Emerging mutations in spike and other structural proteins of SARS-CoV-2

Farwa Mukhtar[1], Muhammad Tahir khan[2], Arif Malik[1], Shaoliang Peng[3], Doaa Darwish[4], Shabbir Muhammad[5], Sami Ullah[5], Ahsan Sheikh[1], and Dongqing Wei[6]

[1]The University of Lahore
[2]Capital University of Science and Technology
[3]Peng cheng lab
[4]University of Tabouk
[5]King Khalid University
[6]Shanghai Jiao Tong University

March 13, 2021

Abstract

The structural proteins, spike (S), nucleocapsid (N), membrane (M), and envelope (E), of severe acute respiratory syndrome (SARS-CoV-2) play a critical role from attachment to replication and virulency. Recently a bulk of genomes have been sequenced from different geographical regions with significant number of variations. Therefore, the current study was aimed to find variations in the structural proteins. This is the first comprehensive study in which we screened 2,95,000 complete genomes in global initiative on sharing all influenza data (GISAID), submitted from December 2019 to December 2020. We detected 4725 non-synonymous mutations in S, 627 in M, 259 in E, and 1631 mutations in N protein, among which the most frequently occurring mutations in S protein are D614G (n=2,66,513), A222V (n=59,697), L18F (n=28,015) and that of M protein are; T175M (n=1286), D3G (n=968), L17I (n=621), A2V (n=463), and A2S (n=460). The most commonly circulating variants in E includes, S68F (n=419), P71S (n=264), and L73F (n=218). Similarly, the N protein also harbored the most common variants which include; R203K (n=82,570), G204R (n=81,858), and A220V (n=39,729). The frequency of N501Y (n=4362) in S is determining a tight interaction of CoV-2 RBD with ACE2. These wide range of mutations in structural proteins may not only affect the therapeutic efforts but also the vaccines efficacy and diagnostics specificity. We suggest that geographically strain specific variations should be investigated for effective drugs, vaccine, and the antibodies combinations. Alternatively, immune boosting compounds might be very useful for successful eradication of CoV-2 infections.

Title: Emerging mutations in spike and other structural proteins of SARS-CoV-2

Running title: Mutations in structural proteins of SARS-CoV-2

Farwa Mukhtar[1], Muhammad Tahir Khan*[1,3], Arif Malik[1], Shaoliang Peng[3], Doaa B. Darwish[5], Shabbir Muhammad[2], Sami Ullah[4], Ahsan Sattar Skeikh[1], Dong Qing Wei[3]*

1-Institute of Molecular Biology and Biotechnology (IMBB), The University of Lahore. KM Defence Road, Lahore, Pakistan Postal code: 58810 Ph: +92 (0)42 111865865 tahirmicrobiologist@gmail.com, muhammad.tahir8@imbb.uol.edu.pk, arif.malik@imbb.uol.edu.pk, farwamukhtar97@gmail.com, ahsan.sattar@imbb.uol.edu.pk

2- Department of Physics, College of Science, King Khalid University, Abha 61413, P.O. Box 9004, Saudi Arabia mshabbir@kku.edu.sa

3-A State Key Laboratory of Microbial Metabolism, Shanghai-Islamabad-Belgrade Joint Innovation Center on Antibacterial Resistances, Joint International Research Laboratory of Metabolic & Developmental Sciences and School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200030, P.R. China

3-B-Peng Cheng Laboratory, Vanke Cloud City Phase I Building 8, Xili Street, Nashan District, Shenzhen, Guangdong, 518055, P.R China.**Email:**slpeng@hnu.edu.cn

4-Department of Chemistry, College of Science, King Khalid University Abha, 61413, Saudi Arabia. Project No. 2-N20/22. samiali@kku.edu.sa

6- Botany Department, Faculty of science, Mansoura University, Egypt and Department of Biology, Faculty of science, University of Tabuk 71491, Saudi Arabia

ddarwish@ut.edu.sa, d_darwish@man.edu.eg

***Corresponding authors**

1-Dong-Qing Wei

1-State Key Laboratory of Microbial Metabolism, Shanghai-Islamabad-Belgrade Joint Innovation Center on Antibacterial Resistances, Joint International Research Laboratory of Metabolic & Developmental Sciences and School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200030, P.R. China. +86 021-34204717.

2-Peng Cheng Laboratory, Vanke Cloud City Phase I Building 8, Xili Street, Nashan District, Shenzhen, Guangdong, 518055, P.R China. **Email:**dqwei@sjtu.edu.cn

*****Co-corresponding**

Muhammad Tahir Khan

Institute of Molecular Biology and Biotechnology (IMBB), The University of Lahore. KM Defence Road, Lahore, Pakistan Postal code: 58810. https://orcid.org/0000-0003-1158-2133

Ph: +92 (0)42 111865865. **Email:**muhammad.tahir8@imbb.uol.edu.pk

Data availability statement

Data in this study is available as supplementary files.

Funding statement

Dong-Qing Wei is supported by grants from the National Science Foundation of China (Grant No. 32070662, 61832019, 32030063), the Key Research Area Grant 2016YFA0501703 of the Ministry of Science and Technology of China, the Science and Technology Commission of Shanghai Municipality (Grant No.: 19430750600), as well as SJTU JiRLMDS Joint Research Fund and Joint Research Funds for Medical and Engineering and Scientific Research at Shanghai Jiao Tong University (YG2021ZD02). The computations were partially performed at the Pengcheng Lab. and the Center for High-Performance Computing, Shanghai Jiao Tong University.

Conflict of interest disclosure.

None

Ethics approval statement

N/A

Patient consent statement

N/A

Abstract

The structural proteins, spike (S), nucleocapsid (N), membrane (M), and envelope (E), of severe acute respiratory syndrome (SARS-CoV-2) play a critical role from attachment to replication and virulency. Recently a bulk of genomes have been sequenced from different geographical regions with significant number of variations. Therefore, the current study was aimed to find variations in the structural proteins. This is the first comprehensive study in which we screened 2,95,000 complete genomes in global initiative on sharing all influenza data (GISAID), submitted from December 2019 to December 2020. We detected 4725 non-synonymous mutations in S, 627 in M, 259 in E, and 1631 mutations in N protein, among which the most frequently occurring mutations in S protein are D614G (n=2,66,513), A222V (n=59,697), L18F (n=28,015) and that of M protein are; T175M (n=1286), D3G (n=968), L17I (n=621), A2V (n=463), and A2S (n=460). The most commonly circulating variants in E includes, S68F (n=419), P71S (n=264), and L73F (n=218). Similarly, the N protein also harbored the most common variants which include; R203K (n=82,570), G204R (n=81,858), and A220V (n=39,729). The frequency of N501Y (n=4362) in S is determining a tight interaction of CoV-2 RBD with ACE2. These wide range of mutations in structural proteins may not only affect the therapeutic efforts but also the vaccines efficacy and diagnostics specificity. We suggest that geographically strain specific variations should be investigated for effective drugs, vaccine, and the antibodies combinations. Alternatively, immune boosting compounds might be very useful for successful eradication of CoV-2 infections.

Keywords: SARS-CoV-2; genome; mutations; spike; nucleocapsid; envelope

1. Introduction

In December, 2019 china reported a disease with pneumonia like conditions, resulting in respiratory malfunctioning due to some viral attack. Later that virus proved lethal and turned into global pandemic. World Health Organization (WHO) named the disease as "coronavirus disease 19 (COVID-19)". Following the international standards of nomenclature, virus was declared as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) due to its taxonomic and genomic relationship with the species of sever acute respiratory syndrome-related coronavirus[1].

In the initial stages of pandemic, it was centered to china only but Spain, Italy, Brazil, France, United States of America, Iran, and India were also severely affected in short period. World had seen the major lockdown of the history in the year 2020 to reduce the spread of CoV-2 that had greatly affected the economy of world powers. Instead of initial precautions taken by the people, the virus affected 108M people around the globe with 80.8M recoveries and 2.31M deaths till January 2021. World had also previously experienced corona as MERS-CoV and SARS-CoV that had affected Middle East and other countries to a large extent in 2012 and 2002. Coronaviruses profoundly spread in humans, other mammals and birds mainly affecting their respiratory, liver, and intestinal and nervous systems[2,3].

Human coronaviruses (HCoVs) were first identified in the mid-1960s. Till now seven HCoVs are known which include two α coronaviruses CoV-229E and CoV-NL63 and four β coronavirus as CoV-OC43, CoV-HKU1, SARS-CoV, MERS-CoV and CoV-2 [4,5]. As CoV-2 belongs to the formerly known family of coronaviruses it holds on to structural formations and show close genomic similarity to the SARS-CoV. The CoV-2 harbors a linear single-stranded positive RNA genome rapidly infecting vertebrates, named for the crown like spikes on their surface[6]. Subsequently after crown like surface projections it has spike protein (S), envelope protein (E), membrane protein (M), and nucleocapsid protein (N)[7]. These structural proteins are responsible for viral replication, virion-receptor attachments and thus involved in pathogenicity, spreading, and entry of virus into host organism.

Within a short period of time the virus shows its mutating ability, giving rise to new resistant more pathogenic strains which could be more difficult to counter. It may void the drugs designed against CoV-2 or may reduce

the vaccine efficacy due to large number of variants. Genomic composition of CoV-2 shows 12 functional ORFs (open reading frames), 11 protein coding genes, with 12 expressed proteins and 5' capped mRNA consist of 38% GC content with poly-A tail at 3' end followed by UTR[6]. The ORFs are arranged on mRNA of CoV-2 as ORF1a, ORF1b, Spike (S), ORF3a, Envelope (E), Membrane (M), ORF6, ORF7a, ORF7b, ORF8, nucleocapsid (N), and ORF10 [8] (Figure 1). The genome of CoV-2 encodes 16 non-structural proteins (NSPs), four structural proteins, and other polyprotein1a and polyprotein1b[9]. Among the NSPs, replicase and protease are important for the viral genome replication, along with structural proteins and also potential drug targets [7,10,11].

Although both, structural and non-structural proteins of CoV-2 are important to investigate, here we investigated the variations, existed only in structural proteins because of their potential drug targets and vaccinal importance. This is the first comprehensive study in which we screened 2,95,000 complete genomes of SARS-CoV-2 for variants in the structural proteins. Exploring the degree of variations in the important target proteins might be helpful in projecting the pathogenicity and transmission of CoV-2 strains around world. Presence of large variations may lead to the confirmational changes in the targets, leading to therapeutic failure. Diagnostic accuracy may also be affected if proper screening has not been performed. Alternatively, geographic strains specific vaccine and antiviral might be more effective.

2. Methodology

2.1 Genome sequence retrieval

The CoV-2 genome sequences were retrieved from global science initiative on sharing all influenza data (GISAID) (Dec, 2019-Dec, 2020) (https://www.gisaid.org/) [12]. About 295000 genomes of CoV-2 reported worldwide, were screened for variants in structural proteins. All the sequences were aligned with reference genome of Wuhan-Hu-1 (Accession NC_045512) using CoVsurver application (https://www.gisaid.org/epiflu-applications/covsurver-mutations-app/). All the identified mutations in structural proteins of CoV-2 were separated and arranged in the form of excels sheets. The statistical analysis was performed to screen the most common variants. The GISAID is sharing the virus data to publish results and the metadata that concerns of public health scientists. The GISAID is an effective server for rapid sharing all kind of data including unpublished influenza data.

2.3 Structural information

The sharing of infectious agent genomic data is critical against disease outbreaks. Scientists are well awared to share the genomic and proteomic data for better management of infectious diseases to develop countermeasures. The SARS-CoV structural protein data was retrieved from protein data bank (PDB) [13], I-TASSER (M=QHD4MOE) [14,15], and SWISS-MODEL (https://swissmodel.expasy.org/repository/species/2697049) (PDB IDs: E=7k3g, N=6m3m, 6yun, S=7l09). The PDB (http://www.rcsb.org/pdb/) is the single structural data archive of biological molecules. All the data is primary, collected from depositors across the globe. The data of macromolecules also contain the coordinates, structures and method of structure determination. The I-TASSER and SWISS-MODEL servers have already modelled the full-length proteins using NCBI reference data (NC_045512) (GenBank MN908947). The SWISS-MODEL data are openly accessible to everyone. The server also provides PDB experimental and complex structural information of each SARS_CoV-2 protein. The I-TASSER developed full-length COV-2 protein which is freely available to the academic community (http://zhang.bioinformatics.ku.edu/I-TASSER).

Structure dynamics

Some of the most common mutations and their effect on structure were computed through online DynaMut server [16]. The server implements normal mode methods, and mutation effect prediction which can be used to analyze the mutations' effect on protein stability and flexibility behind vibrational entropy changes. The impact of a mutation is predicted through normal mode dynamics and graph-based signatures. This approach outperforms (P-value < 0.001) and the results are displayed in good resolution.

3. Results and discussion

We screened 2,95,000 complete genomes of CoV-2, reported worldwide in GISAID (Dec, 2019-Dec, 2020), for variants in S, N, M, and E proteins. A large number of mutations were detected in structural proteins of CoV-2 (Table 1). However, the patterns of mutation are still unknown. This wide range of variations projects toward the number of strains variation of virus present worldwide.

3.1 Genomic Organization of SARS-CoV-2

The CoV-2 is β-coronavirus with small genome size (˜29.9Kb), forming 9860 amino acids (NC_045512.2) [17,18]. Currently about 0.3 million genomes have been sequenced and the data is accessible to researchers from Global Initiative on Sharing All Influenza Data (GISAID)[19] across the globe. The data could be analyzed for variations in viral proteins with reference sequence (GenBank accession number MN908947.3)[18,20]. CoV-2 mutations might affect the susceptibility to CoV-2 infection or severity of COVID-19. In the current study we screened about 0.3 million genome sequences for mutation in S, E, M, and N proteins. All the mutations were analyzed through CoVsurver application (https://www.gisaid.org/epiflu-applications/covsurver-mutations-app/). The structural proteins harbor a large number of mutations in different geographic regions (Supplementary files). However, the level of transmission and pathogenicity is largely unknown behind these variations. Geographic specific vaccines and drugs may be designed after careful analysis of mutations in the target's proteins for better management of CoV-2 infections.

3.3 Structural Proteins of SARS-CoV-2

3.3.1 Spike Protein

The CoV-2 is attached to the host receptors through S proteins. These are glycoproteins that are attached to the surface of virus giving it crown like appearance. Molecular weight of S protein is 141178 kDa and it has 1273 amino acids [21]. Genome sequencing has shown that S protein of COV-2 is 75% similar to SARS-CoV S protein. Ectodomain of virus have two sub-units S1 (residues 13-685), and S2 (residues 686-1273). These subunits gave clove like shape when three S1 subunits join to perform receptor binding and stem of S2 made up of timer that performs membrane fusion [22]. The S1 subunit of CoV-2 is 70% and S2 is 99% similar to the S proteins of SARS-CoV [18]. The S protein is responsible for the formation of attachments between infected and non-infected cells and thus involved in the spread of virus [23]. The Angiotensin-Converting Enzyme 2 (ACE2) is receptor of CoV-2 S protein that is present on the membrane of host cells.

Fusion of viral membrane to the host membrane undergoes structural transformation of the S protein. SARS-CoV-2 is 10 to 20 time more stable toward this viral-host binding because S1 subunit of S protein comprises of Receptor Binding Domain (RBD) [24]. It identifies and attaches to ACE2 receptors present on host cells. This enhanced receptor-spike interaction may be due to polymorphism at 501T resulting in promoting infectivity [25]. The RBD of CoV-2 and ACE2 receptor of host cell are interacted by hydrogen bonding as well as salt bridges. There are 17 RBD residues that interact with 20 ACE2 residues, out of them K417 of RBD form salt bridge with D30 of ACE2 rest of them form hydrogen bondings with respective residues [26,27]. Virus appears to be more proteolytic by the host cells proteases because of the unique prolonged loop formed by S1 and S2 subunits of CoV-2. S1subunit helps S2 subunit to achieve stable configuration during binding with host receptors by shedding and destabilizing itself. The RBD is present near the central pocket of S protein in its downward phase configuration. The S2 subunit consists of a fusion peptide (FP) and proteolytic site with a central FP along with two heptad-repeat before the transmembrane domain [28].

The S protein shows greater number of mutations among all structural proteins. Some of the most frequently occurred mutations reported in S protein are L18F, A222V, N439K, S477N, N501Y, D614G and P681H (Figure 3). Amino acid position 222 to 681 has been found as the most variable part as compared to whole S sequence, this includes S1 subunit of S protein. These frequently occurred mutations are mainly affecting NTD and RBD region of S1 subunit. We detected approximately 4725 different mutations, present along the whole coding region. Some of the most frequently seen mutations are listed in the table along with amino acids and their reported accession IDs (Table 1). The D614G mutation is the most frequent mutation of S protein with frequency of 266513; it has already been term as more infectious as compared to other mutant strains [29]. This mutant D614G strain increases the infectivity as it makes virus entry into the host cells

more efficient as compared to the original strain and also reduce the shedding of S1 domain [30]. According to a recent study, D614G mutant increases the replication of virus in the epithelial lining of the human lungs and other airways of body mainly upper respiratory tract by enhancing the stability as well as infectivity and load of the virus [31,32]. The highest frequency of mutation (A222V) in spike has been found across Europe prevalent in many countries. However, this rise has no relationship with A222V in transmissibility [33].

The S1 subunit has been divided into N-terminal domain (NTD), receptor binding domain (RBD) comprised of residues 319-541 with receptor binding motif (residues 437-508). The RBD harbors a higher frequency of certain mutations than other domains. Mutation S477N has been detected in 16914 genomes, present in RBM. Similarly, the other most common variants present in RBM are N439K (5725), N501Y (4362), and Y453F (968) (Table 3). Among all the RBD and ACE2 binding residues, N501 showed more variations (N501Y (4362), N501T (48), N501S (8), N501R (1), N501I (1), N501G (1)). The N501Y (Table 2) is determining a tight interaction of CoV-2 RBD with ACE2 [34]. We found that mutations N501Y has a stabilization effect on S proteins ($\Delta\Delta$G: 0.535 kcal/mol) (Figure 2) when computed through DynaMut server [16]. Mutation L5F has been detected in 3813 genomes, present in signal peptide. As already reported RBD mutation A348V (7), V367F (155), and A419S (11) that shows high antigenicity were also seen with a notable frequency exhibiting the mutant RBD [35].

Figure 1. SARS CoV-2 genome organization and mutations in structural proteins. Frequency of mutations has been shown.

In the CoV-2 lysine present at position 417 shows stronger interaction toward aspartic acid present on 30^{th} position of ACE2 receptor. This bond actually enhances the interaction between host and receptor by making it more stronger [36], but we have seen three mutations at 417 position of receptor binding residue where it has replaced by Asparagine, Arginine and Proline with a frequency of 218, 2 and 1 respectively that shows K417N mutation may changes the receptor binding patterns of RBD and effects its infectivity (Table 2).

Numerous mutations in all the ACE binding domain of S proteins have been detected. Among which N501Y is present in highest frequency (4362). However, the effect of majority of these mutations on binding affinity is unknown. The Q493K and N501T in RBD might alter the binding affinity of S and ACE2 [34]. Similarly, E484K, present in RBD evade the antibodies. We detected E484R in UK genomes (supplementary file S1) in only two genomes which need further characterization. In a more recent study, Tyr449Asn individual substitutions retracted S2M11 antibody mediated neutralization, whereas the Leu455Phe variant decreased the neutralization potency [37]. The G446V variant in RBD has been detected in 46 genomes. The antibodies combinations, targeting different neutralizing epitopes might a useful strategy.

The transmembrane domain harbors P1263L mutations in 911 genomes sequences (Figure 3). Variants has also been detected in the heptad repeat1 (HR1) (residues 1163-1213), present in S2 subunits. Among them S982A is the most frequent (No. 3670) followed by D936Y (No.1239) and S939F (No. 555). Three variants (D1163Y, G1167V, V1176F) were also detected in high frequency (1903, 1753, 933) in heptad repeat 2 (HR2) region of S protein. Alanine is considered as a best choice in forming helix while on the other hand valine is subjected as a bad choice in helix formation. Therefore, we can assume that any mutation in the helical alanine with valine may results in secondary structure changes as A879S (456), A892V(45), and A930V (128) have seen with a notable frequency range, may result in forming beta-sheet instead of helix [38].

Some studies have shown four unique inserts in CoV-2 S1 subunit. Insert 1 is related to N terminal domain while insert 2 and 3 are implicated by CTD. Insert 4 lies at the intersection of two domains of S1 subunit. And other 1,2 and 3 inserts are similar in configuration to HIV-1 gp-120 while insert 4 is similar to Gag proteins [39]. Polymorphism other than 501T in RBD with certain amino acids in S protein of CoV-2 results in good binding with ACE2 receptor. Amino acids Glu493, Asn501, Leu455, Phe486, and Ser494 are subjected as boosting ACE2 binding [25].

3.3.2 Envelope Protein (E)

The E protein of is composed of 76 amino acids [40] weighing 8 to 12 kDa [41]. The E protein consists of NTD, hydrophobic domain and CTD [42]. These domains are arranged from NTD (aa1-9) to CTD (aa38-75), and hydrophobic region spreads from 10 to 37 (Figure 3) [43]. Ionic pores are formed across the membrane due to the arrangement of hydrophobic tail region. Structurally it shows pentameric configuration having 35 alpha helical regions and 40 looped ring regions formed by the hydrophobic tail. This pentameric configuration of hydrophobic tail can be affected by the interactions with in C terminal domain [44]. These pores act as ion channels that provide movements across the membrane and enhancing the pathogenicity of virus [45]. Some novel mutations that were seen in the E protein sequence, does not exist in already existing coronaviruses. In the E protein of CoV-2 at 69^{th} arginine is mutated by isoleucine, threonine and lysine (R69I, R69T, and R69K). Moreover in amino acid sequence of CoV-2 at positions 55 and 56, serine and phenylalanine have been reported instead of threonine and valine [46].

Being the smallest structural protein (75aa) of CoV-2, all the residues positions of the E protein mutated with non-synonymous (S2). Most of these mutations have been detected in the ending sequence of E, including 68 to 73 amino acids. Some of the most frequently seen mutations in E protein are; T91, S68F, R691, P71L, P71S, and L73F (Figure 3). Some other E protein mutations have been listed in the table along with their amino acids (Table 4 and S2). Mutations in the CTD like, S55F (128), V62F (129), R69I (159) may affects the virus pathogenesis, altering the binding of E protein to tight junction, associated PALS1. The transmembrane variants, T9I (168), F20L (90), L21F (84), V24M (76) and T30I (72) may effects the homo pentameric configuration of the E protein [47]. The E protein may also be an effective drug target as it imparts equally in the pathogenicity and cytotoxicity of virus.

The E protein contributes viral pathogenicity, cytotoxicity, and also responsible for viral assembly and release. It forms viroporins which are hydrophobic in nature [48]. Proline residues facilitates in targeting Cis-Golgi complex by the hydrophobic tail, present in the cytoplasm. Through some golgi complex associating elements, NTD also helps tail region in targeting golgi complex as mutations in the tail region may affects its efficiency. The release of these virion particles are facilitated by the ionic gradient, present in the endoplasmic reticulum and golgi compartment by the E protein [40].

The CTD of E proteins has some common mutations whose stability was predicted. The DynaMut prediction outcome of L73F ($\Delta\Delta$G: -0.417 kcal/mol), P71S ($\Delta\Delta$G: -0.255 kcal/mol) exert a destabilizing effect. However, T9I ($\Delta\Delta$G: 0.190 kcal/mol), P71L ($\Delta\Delta$G: 0.012 kcal/mol), and S68F ($\Delta\Delta$G: 0.362 kcal/mol) shows a stabilizing effect.

3.3.3 Membrane Protein (M)

Non-synonymous mutations are also increasing in M protein, occurred in the whole coding region except position 50, 51, 72, 102, 103, 116, 118, 120, 159, 178, 179, and 187. However, the most common have been detected in NTD and CTD as compared with TM domains (Figure 5). The M is a glycoproteins, and most abundant in CoV particles [49]. It forms four main domains, NTD, triple transmembrane domain, consisting of three transmembrane helices (TMH), attached to CTD and N-linked glycosylated protein with conserved domain of 13 amino acids [50] (Figure 4). The M is present in two configurations; long (Endodomain) and compact form because it undergoes elongation and compression [51]. Although spikes are evenly present on both forms of M proteins but more likely to be present on long form, promoting S insertions. After being translated into the polysomes attached to the membrane of endoplasmic reticulum, M proteins are transported to the golgi complex.

The M protein facilitates in virion formation while interacting with E proteins in the golgi complex. Out of three TMH, the first one provides union of M protein, enhanced membrane affinity, and detention in golgi [52]. The M protein increase the virus transmission by blocking the Nuclear Factor Kappa B (NFkB), required for immune responses against pathogens [53]. The M protein of the SARS CoV also shows the activation of b interferons (IFN-b) in cell lines [54]. Such humoral responses are generated by the M protein and its antigenic epitopes have been found in the TM1 and TM2 region of protein. The M protein interacts with itself (homotypic) and other structural proteins like S, E, and N (heterotypic) helping in budding and

formation of new virus particles. Homotypic interaction is possible with the help of residues present all along the M protein sequence also in TM regions while CTD is involve its heterotypic interactions with E and N protein. The M protein is also involved in the ribonucleoprotein packaging, and dileucine residues present at 219 and 220 positions (L219, L220) are essential for nucleocapsid packaging [55]. Variations in membrane protein may cause adverse effect as it is involved in regulating the virus life cycle. A large number of mutations has been seen in the CTD of the M protein, mostly from 142-209 amino acids. Some of the frequently seen mutations in M protein are A2V, A2S, D3G, L17I, H125S, T175M, and D209Y (Figure 4). However, what kind of effect these variants may produce is still unknown. The crystal structure of M is not available and neither there is any suitable template for homology modeling. We downloaded the 3D structure of M from I-TASSER (ID QHD43419) to compute the effect of common variants on M protein thermodynics.

The effect of common mutations on NTD of M structure is mainly seems stabilizing

We detected 627 in M protein mutations (S3), some of them with greater frequencies are listed in the table (Table 6). Mutations have seen in the residues required for nucleocapsid packaging (L219C (1), L219T (1), L220Y (2) and L220I (1)) respectively. These mutations can alter the ribonucleoprotein packaging which may result in lagging the virion particles formations. The M protein plays important role in viral circulation that suggests it to be a therapeutic drug target to retard the virion particles formations or reduce inflammations in the host cells.

3.3.4 Nucleocapsid Protein (N)

The N protein binds the viral RNA, forming ribonucleoprotein which facilitates virus interacting with the cellular processes and entering into the host cells [56]. The N protein of CoV-2 (419aa) involved in RNA package and virus particle release [57]. It can be detected at the initial stages of infection. Nucleic acid sequences of N protein of CoV-2 is 90% conserved as that of SARS-CoV [58]. It forms Replication Transcription Complexes [59], important for viral genome synthesis.

As N protein is involved in virus replications machinery, any mutation in N protein may affect virus pathogenicity. Variations in amino acid 193 to 235 are more frequently mutating in the NTD and rest of serine rich (SR) linker region. Some of the reported mutations having higher frequencies globally are S194L, P199L, R203K, G204R, A220V, M234I, A376T and A398V (Figure 6).

The N protein has two domains named N-terminal domain (NTD) and C-terminal domain (CTD), connected by a serine rich SR linker region (Figure 4) [60]. RNA bound to the N protein through NTD and more precisely at N45-181 region of NTD that exist as monomer [61]. The most frequently variants; P67S (1860), D103Y (2233), and H145Y (780) have been detected in the NTD. However, the effect of these mutation on its viral RNA binding affinity is still unknown. Amino acids, required for binding of SARS-CoV RNA are present at position R94 and Y122. Binding efficiency of CoV-2 is 6 to 8 time more than the previous viruses as it has dimeric CTD, forming two disordered regions around NTD while previous virus has a single CTD, as a result the combination of linker region, NTD and CTD are important in improved binding capacity of N protein to the RNA genome [62].

Dimers of N protein are formed that play important role during the interaction of SR moieties of linker regions with the central region [63]. The CTD has the residues that self-associate and form homodimers. This basic nature of N terminus shows it as binding site of viral RNA [64]. For viral genome processing The interaction of N and Nsp3 proteins is essential for virus genome processing, the C terminal domain of N protein anchors with the Nsp3 protein and the residues forming this interaction can be a potential drug target [11] [65]. The great force of repulsion present with in domain components that provides electrostatically larger binding surface to RNA genome and it also prevent oligomer and within domain interactions. So when RNA binds to these domains it neutralizes the charges and as a result protein molecules are attracted toward the genome and oligomerize to form nucleocapsid [62]. About 1034 mutations had identified globally, out of which 367 are primer binding sites, 684 are AA substitutions across 317 unique positions, also having 82, 21, 83 of NTD, CTD and SR linker region, and 11 in-frame deletion in the linker region and other were in NTD region

[66]. We detected 1632 different kinds of mutation in N, prevalent in all domains. The SR linker region harbors the highest frequency of mutations (Figure 6). The R203K has been detected in the 82570 genomes (S4) followed by G204R (81858). The frequency of mutations in CTD is fewer than NTD and SR region (S4).

The CTD of N protein has also shown a number of mutations. Some of these mutations having higher frequencies are listed in the table (Table 4). The CoV-2 has two important sites in N protein which include RNA binding domain and other include Phosphorylation sites, both plays an important role in binding with RNA and its replication, transcription and packaging processes as well as in cell cycle. Any mutations in these regions would be of great importance like S186Y (549), S197L (2165), S202N (756), R203K (82570), and G204R (81858) are phosphorylation sites in N protein, have undergone variations [67] (Table 6). The RNA binding domain (RBD) (aa40-180) shows high frequency variations including P67S (1860), D103Y (2233), and H145Y (780) [68]. These mutations can alter the RNA binding patterns and may affect virus replication and transcription processes. The N protein also shows its importance in viral proliferation and functioning, which necessitates for developing effective therapeutics against N to prevent virus proliferation.

The N has been given a considerable importance in diagnosis, and being proposed as alternative to spike, for designing and synthesis of vaccine and drug target. However, the emergence of such large number of variants may challenge diagnostics and vaccine designing efforts. We therefore, propose a continuous screening of genome for better management of ongoing structural proteins evolution of CoV-2.

4. Conclusion

A large number of different kinds of non-synonymous mutations in CoV-2 structural proteins, S (4725), E (259), M (627), and N (1631) are present, covering the entire coding sequences, showing that antiviral and vaccines efficacy might be compromised. This will make it difficult to design particular drugs against structural targets. Further, investigating the stability of these mutation on structural dynamics will enhance our understanding about the viral pathogenicity and transmission pattern. The variants may also affect drug interactions, diagnostics, and virulence of CoV-2, especially the N which is being considered as alternative to S. Geographic genome specific therapy might be useful. Alternatively, immune boosting agents against CoV-2 might be a more successful strategy in the current scenario.

5. Competing interest

All the authors have no competing interests.

6. Funding

Dong-Qing Wei is supported by grants from the National Science Foundation of China (Grant No. 32070662, 61832019, 32030063), the Key Research Area Grant 2016YFA0501703 of the Ministry of Science and Technology of China, the Science and Technology Commission of Shanghai Municipality (Grant No.: 19430750600), as well as SJTU JiRLMDS Joint Research Fund and Joint Research Funds for Medical and Engineering and Scientific Research at Shanghai Jiao Tong University (YG2021ZD02). The computations were partially performed at the Pengcheng Lab. and the Center for High-Performance Computing, Shanghai Jiao Tong University.

References

[1] S. Satarker, M. Nampoothiri, Structural Proteins in Severe Acute Respiratory Syndrome Coronavirus-2, Arch. Med. Res. 51 (2020) 482–491. https://doi.org/10.1016/j.arcmed.2020.05.012.

[2] M.J. Bakkers, Y. Lang, L.J. Feitsma, R.J. Hulswit, S.A. de Poot, A.L. van Vliet, I. Margine, J.D. de Groot-Mijnes, F.J. van Kuppeveld, M.A. Langereis, Betacoronavirus adaptation to humans involved progressive loss of hemagglutinin-esterase lectin activity, Cell Host Microbe. 21 (2017) 356–366.

[3] J. Rasheed, A.A. Hameed, C. Djeddi, A. Jamil, F. Al-Turjman, A machine learning-based framework for diagnosis of COVID-19 from chest X-ray images, Interdiscip. Sci. Comput. Life Sci. (2021).

https://doi.org/10.1007/s12539-020-00403-6.

[4] C. Yin, Genotyping coronavirus SARS-CoV-2_ methods and implications, (2020) 9.

[5] S.M. Nur, Md.A. Hasan, M.A. Amin, M. Hossain, T. Sharmin, Design of Potential RNAi (miRNA and siRNA) Molecules for Middle East Respiratory Syndrome Coronavirus (MERS-CoV) Gene Silencing by Computational Method, Interdiscip. Sci. Comput. Life Sci. 7 (2015) 257–265. https://doi.org/10.1007/s12539-015-0266-9.

[6] A.A.T. Naqvi, K. Fatima, T. Mohammad, U. Fatima, I.K. Singh, A. Singh, S.M. Atif, G. Hariprasad, G.M. Hasan, Md.I. Hassan, Insights into SARS-CoV-2 genome, structure, evolution, pathogenesis and therapies: Structural genomics approach, Biochim. Biophys. Acta BBA - Mol. Basis Dis. 1866 (2020) 165878. https://doi.org/10.1016/j.bbadis.2020.165878.

[7] A. Khan, M. Tahir Khan, S. Saleem, M. Junaid, A. Ali, S. Shujait Ali, M. Khan, D.-Q. Wei, Structural insights into the mechanism of RNA recognition by the N-terminal RNA-binding domain of the SARS-CoV-2 nucleocapsid phosphoprotein, Comput. Struct. Biotechnol. J. 18 (2020) 2174–2184. https://doi.org/10.1016/j.csbj.2020.08.006.

[8] Y.-Z. Zhang, E.C. Holmes, A Genomic Perspective on the Origin and Emergence of SARS-CoV-2, Cell. 181 (2020) 223–227. https://doi.org/10.1016/j.cell.2020.03.035.

[9] Y.-R. Guo, Q.-D. Cao, Z.-S. Hong, Y.-Y. Tan, S.-D. Chen, H.-J. Jin, K.-S. Tan, D.-Y. Wang, Y. Yan, The origin, transmission and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak–an update on the status, Mil. Med. Res. 7 (2020) 1–10.

[10] M.T. Khan, A. Ali, Q. Wang, M. Irfan, A. Khan, M.T. Zeb, Y.-J. Zhang, S. Chinnasamy, D.-Q. Wei, Marine natural compounds as potents inhibitors against the main protease of SARS-CoV-2. A molecular dynamic study, J. Biomol. Struct. Dyn. 0 (2020) 1–14. https://doi.org/10.1080/07391102.2020.1769733.

[11] M.T. Khan, SARS-CoV-2 nucleocapsid and Nsp3 binding: an in silico study, Arch. Microbiol. (n.d.) 8.

[12] S. Elbe, G. Buckland-Merrett, Data, disease and diplomacy: GISAID's innovative contribution to global health, Glob. Chall. 1 (2017) 33–46. https://doi.org/10.1002/gch2.1018.

[13] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The Protein Data Bank, Nucleic Acids Res. 28 (2000) 235–242.

[14] A. Roy, A. Kucukural, Y. Zhang, I-TASSER: a unified platform for automated protein structure and function prediction., Nat. Protoc. 5 (2010) 725–38.

[15] Y. Zhang, I-TASSER server for protein 3D structure prediction, BMC Bioinformatics. 9 (2008) 40–40. https://doi.org/10.1186/1471-2105-9-40.

[16] C.H. Rodrigues, D.E. Pires, D.B. Ascher, DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability, Nucleic Acids Res. 46 (2018) W350–W355. https://doi.org/10.1093/nar/gky300.

[17] R. Lu, X. Zhao, J. Li, P. Niu, B. Yang, H. Wu, W. Wang, H. Song, B. Huang, N. Zhu, Y. Bi, X. Ma, F. Zhan, L. Wang, T. Hu, H. Zhou, Z. Hu, W. Zhou, L. Zhao, J. Chen, Y. Meng, J. Wang, Y. Lin, J. Yuan, Z. Xie, J. Ma, W.J. Liu, D. Wang, W. Xu, E.C. Holmes, G.F. Gao, G. Wu, W. Chen, W. Shi, W. Tan, Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding, The Lancet. 395 (2020) 565–574. https://doi.org/10.1016/S0140-6736(20)30251-8.

[18] J.F.-W. Chan, K.-H. Kok, Z. Zhu, H. Chu, K.K.-W. To, S. Yuan, K.-Y. Yuen, Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan, Emerg. Microbes Infect. 9 (2020) 221–236.

[19] Y. Shu, J. McCauley, GISAID: Global initiative on sharing all influenza data – from vision to reality, Eurosurveillance. 22 (2017). https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494.

[20] P. Zhou, X.-L. Yang, X.-G. Wang, B. Hu, L. Zhang, W. Zhang, H.-R. Si, Y. Zhu, B. Li, C.-L. Huang, H.-D. Chen, J. Chen, Y. Luo, H. Guo, R.-D. Jiang, M.-Q. Liu, Y. Chen, X.-R. Shen, X. Wang, X.-S. Zheng, K. Zhao, Q.-J. Chen, F. Deng, L.-L. Liu, B. Yan, F.-X. Zhan, Y.-Y. Wang, G.-F. Xiao, Z.-L. Shi, A pneumonia outbreak associated with a new coronavirus of probable bat origin, Nature. 579 (2020) 270–273. https://doi.org/10.1038/s41586-020-2012-7.

[21] S. Kumar, Drug and vaccine design against Novel Coronavirus (2019-nCoV) spike protein through Computational approach, Prepr. Www Prepr. OrgInternet. (2020).

[22] Y. Cai, J. Zhang, T. Xiao, H. Peng, S.M. Sterling, R.M.W. Jr, S. Rawson, S. Rits-Volloch, B. Chen, Distinct conformational states of SARS-CoV-2 spike protein, (2020) 8.

[23] D. Schoeman, B.C. Fielding, Coronavirus envelope protein: current knowledge, Virol. J. 16 (2019) 1–22.

[24] A. Sternberg, C. Naujokat, Structural features of coronavirus SARS-CoV-2 spike protein: Targets for vaccination, Life Sci. 257 (2020) 118056. https://doi.org/10.1016/j.lfs.2020.118056.

[25] Y. Wan, J. Shang, R. Graham, R.S. Baric, F. Li, Receptor recognition by the novel coronavirus from Wuhan: an analysis based on decade-long structural studies of SARS coronavirus, J. Virol. 94 (2020).

[26] J. Lan, J. Ge, J. Yu, S. Shan, H. Zhou, S. Fan, Q. Zhang, X. Shi, Q. Wang, L. Zhang, X. Wang, Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor, Nature. 581 (2020) 215–220. https://doi.org/10.1038/s41586-020-2180-5.

[27] Q. Wang, Y. Zhang, L. Wu, S. Niu, C. Song, Z. Zhang, G. Lu, C. Qiao, Y. Hu, K.-Y. Yuen, Q. Wang, H. Zhou, J. Yan, J. Qi, Structural and Functional Basis of SARS-CoV-2 Entry by Using Human ACE2, Cell. 181 (2020) 894-904.e9. https://doi.org/10.1016/j.cell.2020.03.045.

[28] B. Coutard, C. Valle, X. de Lamballerie, B. Canard, N.G. Seidah, E. Decroly, The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade, Antiviral Res. 176 (2020) 104742.

[29] Q. Li, J. Wu, J. Nie, L. Zhang, H. Hao, S. Liu, C. Zhao, Q. Zhang, H. Liu, L. Nie, H. Qin, M. Wang, Q. Lu, X. Li, Q. Sun, J. Liu, L. Zhang, X. Li, W. Huang, Y. Wang, The Impact of Mutations in SARS-CoV-2 Spike on Viral Infectivity and Antigenicity, Cell. 182 (2020) 1284-1294.e9. https://doi.org/10.1016/j.cell.2020.07.012.

[30] L. Zhang, C.B. Jackson, H. Mou, A. Ojha, H. Peng, B.D. Quinlan, E.S. Rangarajan, A. Pan, A. Vanderheiden, M.S. Suthar, W. Li, T. Izard, C. Rader, M. Farzan, H. Choe, SARS-CoV-2 spike-protein D614G mutation increases virion spike density and infectivity, Nat. Commun. 11 (2020) 6013. https://doi.org/10.1038/s41467-020-19808-4.

[31] J.A. Plante, Y. Liu, J. Liu, H. Xia, B.A. Johnson, K.G. Lokugamage, X. Zhang, A.E. Muruato, J. Zou, C.R. Fontes-Garfias, D. Mirchandani, D. Scharton, J.P. Bilello, Z. Ku, Z. An, B. Kalveram, A.N. Freiberg, V.D. Menachery, X. Xie, K.S. Plante, S.C. Weaver, P.-Y. Shi, Spike mutation D614G alters SARS-CoV-2 fitness, Nature. (2020). https://doi.org/10.1038/s41586-020-2895-3.

[32] B. Korber, W.M. Fischer, S. Gnanakaran, H. Yoon, J. Theiler, W. Abfalterer, N. Hengartner, E.E. Giorgi, T. Bhattacharya, B. Foley, et al., Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus, Cell. 182 (2020) 812-827.e19. https://doi.org/10.1016/j.cell.2020.06.043.

[33] E.B. Hodcroft, M. Zuber, S. Nadeau, K.H.D. Crawford, J.D. Bloom, D. Veesler, T.G. Vaughan, I. Comas, F.G. Candelas, T. Stadler, R.A. Neher, Emergence and spread of a SARS-CoV-2 variant through Europe in the summer of 2020, MedRxiv. (2020). https://doi.org/10.1101/2020.10.25.20219063.

[34] S. Fiorentini, S. Messali, A. Zani, F. Caccuri, M. Giovanetti, M. Ciccozzi, A. Caruso, First detection of SARS-CoV-2 spike protein N501 mutation in Italy in August, 2020, Lancet Infect. Dis. 0 (2021).

https://doi.org/10.1016/S1473-3099(21)00007-4.

[35] P.K. Singh, U. Kulsum, S.B. Rufai, S.R. Mudliar, S. Singh, Mutations in SARS-CoV-2 Leading to Antigenic Variations in Spike Protein: A Challenge in Vaccine Development, J. Lab. Physicians. 12 (2020) 154–160. https://doi.org/10.1055/s-0040-1715790.

[36] B. Durmaz, O. Abdulmajed, R. Durmaz, Mutations Observed in the SARS-CoV-2 Spike Glycoprotein and Their Effects in the Interaction of Virus with ACE-2 Receptor, Medeni. Med. J. (2020). https://doi.org/10.5222/MMJ.2020.98048.

[37] M.A. Tortorici, M. Beltramello, F.A. Lempp, D. Pinto, H.V. Dang, L.E. Rosen, M. McCallum, J. Bowen, A. Minola, S. Jaconi, F. Zatta, A.D. Marco, B. Guarino, S. Bianchi, E.J. Lauron, H. Tucker, J. Zhou, A. Peter, C. Havenar-Daughton, J.A. Wojcechowskyj, J.B. Case, R.E. Chen, H. Kaiser, M. Montiel-Ruiz, M. Meury, N. Czudnochowski, R. Spreafico, J. Dillen, C. Ng, N. Sprugasci, K. Culap, F. Benigni, R. Abdelnabi, S.-Y.C. Foo, M.A. Schmid, E. Cameroni, A. Riva, A. Gabrieli, M. Galli, M.S. Pizzuto, J. Neyts, M.S. Diamond, H.W. Virgin, G. Snell, D. Corti, K. Fink, D. Veesler, Ultrapotent human antibodies protect against SARS-CoV-2 challenge via multiple mechanisms, Science. 370 (2020) 950–957. https://doi.org/10.1126/science.abe3354.

[38] G.B. Chand, A. Banerjee, G.K. Azad, Identification of twenty-five mutations in surface glycoprotein (Spike) of SARS-CoV-2 among Indian isolates and their impact on protein dynamics, Gene Rep. 21 (2020) 100891. https://doi.org/10.1016/j.genrep.2020.100891.

[39] P. Pradhan, A.K. Pandey, A. Mishra, P. Gupta, P.K. Tripathi, M.B. Menon, J. Gomes, P. Vivekanandan, B. Kundu, Uncanny similarity of unique inserts in the 2019-nCoV spike protein to HIV-1 gp120 and Gag, Evolutionary Biology, 2020. https://doi.org/10.1101/2020.01.30.927871.

[40] D.X. Liu, Q. Yuan, Y. Liao, Coronavirus envelope protein: A small membrane protein with multiple functions, Cell. Mol. Life Sci. 64 (2007) 2043–2048. https://doi.org/10.1007/s00018-007-7103-1.

[41] T.S. Fung, D.X. Liu, Post-translational modifications of coronavirus proteins: roles and function, Future Virol. 13 (2018) 405–430. https://doi.org/10.2217/fvl-2018-0008.

[42] T.R. Ruch, C.E. Machamer, The Hydrophobic Domain of Infectious Bronchitis Virus E Protein Alters the Host Secretory Pathway and Is Important for Release of Infectious Virus, J. Virol. 85 (2011) 675–685. https://doi.org/10.1128/JVI.01570-10.

[43] C. Verdiá-Báguena, J.L. Nieto-Torres, A. Alcaraz, M.L. DeDiego, L. Enjuanes, V.M. Aguilella, Analysis of SARS-CoV E protein ion channel activity by tuning the protein and lipid charge, Biochim. Biophys. Acta BBA - Biomembr. 1828 (2013) 2026–2031. https://doi.org/10.1016/j.bbamem.2013.05.008.

[44] W. Surya, Y. Li, J. Torres, Structural model of the SARS coronavirus E channel in LMPG micelles, Biochim. Biophys. Acta BBA - Biomembr. 1860 (2018) 1309–1317. https://doi.org/10.1016/j.bbamem.2018.02.017.

[45] M.K. Gupta, S. Vemula, R. Donde, G. Gouda, L. Behera, R. Vadde,*In-silico* approaches to detect inhibitors of the human severe acute respiratory syndrome coronavirus envelope protein ion channel, J. Biomol. Struct. Dyn. (2020) 1–11. https://doi.org/10.1080/07391102.2020.1751300.

[46] M. Bianchi, D. Benvenuto, M. Giovanetti, S. Angeletti, M. Ciccozzi, S. Pascarella, Sars-CoV-2 Envelope and Membrane Proteins: Structural Differences Linked to Virus Characteristics?, BioMed Res. Int. 2020 (2020) 1–6. https://doi.org/10.1155/2020/4389089.

[47] M.S. Rahman, M.N. Hoque, M.R. Islam, I. Islam, I.D. Mishu, Md.M. Rahaman, M. Sultana, M.A. Hossain, Mutational insights into the envelope protein of SARS-CoV-2, Gene Rep. 22 (2021) 100997. https://doi.org/10.1016/j.genrep.2020.100997.

[48] Y. Ye, B.G. Hogue, Role of the Coronavirus E Viroporin Protein Transmembrane Domain in Virus Assembly, J. Virol. 81 (2007) 3597–3607. https://doi.org/10.1128/JVI.01472-06.

[49] E.A. J Alsaadi, I.M. Jones, Membrane binding proteins of coronaviruses, Future Virol. 14 (2019) 275–286. https://doi.org/10.2217/fvl-2018-0144.

[50] A.L. Arndt, B.J. Larson, B.G. Hogue, A Conserved Domain in the Coronavirus Membrane Protein Tail Is Important for Virus Assembly, J. Virol. 84 (2010) 11418–11428. https://doi.org/10.1128/JVI.01131-10.

[51] S. Thomas, The Structure of the Membrane Protein of SARS-CoV-2 Resembles the Sugar Transporter SemiSWEET, Pathog. Immun. 5 (2020) 342. https://doi.org/10.20411/pai.v5i1.377.

[52] Y.-T. Tseng, S.-M. Wang, K.-J. Huang, A.I.-R. Lee, C.-C. Chiang, C.-T. Wang, Self-assembly of Severe Acute Respiratory Syndrome Coronavirus Membrane Protein, J. Biol. Chem. 285 (2010) 12862–12872. https://doi.org/10.1074/jbc.M109.030270.

[53] X. Fang, J. Gao, H. Zheng, B. Li, L. Kong, Y. Zhang, W. Wang, Y. Zeng, L. Ye, The membrane protein of SARS-CoV suppresses NF-κB activation, J. Med. Virol. 79 (2007) 1431–1439. https://doi.org/10.1002/jmv.20953.

[54] Y. Wang, L. Liu, The Membrane Protein of Severe Acute Respiratory Syndrome Coronavirus Functions as a Novel Cytosolic Pathogen-Associated Molecular Pattern To Promote Beta Interferon Induction via a Toll-Like-Receptor-Related TRAF3-Independent Mechanism, MBio. 7 (2016) e01872-15, /mbio/7/1/e01872-15.atom. https://doi.org/10.1128/mBio.01872-15.

[55] R. Arya, S. Kumari, B. Pandey, H. Mistry, S.C. Bihani, A. Das, V. Prashar, G.D. Gupta, L. Panicker, M. Kumar, Structural insights into SARS-CoV-2 proteins, J. Mol. Biol. 433 (2021) 166725. https://doi.org/10.1016/j.jmb.2020.11.024.

[56] N.K. Dutta, K. Mazumdar, J.T. Gordy, The Nucleocapsid Protein of SARS–CoV-2: a Target for Vaccine Development, 94 (2020) 2.

[57] W. Zeng, Biochemical characterization of SARS-CoV-2 nucleocapsid protein, Biochem. Biophys. Res. Commun. (2020) 6.

[58] L.E. Gralinski, V.D. Menachery, Return of the Coronavirus: 2019-nCoV, Viruses. 12 (2020) 135.

[59] P. V'kovski, M. Gerber, J. Kelly, S. Pfaender, N. Ebert, S. Braga Lagache, C. Simillion, J. Portmann, H. Stalder, V. Gaschen, R. Bruggmann, M.H. Stoffel, M. Heller, R. Dijkman, V. Thiel, Determination of host proteins composing the microenvironment of coronavirus replicase complexes by proximity-labeling, ELife. 8 (2019) e42037. https://doi.org/10.7554/eLife.42037.

[60] S. Kang, M. Yang, Z. Hong, L. Zhang, Z. Huang, X. Chen, S. He, Z. Zhou, Z. Zhou, Q. Chen, Y. Yan, C. Zhang, H. Shan, S. Chen, Crystal structure of SARS-CoV-2 nucleocapsid protein RNA binding domain reveals potential unique drug targeting sites, Acta Pharm. Sin. B. 10 (2020) 1228–1238. https://doi.org/10.1016/j.apsb.2020.04.009.

[61] C. Chang, S.-C. Sue, T. Yu, C.-M. Hsieh, C.-K. Tsai, Y.-C. Chiang, S. Lee, H. Hsiao, W.-J. Wu, W.-L. Chang, C.-H. Lin, T. Huang, Modular organization of SARS coronavirus nucleocapsid protein, J. Biomed. Sci. 13 (2006) 59–72. https://doi.org/10.1007/s11373-005-9035-9.

[62] C.-K. Chang, Y.-L. Hsu, Y.-H. Chang, F.-A. Chao, M.-C. Wu, Y.-S. Huang, C.-K. Hu, T.-H. Huang, Multiple Nucleic Acid Binding Sites and Intrinsic Disorder of Severe Acute Respiratory Syndrome Coronavirus Nucleocapsid Protein: Implications for Ribonucleocapsid Protein Packaging, J. Virol. 83 (2009) 2255–2264. https://doi.org/10.1128/JVI.02001-08.

[63] H. Luo, F. Ye, K. Chen, X. Shen, H. Jiang, SR-Rich Motif Plays a Pivotal Role in Recombinant SARS Coronavirus Nucleocapsid Protein Multimerization, (n.d.) 8.

[64] C.-Y. Chen, C. Chang, Y.-W. Chang, S.-C. Sue, H.-I. Bai, L. Riang, C.-D. Hsiao, T. Huang, Structure of the SARS Coronavirus Nucleocapsid Protein RNA-binding Dimerization Domain Sug-

gests a Mechanism for Helical Packaging of Viral RNA, J. Mol. Biol. 368 (2007) 1075–1086. https://doi.org/10.1016/j.jmb.2007.02.069.

[65] A. Savastano, A. Ibáñez de Opakua, M. Rankovic, M. Zweckstetter, Nucleocapsid protein of SARS-CoV-2 phase separates into RNA-rich polymerase-containing condensates, Nat. Commun. 11 (2020) 6041. https://doi.org/10.1038/s41467-020-19843-1.

[66] M.S. Rahman, M.R. Islam, A.S.M.R.U. Alam, I. Islam, M.N. Hoque, S. Akter, Md.M. Rahaman, M. Sultana, M.A. Hossain, Evolutionary dynamics of SARS-CoV-2 nucleocapsid protein and its consequences, J. Med. Virol. (2020) jmv.26626. https://doi.org/10.1002/jmv.26626.

[67] H.Y.L. Tung, P. Limtung, Mutations in the phosphorylation sites of SARS-CoV-2 encoded nucleocapsid protein and structure model of sequestration by protein 14-3-3, Biochem. Biophys. Res. Commun. 532 (2020) 134–138. https://doi.org/10.1016/j.bbrc.2020.08.024.

[68] G.K. Azad, Identification and molecular characterization of mutations in nucleocapsid phosphoprotein of SARS-CoV-2, PeerJ. 9 (2021) e10666. https://doi.org/10.7717/peerj.10666.

Table 1. Number of mutations in structural protein of CoV-2

Structural Proteins	**No. of mutations**	**Prevalent Countries**
Spike protein	4725	48
Envelope protein	259	34
Membrane protein	627	40
Nucleocapsid protein	1631	37

Table 2. Residues interacting at the CoV-2 RBD-ACE2 interface and mutations frequencies.

S. No	**ACE2**	**CoV-2 RBD**	**Mutations with frequency**
1	D30	K417	K417N *(218), K417R (2), K417P (1)
2	T27	G446	G446V (58), G446S (8), G446D (2), G446A (2)
3	E28	Y449	Y449N (1), Y449F (1)
4	Q24	Y453	Y453F (968)
5	K31	L455	L455F (45), L455W (1)
6	H34	F456	F456L (34)
7	E35	A475	A475V (37), A475S (1)
8	E37	F486	F486L (34), F486G (1), F486I (1)
9	D38	N487	N487D (1), N487I (3), N487K (1), N487K (1)
10	Y41	Y489	Y489V (1), Y489H (2)
11	Q42	Q493	Q493L (12), Q493K (3), Q493H (1), Q493R (1), Q493D (1)
12	L79	G496	G496C (1), G496R (1)
13	M82	Q498	Q498P (1), Q498H (2), Q498I (1)
14	Y83	T500	T500I (2), T500P (2), T500S (1)
15	N330	N501	N501Y (4362), N501T (48), N501S (8), N501R (1), N501I (1), N501G (1)
16	K353	G502	G502R (1), G502C (1), G502Q (1), G502D (1)
17	G354	Y505	Y505W (19), Y505E (1), Y505K (1)
18	D355		
19	R357		
20	R393		

*Number of genomes/frequencies, #: Mutations

Table 3. Frequency of mutation in S protein of CoV-2

Accession	Wild type AA	Position	Mutated AA	Frequency	Mutation
EPI_ISL_-732828	L	5	F	3813	L5F
EPI_ISL_-705553	L	18	F	28015	L18F
EPI_ISL_-627142	R	21	I	1741	R21I
EPI_ISL_-430632	H	49	Y	576	H49Y
EPI_ISL_-472372	L	54	F	998	L54F
EPI_ISL_-679344	D	80	Y	1530	D80Y
EPI_ISL_-732880	S	98	F	2601	S98F
EPI_ISL_-706875	H	146	Y	531	H146Y
EPI_ISL_-708736	M	153	T	1673	M153T
EPI_ISL_-679344	N	164	T	584	N164T
EPI_ISL_-650411	L	176	F	1017	L176F
EPI_ISL_-593832	D	215	H	918	D215H
EPI_ISL_-705553	A	222	V	59697	A222V
EPI_ISL_-696019	D	253	G	874	D253G
EPI_ISL_-706880	A	262	S	3092	A262S
EPI_ISL_-706880	P	272	L	2213	P272L
EPI_ISL_-732841	N	439	K	5725	N439K
EPI_ISL_-523390	Y	453	F	968	Y453F
EPI_ISL_-640800	S	477	N	16914	S477N
EPI_ISL_-519331	N	501	Y	4362	N501Y
EPI_ISL_-727754	A	570	D	3676	A570D
EPI_ISL_-705553	D	614	G	266513	D614G
EPI_ISL_-732827	A	626	S	881	A626S
EPI_ISL_-448554	H	655	Y	1131	H655Y

Accession	Wild type AA	Position	Mutated AA	Frequency	Mutation
EPI_ISL_-602412	Q	677	H	888	Q677H
EPI_ISL_-727754	P	681	H	4166	P681H
EPI_ISL_-627066	A	688	V	1053	A688V
EPI_ISL_-727754	T	716	I	3770	T716I
EPI_ISL_-649929	T	723	I	909	T723I
EPI_ISL_-433973	G	769	V	624	G769V
EPI_ISL_-452908	D	936	Y	1239	D936Y
EPI_ISL_-704615	S	939	F	555	S939F
EPI_ISL_-727754	S	982	A	3670	S982A
EPI_ISL_-732841	K	1073	N	1055	K1073N
EPI_ISL_-727754	D	1118	H	3676	D1118H
EPI_ISL_-660552	D	1163	Y	1903	D1163Y
EPI_ISL_-660552	G	1167	V	1753	G1167V
EPI_ISL_-584645	V	1176	F	933	V1176F
EPI_ISL_-457026	P	1263	L	911	P1263L

Table 4. Frequency of some common mutation in E protein of CoV-2.

Accession	Wild type AA	Position	Mutated AA	Frequency	Mutation
EPI_ISL_476911	T	9	I	168	T9I
EPI_ISL_526912	F	20	L	90	F20L
EPI_ISL_542279	L	21	F	84	L21F
EPI_ISL_604556	V	24	M	76	V24M
EPI_ISL_416354	T	30	I	72	T30I
EPI_ISL_636718	I	46	V	66	I46V
EPI_ISL_478733	V	49	L	93	V49L
EPI_ISL_424214	S	55	F	128	S55F
EPI_ISL_538676	V	62	F	129	V62F
EPI_ISL_448073	S	68	F	419	S68F
EPI_ISL_452908	R	69	I	159	R69I
EPI_ISL_577907	P	71	L	158	P71L
EPI_ISL_660339	P	71	S	264	P71S
EPI_ISL_414686	D	72	Y	76	D72Y

Accession	Wild type AA	Position	Mutated AA	Frequency	Mutation
EPI_ISL_478788	L	73	F	218	L73F

Table 5. The M protein common mutations and their effect on structure.

Mutation	Σταβιλιτψ: ΔΔΓ: (κςαλ/μολ)	Φλεξιβιλιτψ: ΔΔΣϊβ ΕΝδΜ (κςαλ.μολ^{-1}Κ^{-1})
A2V	-0.015 (Destabilizing)	0.129 (Increase of molecule flexibility)
A2S	-0.091 (Destabilizing)	0.080 (Increase of molecule flexibility)
D3G	-0.975 (Destabilizing)	0.464 (Increase of molecule flexibility)
L17I	0.468 (Stabilizing)	-0.022 (Decrease of molecule flexibility)
H125S	-1.279 (Destabilizing)	0.582 (Increase of molecule flexibility)
T175M	2.248 (Stabilizing)	-0.780 (Decrease of molecule flexibility)
D209Y	1.702 (Stabilizing)	-0.674 (Decrease of molecule flexibility)

Table 6. Frequency of mutation in M protein of CoV-2

Accession	Wild type AA	Position	Mutated AA	Frequency	Mutation
EPI_ISL_454390	A	2	V	463	A2V
EPI_ISL_663841	A	2	S	460	A2S
EPI_ISL_523400	D	3	G	968	D3G
EPI_ISL_584320	S	4	F	68	S4F
EPI_ISL_542207	T	7	I	64	T7I
EPI_ISL_575047	K	15	N	63	K15N
EPI_ISL_663873	L	17	F	162	L17F
EPI_ISL_650389	L	17	I	621	L17I
EPI_ISL_464932	V	23	L	165	V23L
EPI_ISL_490511	L	29	F	78	L29F
EPI_ISL_433327	W	31	C	71	W31C
EPI_ISL_417242	L	34	F	55	L34F
EPI_ISL_461174	L	46	F	53	L46F
EPI_ISL_672005	I	48	V	81	I48V
EPI_ISL_413017	I	52	T	56	I52T
EPI_ISL_596718	V	60	L	221	V60L
EPI_ISL_671336	A	69	S	109	A69S
EPI_ISL_483066	V	70	F	220	V70F
EPI_ISL_591472	V	70	I	54	V70I
EPI_ISL_425553	W	75	L	50	W75L
EPI_ISL_532006	I	80	F	89	I80F
EPI_ISL_483441	A	83	S	60	A83S
EPI_ISL_542158	C	86	F	235	C86F
EPI_ISL_460103	M	109	I	103	M109I
EPI_ISL_649900	H	125	Y	395	H125Y
EPI_ISL_604474	A	142	V	65	A142V
EPI_ISL_465103	H	155	Y	114	H155Y
EPI_ISL_422292	K	162	N	52	K162N
EPI_ISL_479375	I	168	L	142	I168L
EPI_ISL_579281	T	175	M	1286	T175M
EPI_ISL_464659	T	208	I	54	T208I

Accession	Wild type AA	Position	Mutated AA	Frequency	Mutation
EPI_ISL_419249	D	209	Y	386	D209Y

Table 7. Frequency of mutation in Nucleocapsid protein of CoV-2

Accession	Wild type AA	Position	Mutated AA	Frequency	Mutation
EPI_ISL_440111	D	3	L	532	D3L
EPI_ISL_605792	Q	9	H	861	Q9H
EPI_ISL_480694	P	13	L	2611	P13L
EPI_ISL_627662	P	13	T	548	P13T
EPI_ISL_579246	S	33	I	590	S33I
EPI_ISL_421434	P	67	S	1860	P67S
EPI_ISL_461522	D	103	Y	2233	D103Y
EPI_ISL_671336	H	145	Y	780	H145Y
EPI_ISL_579513	S	183	Y	1401	S183Y
EPI_ISL_577852	S	186	Y	549	S186Y
EPI_ISL_561334	S	187	L	561	S187L
EPI_ISL_461513	S	188	L	1029	S188L
EPI_ISL_490497	S	190	I	913	S190I
EPI_ISL_538513	S	193	I	1024	S193I
EPI_ISL_490428	S	194	L	11915	S194L
EPI_ISL_547472	S	197	L	2165	S197L
EPI_ISL_577955	P	199	L	3483	P199L
EPI_ISL_664658	S	202	N	756	S202N
EPI_ISL_640800	R	203	K	82570	R203K
EPI_ISL_640800	G	204	R	81858	G204R
EPI_ISL_456596	T	205	I	1466	T205I
EPI_ISL_626926	M	210	I	575	M210I
EPI_ISL_665951	A	220	V	39729	A220V
EPI_ISL_593983	M	234	I	5032	M234I
EPI_ISL_574824	S	235	F	1081	S235F
EPI_ISL_548689	I	292	T	711	I292T
EPI_ISL_637214	P	365	S	2369	P365S
EPI_ISL_593983	A	376	T	4183	A376T
EPI_ISL_490431	D	377	Y	2363	D377Y
EPI_ISL_577921	P	383	L	534	P383L
EPI_ISL_593983	R	385	K	1254	R385K
EPI_ISL_664987	A	398	V	3552	A398V

Figure legends

Figure 1. SARS CoV-2 genome organization and mutations in structural proteins. Frequency of mutations has been shown.

Figure 2. Stability prediction of N501Y. The mutant (Y501) seems more stable than wild type residues (N501). The Y501 has been detected in 4362 genome sequences, prevalent in 48 countries. The prediction was performed through DynaMut server[16].

Figure 3. Frequency of mutations in SARS-CoV-2 spike proteins. (A) SP: signal peptide, NTD: N-terminal domain, RBD: receptor-binding domain, RBM: receptor binding motif, FP: fusion peptide, HR:

heptapeptide repeat sequence, TM: transmembrane region, CT: cytoplasmic tail. S1 (14–685aa) includes, NTD (14–305aa), RBD (319–541aa), RBM (aa437–508). S2 (686–1273aa) include FP (788–806aa), HR1 (912–984aa), HR2 (1163–1213aa), TM (1214–1237aa) and CP (1238–1273aa) S2 subunit. **(B)** Domains with mutations. RBD: ACE binding residues and mutations.

Figure 4. Domain organization of E proteins and the location and frequency of most common mutations. TM: transmembrane, NTD: N-terminal region, CTD: C-terminal region. (B). Pentameric representation of E proteins and common mutation T9. (C). Full length E protein most common mutations in the loop region.

Figure 5. D omains organization of M protein along with most common mutations and their frequencies. (A). NTD: N-terminal region, TM: transmembrane helices, CTD: C-terminal region. (B). Mutations most commonly occurs at NTD and CTD.

Figure 6. Location and frequencies of some frequently occurring mutation in N protein. (A). NTD: N-terminal domain (43-174aa), SR-Linker: serine and arginine rich region (194-225aa), CTD: C-terminal domain (257-364aa), RBD: RNA binding domain (40-180aa), PS: phosphorylation sites (186, 197, 202, 204 aa). (B). Crystal structure of NTD and most common variants.

Graphical abstract