or neutral type of mutations. Next, we analysed consecutive mutations within a protein sequence on the basis of chronological order of the time-line of sample collection from COVID-19 infected patients.

We further went on to analyse the mutations responsible for tetramerization, and dimerization with respect to structure and found that there were no significant structural changes observed by homology modelling method. So, other method should be used to detect the effect of mutations on the 3D structure of the protein and results need to be experimentally validated. Finally, twenty clusters are formed from 296 distinct variants of ORF3a of SARS-CoV-2 based on the amino acid compositions of the proteins. It also shows wide variety of compositions of ORF3a variants in the USA which is further quantitatively supported by the SE. This study of comprehensive 175 novel mutations would help in understanding the pathogenetic contribution of the ORF3a proteins. This understanding is an important aspect in devising vaccine for COVID-19.

Author contributions

SH conceived the problem. DA, SG, and SH examined the mutations. All the authors analysed the data and result. SH wrote the initial draft which was checked, and edited by all other authors to generate the final version.

Conflict of interests

The authors do not have any conflicts of interest to declare.

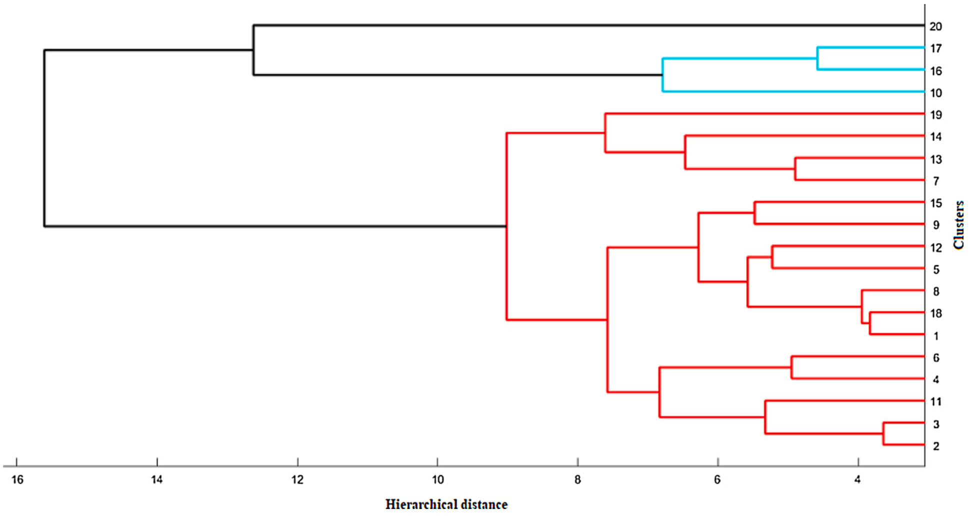


Fig. 17. Dendogram of the twenty clusters.

Table 10
Shannon entropy of amino acid conservations of the 296 ORF3a distinct variants across the world.

Protein ID	Geo-location	SE	Protein ID	Geo-location	SE	Protein ID	Geo-location	SE
QJR88390.1	*Australia*	0.957	QJD20838.1	*Sri Lanka*	0.959	QJD47203.1	*USA*	0.957
QJR88822.1	*Australia*	0.956	QJD47849.1	*Taiwan*	0.955	QKQ63773.1	*USA*	0.958
QKV38281.1	*Australia*	0.957	QJD47873.1	*Taiwan*	0.957	QKU32202.1	*USA*	0.958
QJR88306.1	*Australia*	0.958	QKS66053.1	*USA*	0.958	QKV40716.1	*USA*	0.958
QJR89110.1	*Australia*	0.958	QJD47419.1	*USA*	0.958	QKE45861.1	*USA*	0.955
QLG76542.1	*Australia*	0.958	QJC19648.1	*USA*	0.959	QKV35688.1	*USA*	0.957
QJR95110.1	*Australia*	0.958	QKW89480.1	*USA*	0.958	QLC93357.1	*USA*	0.955
QLG75942.1	*Australia*	0.955	QLH01250.1	*USA*	0.957	QII57239.2	*USA*	0.958
QKV37633.1	*Australia*	0.957	QLH01334.1	*USA*	0.957	QKU53854.1	*USA*	0.958
QJR87730.1	*Australia*	0.957	QLB39261.1	*USA*	0.958	QKU31806.1	*USA*	0.958
QJR89278.1	*Australia*	0.959	QLI46290.1	*USA*	0.958	QKS66041.1	*USA*	0.960
QJR89446.1	*Australia*	0.958	QKV40164.1	*USA*	0.956	QKV07340.1	*USA*	0.958
QKV38005.1	*Australia*	0.958	QLG97460.1	*USA*	0.957	QLI50570.1	*USA*	0.958
QKV38209.1	*Australia*	0.955	QLH00290.1	*USA*	0.957	QLH59007.1	*USA*	0.958
QLG75930.1	*Australia*	0.958	QJV21807.1	*USA*	0.958	QLC92601.1	*USA*	0.958
QKV38257.1	*Australia*	0.958	QKG81932.1	*USA*	0.955	QKK14612.1	*USA*	0.957
QJR89362.1	*Australia*	0.958	QLC46986.1	*USA*	0.957	QKU28463.1	*USA*	0.958
QLG76026.1	*Australia*	0.957	QLI50414.1	*USA*	0.957	QKX46204.1	*USA*	0.958
QLG76386.1	*Australia*	0.957	QKV41616.1	*USA*	0.958	QJR84790.1	*USA*	0.958
QKV38401.1	*Australia*	0.960	QLF95641.1	*USA*	0.958	QLH01382.1	*USA*	0.958
QJR87598.1	*Australia*	0.958	QJD23478.1	*USA*	0.958	QJY78272.1	*USA*	0.956
QLG75678.1	*Australia*	0.957	QKE45885.1	*USA*	0.958	QKU52834.1	*USA*	0.958
QJR91282.1	*Australia*	0.959	QKS65621.1	*USA*	0.958	QKU31182.1	*USA*	0.956
QJR87574.1	*Australia*	0.957	QKU29039.1	*USA*	0.958	QJX70592.1	*USA*	0.959
QJR86050.1	*Australia*	0.957	QKG64052.1	*USA*	0.958	QLC92421.1	*USA*	0.956
QJR91354.1	*Australia*	0.958	QJW28665.1	*USA*	0.959	QKC05357.1	*USA*	0.958
QLG75822.1	*Australia*	0.957	QLA47500.1	*USA*	0.958	QJX45032.1	*USA*	0.958
QLG75126.1	*Bahrain*	0.957	QKN20812.1	*USA*	0.957	QKV36900.1	*USA*	0.958
QLH93429.1	*Bangladesh*	0.957	QLF95245.1	*USA*	0.958	QKV07184.1	*USA*	0.958
QLF98036.1	*Bangladesh*	0.956	QLI50222.1	*USA*	0.957	QLH57846.1	*USA*	0.957
QLH93441.1	*Bangladesh*	0.956	QLC94737.1	*USA*	0.958	QLH01502.1	*USA*	0.957
QLF97772.1	*Bangladesh*	0.958	QKG87087.1	*USA*	0.958	QKV06236.1	*USA*	0.958
QLH93453.1	*Bangladesh*	0.957	QIZ13838.1	*USA*	0.958	QJF75396.1	*USA*	0.957
QKO25747.1	*Bangladesh*	0.955	QJQ84173.1	*USA*	0.958	QKV06920.1	*USA*	0.957
QKX47995.1	*Bangladesh*	0.957	QKG88539.1	*USA*	0.958	QJD47956.1	*USA*	0.957
QKX49024.1	*Bangladesh*	0.957	QJY40110.1	*USA*	0.958	QKV42204.1	*USA*	0.958
QLH55816.1	*Bangladesh*	0.958	QJD47551.1	*USA*	0.958	QJQ38625.1	*USA*	0.959
QLF97844.1	*Bangladesh*	0.959	QJD25758.1	*USA*	0.957	QKG90867.1	*USA*	0.958
QLE11150.1	*Bangladesh*	0.957	QKU30570.1	*USA*	0.957	QKY77929.1	*USA*	0.958
QLH55840.1	*Bangladesh*	0.958	QKG90399.1	*USA*	0.957	QKV40440.1	*USA*	0.954
QLH56279.1	*Bangladesh*	0.958	QIZ13336.1	*USA*	0.958	QLH01298.1	*USA*	0.958
QLF97736.1	*Bangladesh*	0.957	QKS66305.1	*USA*	0.959	QLG99737.1	*USA*	0.958
QKO25735.1	*Bangladesh*	0.957	QKS65597.1	*USA*	0.960	QLG99677.1	*USA*	0.957
QKK12852.1	*Bangladesh*	0.958	QKV06224.1	*USA*	0.957	QKU31266.1	*USA*	0.957
QLF98048.1	*Bangladesh*	0.957	QLH58601.1	*USA*	0.957	QLI51782.1	*USA*	0.957
QLH55768.1	*Bangladesh*	0.957	QLH58947.1	*USA*	0.958	QLC92097.1	*USA*	0.957
QLH55720.1	*Bangladesh*	0.957	QKU53050.1	*USA*	0.958	QIS61315.1	*USA*	0.956
QLF80217.1	*Brazil*	0.957	QJI07211.1	*USA*	0.958	QJF77147.1	*USA*	0.956
QKS67456.1	*China*	0.958	QKV26659.1	*USA*	0.958	QKG90147.1	*USA*	0.945
YP_009724391.1	*China*	0.958	QLH58085.1	*USA*	0.957	QJE38451.1	*USA*	0.956
QKE10935.1	*Czech Republic*	0.957	QKV39588.1	*USA*	0.957	QJI54254.1	*USA*	0.943
QKS66941.1	*Egypt*	0.958	QKV38810.1	*USA*	0.958	QKS67001.1	*USA*	0.957
QJY78153.1	*Egypt*	0.957	QJS54923.1	*USA*	0.959	QLC91617.1	*USA*	0.958
QKR84274.1	*Egypt*	0.957	QLC47346.1	*USA*	0.957	QJI54123.1	*USA*	0.961
QKR84421.1	*Egypt*	0.957	QKG91107.1	*USA*	0.958	QJQ39741.1	*USA*	0.958
QJT72507.1	*France*	0.959	QJQ39045.1	*USA*	0.958	QJR84550.1	*USA*	0.960
QJT72327.1	*France*	0.958	QJU70306.1	*USA*	0.958	QLG98012.1	*USA*	0.955
QJT72471.1	*France*	0.957	QIZ16438.1	*USA*	0.957	QKE44990.1	*USA*	0.958
QJT72387.1	*France*	0.958	QLI51038.1	*USA*	0.956	QKU32934.1	*USA*	0.959
QJT72951.1	*France*	0.957	QKG86518.1	*USA*	0.956	QLF95773.1	*USA*	0.957
QJW69308.1	*Germany*	0.956	QJC20380.1	*USA*	0.958	QLC94473.1	*USA*	0.958
QKM76547.1	*Germany*	0.957	QKU28847.1	*USA*	0.958	QLH00578.1	*USA*	0.958
QKM76907.1	*Germany*	0.958	QJQ39081.1	*USA*	0.957	QLC93129.1	*USA*	0.958
QJS54155.1	*Greece*	0.957	QKG90495.1	*USA*	0.958	QKS66737.1	*USA*	0.968
QIZ16548.1	*Greece*	0.958	QLH58037.1	*USA*	0.957	QKS65777.1	*USA*	0.965
QJS53735.1	*Greece*	0.955	QJX68859.1	*USA*	0.958	QKS89844.1	*USA*	0.956
QJS54383.1	*Greece*	0.958	QKV08048.1	*USA*	0.957	QJD47539.1	*USA*	0.957
QJS54191.1	*Greece*	0.956	QKE45933.1	*USA*	0.957	QIZ14498.1	*USA*	0.957
QJS53831.1	*Greece*	0.955	QLI51746.1	*USA*	0.958	QKE45765.1	*USA*	0.957
QJS54023.1	*Greece*	0.954	QLG99773.1	*USA*	0.957	QLH00026.1	*USA*	0.957
QKO00487.1 truncated	*India*	0.957	QLH00362.1	*USA*	0.958	QJD23730.1	*USA*	0.958
QLA10225.1 truncated	*India*	0.957	QKV38894.1	*USA*	0.957	QKU52870.1	*USA*	0.957
QLA10069.1 truncated	*India*	0.957	QKU37646.1	*USA*	0.958	QLC92553.1	*USA*	0.957
QKE61733.1	*India*	0.958	QKS65849.1	*USA*	0.959	QKU31818.1	*USA*	0.957

(*continued on next page*)

Table 10 (*continued*)

Protein ID	Geo-location	SE	Protein ID	Geo-location	SE	Protein ID	Geo-location	SE
QLF97952.1	*India*	0.957	QLI51614.1	*USA*	0.957	QKV35400.1	*USA*	0.954
QJX44383.1	*India*	0.958	QJD47299.1	*USA*	0.951	QKV42947.1	*USA*	0.952
QLF98201.1	*India*	0.955	QJC20500.1	*USA*	0.954	QKU37202.1	*USA*	0.956
QLI49698.1	*India*	0.958	QKG87267.1	*USA*	0.957	QKV39324.1	*USA*	0.957
QJX44407.1	*India*	0.957	QKU32046.1	*USA*	0.956	QIS30116.1	*USA*	0.957
QJW00412.1	*India*	0.959	QJS57052.1	*USA*	0.958	QKU32982.1	*USA*	0.957
QLF98261.1	*India*	0.958	QKG87195.1	*USA*	0.957	QJA17681.1	*USA*	0.957
QKY59990.1	*India*	0.958	QJU11458.1	*USA*	0.957	QLC94305.1	*USA*	0.957
QLF98084.1	*India*	0.957	QIS61075.1	*USA*	0.957	QJD48484.1	*USA*	0.958
QLH64816.1	*India*	0.958	QJW28449.1	*USA*	0.957	QLF95737.1	*USA*	0.958
QJY40506.1	*India*	0.958	QLC91905.1	*USA*	0.957	QKN20740.1	*USA*	0.958
QLH93202.1	*India*	0.957	QKG87159.1	*USA*	0.958	QLH57751.1	*USA*	0.959
QLA10165.1	*India*	0.957	QKU31638.1	*USA*	0.957	QLC46314.1	*USA*	0.958
QLG97055.1	*Italy*	0.958	QKU31746.1	*USA*	0.958	QKG81824.1	*USA*	0.958
QJS39568.1	*Netherlands*	0.958	QLF99991.1	*USA*	0.959	QKU53650.1	*USA*	0.957
QJS39520.1	*Netherlands*	0.958	QJX70192.1	*USA*	0.961	QKN19672.1	*USA*	0.957
QJS39616.1	*Netherlands*	0.957	QKG88935.1	*USA*	0.963	QLA47776.1	*USA*	0.957
QLF78310.1	*Poland*	0.957	QLC91545.1	*USA*	0.957	QLG97484.1	*USA*	0.957
QJX45308.1	*Poland*	0.957	QLH01238.1	*USA*	0.957	QLI50282.1	*USA*	0.957
QLH56255.1	*Saudi Arabia*	0.958	QKN20824.1	*USA*	0.957	QKW88844.1	*USA*	0.958
QLH56231.1	*Saudi Arabia*	0.957	QJQ39297.1	*USA*	0.958	QKS90192.1	*USA*	0.958
QKU37034.1	*Saudi Arabia*	0.958	QLB39321.1	*USA*	0.958	QKV39840.1	*USA*	0.960
QLH56099.1	*Saudi Arabia*	0.957	QKV41592.1	*USA*	0.958	QIU81286.1	*USA*	0.957
QHZ00380.1	*South Korea*	0.955	QKV42875.1	*USA*	0.957	QKV07400.1	*USA*	0.957
QIU78768.1	*Spain*	0.956	QLA09656.1	*USA*	0.958			

Table 11
Maxima and minima of SEs across geo-locations.

Geo-location	Min	Max	Range
Australia	0.955	0.96	0.005
India	0.955	0.959	0.004
USA	**0.943**	**0.968**	**0.025**
Bangladesh	0.955	0.959	0.004

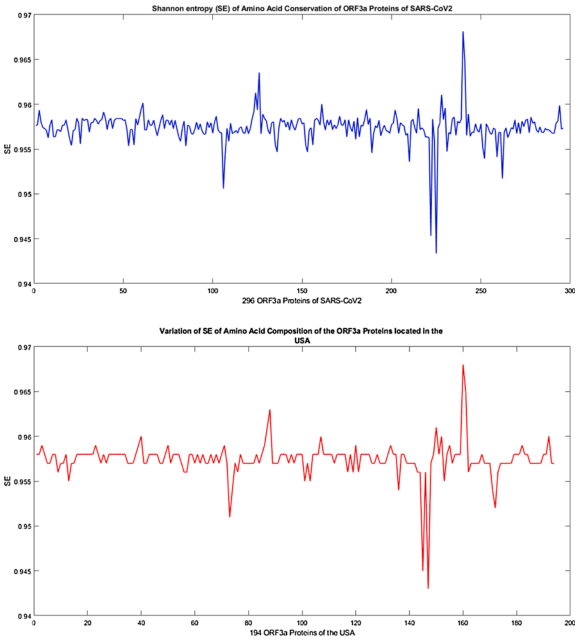


Fig. 18. SE of amino acid compositions of ORF3a proteins.

Acknowledgement

Ms. Diksha Attrish and Ms. Shinjini Ghosh are Interns under the supervision of Dr. Sk. Sarif Hassan through Virtual Internship with Science Leader (VISL) Programme, 2020. Authors thank to the Virtual Internship with Science Leader (VISL) program for their support. Authors would like to thank the reviewers for their constructive suggestions.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.virusres.2021.198441.

References

Al-Osail, A.M., Al-Wazzah, M.J., 2017. The history and epidemiology of middle east respiratory syndrome corona virus. Multidiscip. Respir. Med. 12 (1), 20.

Andersen, K.G., Rambaut, A., Lipkin, W.I., Holmes, E.C., Garry, R.F., 2020. The proximal origin of sars-cov-2. Nat. Med. 26 (4), 450–452.

Brooks, D.J., Fresco, J.R., Lesk, A.M., Singh, M., 2002. Evolution of amino acid frequencies in proteins over deep time: inferred order of introduction of amino acids into the genetic code. Mol. Biol. Evolut. 19 (10), 1645–1655.

Buchholz, U.J., Bukreyev, A., Yang, L., Lamirande, E.W., Murphy, B.R., Subbarao, K., Collins, P.L., 2004. Contributions of the structural proteins of severe acute respiratory syndrome coronavirus to protective immunity. Proc. Natl. Acad. Sci. USA 101 (26), 9804–9809.

Capriotti, E., Fariselli, P., Casadio, R., 2005. I-mutant2. 0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res. 33 (suppl_2), W306–W310.

Capriotti, E., Altman, R.B., Bromberg, Y., 2013. Collective judgment predicts disease-associated single nucleotide variants. BMC Genomics 14 (S3), S2.

Fiorino, G., Allocca, M., Furfaro, F., Gilardi, D., Zilli, A., Radice, S., Spinelli, A., Danese, S., 2020. Inflammatory bowel disease care in the covid-19 pandemic era: the humanitas, milan, experience. J. Crohn's Colitis.

Gao, Y., Yan, L., Huang, Y., Liu, F., Zhao, Y., Cao, L., Wang, T., Sun, Q., Ming, Z., Zhang, L., et al., 2020. Structure of the rna-dependent rna polymerase from covid-19 virus. Science 368 (6492), 779–782.

Guarner, J., 2020. Three Emerging Coronaviruses in Two Decades: The Story of Sars, Mers, and Now Covid-19.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H., 2009. The weka data mining software: an update. ACM SIGKDD Explor. Newslett. 11 (1), 10–18.

Harapan, H., Itoh, N., Yufika, A., Winardi, W., Keam, S., Te, H., Megawati, D., Hayati, Z., Wagner, A.L., Mudatsir, M., 2020. Coronavirus disease 2019 (covid-19): a literature review. J. Infect. Public Health.

Hassan, S.S., Choudhury, P.P., Basu, P., Jana, S.S., 2020. Molecular conservation and differential mutation on orf3a gene in Indian sars-cov2 genomes. Genomics.

Hintze, J.M., Fitzgerald, C.W., Lang, B., Lennon, P., Kinsella, J.B., 2020. Mortality risk in post-operative head and neck cancer patients during the sars-cov2 pandemic: early experiences. Eur. Arch. Oto-Rhino-Laryngol. 1–4.

Huang, Y., et al., 2004. The sars epidemic and its aftermath in China: a political perspective. Learning from SARS: Preparing For the Next Disease Outbreak: Workshop Summary 116–136.

Issa, E., Merhi, G., Panossian, B., Salloum, T., Tokajian, S., 2020. Sars-cov-2 and orf3a: Nonsynonymous mutations, functional domains, and viral pathogenesis. Msystems 5 (3).

Johansson, F., Toh, H., 2010. Relative von neumann entropy for evaluating amino acid conservation. J. Bioinform. Comput. Biol. 8 (05), 809–823.

Kern, D.M., Sorum, B., Hoel, C.M., Sridharan, S., Remis, J.P., Toso, D.B., Brohawn, S.G., 2020. Cryo-em structure of the sars-cov-2 3a ion channel in lipid nanodiscs. BioRxiv.

Law, P.T., Wong, C.-H., Au, T.C., Chuck, C.-P., Kong, S.-K., Chan, P.K., To, K.-F., Lo, A.W., Chan, J.Y., Suen, Y.-K., et al., 2005. The 3a protein of severe acute respiratory syndrome-associated coronavirus induces apoptosis in vero e6 cells. J. Gen. Virol. 86 (7), 1921–1930.

Likas, A., Vlassis, N., Verbeek, J.J., 2003. The global k-means clustering algorithm. Pattern Recognit. 36 (2), 451–461.

Lu, W., Zheng, B.-J., Xu, K., Schwarz, W., Du, L., Wong, C.K., Chen, J., Duan, S., Deubel, V., Sun, B., 2006. Severe acute respiratory syndrome-associated coronavirus 3a protein forms an ion channel and modulates virus release. Proc. Natl. Acad. Sci. USA 103 (33), 12540–12545.

Lu, W., Xu, K., Sun, B., 2010. Sars accessory proteins orf3a and 9b and their functional analysis. Molecular Biology of the SARS-Coronavirus. Springer, pp. 167–175.

Madeira, F., Park, Y.M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., Basutkar, P., Tivey, A.R., Potter, S.C., Finn, R.D., et al., 2019. The embl-ebi search and sequence analysis tools apis in 2019. Nucleic Acids Res. 47 (W1), W636–W641.

Meitzler, J.L., Hinde, S., Bánfi, B., Nauseef, W.M., de Montellano, P.R.O., 2013. Conserved cysteine residues provide a protein-protein interaction surface in dual oxidase (duox) proteins. J. Biol. Chem. 288 (10), 7147–7157.

Minakshi, R., Padhan, K., 2014. The yxxϕ motif within the severe acute respiratory syndrome coronavirus (sars-cov) 3a protein is crucial for its intracellular transport. Virol. J. 11 (1), 75.

Padhan, K., Tanwar, C., Hussain, A., Hui, P.Y., Lee, M.Y., Cheung, C.Y., Peiris, J.S.M., Jameel, S., 2007. Severe acute respiratory syndrome coronavirus orf3a protein interacts with caveolin. J. Gen. Virol. 88 (11), 3067–3077.

Perrella, A., Carannante, N., Berretta, M., Rinaldi, M., Maturo, N., Rinaldi, L., 2020. Editorial-novel coronavirus 2019 (sars-cov2): a global emergency that needs new approaches. Eur. Rev. Med. Pharmacol. 24, 2162–2164.

Phan, T., 2020. Genetic diversity and evolution of sars-cov-2. Infect. Genet. Evolut. 81, 104260.

Ren, Y., Shu, T., Wu, D., Mu, J., Wang, C., Huang, M., Han, Y., Zhang, X.-Y., Zhou, W., Qiu, Y., et al., 2020. The orf3a protein of sars-cov-2 induces apoptosis in cells. Cell. Mol. Immunol. 1–3.

Shen, Z., Xiao, Y., Kang, L., Ma, W., Shi, L., Zhang, L., Zhou, Z., Yang, J., Zhong, J., Yang, D., et al., 2020. Genomic diversity of sars-cov-2 in coronavirus disease 2019 patients. Clin. Infect. Dis.

Siu, K.-L., Yuen, K.-S., Castano-Rodriguez, C., Ye, Z.-W., Yeung, M.-L., Fung, S.-Y., Yuan, S., Chan, C.-P., Yuen, K.-Y., Enjuanes, L., et al., 2019. Severe acute respiratory syndrome coronavirus orf3a protein activates the nlrp3 inflammasome by promoting traf3-dependent ubiquitination of asc. FASEB J. 33 (8), 8865–8877.

Smirnova, E., Firth, A.E., Miller, W.A., Scheidecker, D., Brault, V., Reinbold, C., Rakotondrafara, A.M., Chung, B.Y.-W., Ziegler-Graff, V., 2015. Discovery of a small non-aug-initiated orf in poleroviruses and luteoviruses that is required for long-distance movement. PLOS Pathog 11 (5), e1004868.

Strait, B.J., Dewey, T.G., 1996. The shannon information entropy of protein sequences. Biophys. J. 71 (1), 148–155.

Tang, X., Wu, C., Li, X., Song, Y., Yao, X., Wu, X., Duan, Y., Zhang, H., Wang, Y., Qian, Z., et al., 2020. On the origin and continuing evolution of sars-cov-2. Natl. Sci. Rev.

The Mathworks, Inc, 2020. Natick, Massachusetts. MATLAB version 9.3.0.713579 (R2020a).

To, J., Torres, J., 2018. Beyond channel activity: protein-protein interactions involving viroporins. Virus Protein and Nucleoprotein Complexes. Springer, pp. 329–377.

Van Doremalen, N., Bushmaker, T., Morris, D.H., Holbrook, M.G., Gamble, A., Williamson, B.N., Tamin, A., Harcourt, J.L., Thornburg, N.J., Gerber, S.I., et al., 2020. Aerosol and surface stability of sars-cov-2 as compared with sars-cov-1. N. Engl. J. Med. 382 (16), 1564–1567.

Wang, K., Xie, S., Sun, B., 2011. Viral proteins function as ion channels. Biochim. Biophys. Acta (BBA)-Biomembr. 1808 (2), 510–515.

Wang, X., Zhou, Q., He, Y., Liu, L., Ma, X., Wei, X., Jiang, N., Liang, L., Zheng, Y., Ma, L., et al., 2020. Nosocomial outbreak of covid-19 pneumonia in Wuhan, China. Eur. Respir. J. 55 (6).

Zhang, Y.-Z., Holmes, E.C., 2020. A genomic perspective on the origin and emergence of sars-cov-2. Cell.

Zhong, W., Altun, G., Harrison, R., Tai, P.C., Pan, Y., 2005. Improved k-means clustering algorithm for exploring local protein sequence motifs representing common structural property. IEEE Trans. Nanobiosci. 4 (3), 255–265.