Whole Genome Sequencing for Revealing the Point Mutations of SARS-CoV-2 Genome in Bangladeshi Isolates and their Structural Effects on Viral Proteins

Mohammad Uzzal Hossain[1], Ishtiaque Ahammad[1], Md. Tabassum Hossain Emon[2], Arittra Bhattacharjee[1,3], Zeshan Mahmud Chowdhury[3], Md. Golam Mosaib[4], Keshob Chandra Das[5], Chaman Ara Keya[3], Md. Salimullah[5*]

[1]Bioinformatics Division, [5]Molecular Biotechnology Division, National Institute of Biotechnology, Ganakbari, Ashulia, Savar, Dhaka-1349, Bangladesh

[2]Department of Biotechnology and Genetic Engineering, Life Science Faculty, Mawlana Bhashani Science and Technology University, Santosh, Tangail-1902, Bangladesh.

[3]Department of Biochemistry and Microbiology, North South University, Bashundhara, Dhaka-1229, Bangladesh

[4]Department of Biochemistry and Molecular Biology, Faculty of Health & Medical Sciences, Gono Bishwabidyaloy,Ashulia, Savar, Dhaka-1344, Bangladesh

***Correspondence:**

Dr. Md. Salimullah

Chief Scientific Officer

Molecular Biotechnology Division

National Institute of Biotechnology

Ganakbari, Ashulia, Savar, Dhaka-1349, Bangladesh

Tel: 880-2-7788443

E-mail: salim2969@gmail.com

Abstract

Coronavirus disease-19 (COVID-19) is the recent global pandemic caused by the virus Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2). The virus has already killed more than one million people worldwide and billions are at risk of getting infected. As of now, there is neither any drug nor any vaccine in sight with conclusive scientific evidence that it can cure or provide protection against the illness. Since novel coronavirus is a new virus, mining its genome sequence is of crucial importance for drug/vaccine(s) development. Whole genome sequencing is a helpful tool in identifying genetic changes that occur in a virus when it spreads through the population. In this study, we performed complete genome sequencing of SARS-CoV-2 to unveil the genomic variation and indel, if present. We discovered thirteen (13) mutations in Orf1ab, S and N gene where seven (7) of them turned out to be novel mutations from our sequenced isolate. Besides, we found one (1) insertion and seven (7) deletions from the indel analysis among the 323 Bangladeshi isolates. However, the indel did not show any effect on proteins. Our energy minimization analysis showed both stabilizing and destabilizing impact on viral proteins depending on the mutation. Interestingly, all the variants were located in the binding site of the proteins. Furthermore, drug binding analysis revealed marked difference in interacting residues in mutants when compared to the wild type. Our analysis also suggested that eleven (11) mutations could exert damaging effects on their corresponding protein structures. The analysis of SARS-CoV-2 genetic variation and their impacts presented in this study might be helpful in gaining a better understanding of the pathogenesis of this deadly virus.

Keywords

SARS-CoV-2; Genome sequencing; Mutations; Insertions; Deletions; Drug Binding

Introduction

COVID-19 can be currently considered an menace to humankind brought about by the novel Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2) which began its journey from the Wuhan province of the People's Republic of China [1, 2]. The infection basically targets the respiratory framework of its host causing influenza-like sickness with symptoms such as cough, fever, and in progressively serious cases, troubled breathing [3]. According to the data available, mortality is higher in individuals of advanced ages (> 60 years) and the ones with comorbidities. Apart from intense respiratory problems, COVID-19 has been shown to cause systemic irritation prompting sepsis, intense cardiovascular injury, cardiovascular breakdown and multiorgan failure in critical patients [4]. COVID-19 has been rightly announced as a global pandemic by the World Health Organization (WHO) as it has spanned over 200 countries and territories around the world. As of Oct 20, 2020, more than 40,648,527 individuals have been infected up until this point, with around 9,172,183 active cases and 31,476,344 closed cases and 1,122,992 deaths (https://www.worldometers.info/coronavirus/). While the USA has the highest number of confirmed cases (with over 8,456,653 so far) and has more than 225,222 deaths to date. China, the nation where everything started in November, is presently on the downslope sustaining just above 200 active cases. The vast majority of the countries around the world were in lockdown to stop human to human transmission and halt the spread of the disease.

Coronaviruses (CoVs) are enveloped, single-stranded, positive RNA viruses that are pathogenic to their hosts [6]. SARS-CoV-2 is the causative agent behinds CODIV-19 and is more pathogenic in contrast with previously observed SARS-CoV (2002) and Middle East respiratory syndrome coronavirus (MERS-CoV, 2013). There is a dire need to examine the virus more comprehensively to comprehend the system of pathogenesis, its destructiveness to develop powerful therapeutic measures [7]. CoVs belongs to the Coronaviridae family under Nidovirales order. They have been grouped into four genera that belong to α-, β-, γ-, and δ-coronaviruses [8]. Among them, α-and β-COVs infect vertebrates, γ-coronaviruses avians, while the δ-coronaviruses infect both. SARS-CoV, mouse hepatitis coronavirus (MHV), MERS-CoV, Bovine coronavirus (BCoV), bat coronavirus HKU4, and human coronavirus OC43, including SARS-CoV-2, are β-coronaviruses [9]. Zoonotic transmission is the medium of transmission for each of the three CoVs, SARS-,

MERS-, and SARS-CoV-2, and they spread through close contact. The essential multiplication number (R0) of the individual-to-individual spread of SARS-CoV-2 is around 2.6, which implies that the confirmed cases develop at a striking exponential rate [10]. CoVs being 26 to 32 kb long have the biggest RNA viral genome [11]. The SARS-CoV-2 genome share approximately 90% identity with essential enzymes and structural proteins of SARS-CoV. Fundamentally, SARS-CoV-2 contains four basic proteins known as- spike (S), envelope (E), membrane (M), and nucleocapsid (N) proteins. These proteins share high sequence similarity with the sequence of the corresponding proteins in SARS– CoV, and MERS-CoV. Hence, it is vital to scrutinize the SARS-CoV-2 genome to determine why this infection is progressively inclined to be more infectious and lethal than its predecessors.

Utilizing Sanger sequencing and cutting-edge whole genome sequencing of SARS-CoV-2 isolates from oropharyngeal samples, we depicted the genomic portraits of two genomes alongside other Bangladeshi strains [reference].

In this study, we have analyzed the genomic arrangements of SARS-CoV-2 to identify the mutations found within the genomes and anticipate their effect on the protein structure from a structural biology perspective in order to shed light on the suitable therapeutics against this deadly virus.

Methods

Virus Isolation

The patient's oropharyngeal samples SARS-CoV-2/human/BGD/NIB_01/2020 and SARS-CoV-2/human/BGD/NIB-BCSIR_02/2020 were collected using the UTM™ kit containing 1 mL of viral transport media (Copan Diagnostics Inc., Murrieta, CA, USA) on day 7 of the patient's illness with symptoms of cough, mild fever, and throat congestion. The specimen was tested positive for SARS-CoV-2 by real-time reverse transcriptase PCR (rRT-PCR). Then, the viral RNA was extracted directly from the patient's swab using PureLink Viral RNA/DNA Mini Kit (Invitrogen). The viral RNA was then converted into cDNA using SuperScript™ VILO™ cDNA Synthesis Kit (Invitrogen) according to the manufacturer's instructions.

DNA Sequencing

Sanger Dideoxy sequencing

The forty eight (48) pair primers were designed to cover the whole genome of the virus by following two conditions: (1) their sequence is conserved among all the available SARS-CoV-2 isolates and (2) the terminal of the amplicons will overlap with neighboring amplicons. The polymerase chain reaction (PCR) was performed and the 48 primers then generated 47 amplicons which were visualized in 1.5% Agarose gel electrophoresis. The amplicons were further purified using Purelink PCR purification kit (ThermoFisher Scientific, USA). These purified amplicons were sequenced using Sanger dideoxy method by "ABI 3500" with BigDye Terminator version 3.1 cycle sequencing kit (Applied Biosystems, USA). The raw reads were assembled by DNA Baser (https://www.dnabaser.com) [12] and verified by SeqMan Pro®. Version 14.1. DNASTAR [13]. Madison, WI. These overlapping regions were visualized by CLC Genomics Workbench 20.0.4 (https://digitalinsights.qiagen.com) and merged with EMBOSS: merger (https://www.bioinformatics.nl/cgi-bin/emboss/merger).

NGS Sequencing

Illumina Nextseq 550 next-generation sequencing technology was used to sequence the complete genome of the SARS-CoV-2/human/BGD/NIB-BCSIR_02/2020 isolate to where Nextera DNA Flex was utilized as library preparation kit for the synthesis of the nucleotides [14]. To cover the 300 cycle, the NextSeq High Output Kit was utilized as the reagent cartridge. To generate the FASTQ data workflow the run mode was set as local run manager in every NextSeq 4-channel chemistry. Analysis and Quality check was performed using a customized version of the DRAGEN RNA pipeline, which was also available on local DRAGEN server hardware. The Illumina® DRAGEN RNA Pathogen Detection App uses a combined human and virus reference to analyze pathogen data. The raw reads were cleaned by trimming low-quality bases with Trimmomatic 0.36 (-phred33, LEADING:20, TRAILING:20, SLIDCitation. The assembly was performed by the utilization of SPAdes using default parameters as well as used to cross-validate with the reference-based method as an internal control.

Variant Identification

Basic Local Alignment Search Tools (BLAST) [15] was employed to identify possible mutations in Sanger Based sequenced nucleotide sequences. Nucleotide program of blast was selected for

this identification. The mapped polymorphisms were investigated for their occurrence frequency worldwide and checked for their profile at CNCB[3] resource. Chimera was utilized to visualize the mapped polymorphisms. Besides, all the available Bangladeshi strains (n=323) of SARS-CoV-2 were retrieved from GISAID [16] and further explored to find out the most common mutations.

Indel Analysis

We ran BLAST for all Bangladeshi SARS CoV-2 genome profiles, to get a view on the pairwise alignment comparison. Deleted sequences were closely investigated in Artemis Comparison Tool window [17]. Gene regions are observed discreetly.

Mutational Effect Analysis

To observe the mutational effect of mapped polymorphisms, 3 dimensional (3D) structure was built through the ROBETTA prediction server [18]. Later, the difference of the energy was calculated by GROMACS in both wild type and mutant 3D structures to estimate the structural abnormality and change in stability [19]. Binding site of both wild and mutant structure was analyzed to check whether the amino acid residues are into the binding site region or not.

Drug-Binding Analysis

We have performed Molecular Docking using Autodock vina [20] for the analysis of interacting residues to the druggable targets. We retrieved the structures of all the interacting drugs (.pdb files) by virtual screening of the Drugbank database. Then we generated the .pdbqt files of the targets of polymorphisms for docking experiments. Blind docking was performed for the identification of the most effective binding site of these drugs. A grid box parameter for covering the whole protein was set for all docking runs.

Results

Revealing the SARS-CoV-2 Genome

The workflow of this manuscript has been shown in **Fig 1.**

In case of NIB-01 isolates forty eight (48) contigs with ninety four (94) overlapping regions were obtained. The sequence coverage was 2X the raw reads. It was then assembled by SeqMan Pro and EMBOSS merger [21]. The assembled viral genome consisted of a single stranded positive

(+) RNA that is 29,724 nucleotides long with: 8,882 Adenosines (29.88%), 5,455 Cytosine (18.35%), 5,836 Guanine (19.63%), and 9,551 Thymine (32.13%). The GC content of the whole genome was 38%. A total of 1,78,22,898 reads were produced in the reference-based alignment after trimming 99% of them were mapped to the SARS-CoV-2 reference genome.

The complete nucleotide sequence of SARS-CoV-2 isolate SARS-CoV-2/human/BGD/NIB_01/2020 from the Sanger sequencing has been deposited in GenBank under the accession number MT509958 (https://www.ncbi.nlm.nih.gov/nuccore/MT509958).The complete nucleotide sequence of SARS-CoV-2/human/BGD/NIB-BCSIR_02/2020 isolate from the NGS has been also deposited under the accession number MT568643 (https://www.ncbi.nlm.nih.gov/nuccore/MT568643.1?report=genbank).

Exploration of variant in the sequenced isolates

We have found thirteen (13) mutations at the 93rd; C → T, 479th; T → A, 481th; C→A; 1015th; A → T, 2889th; C → T, 5098th; G → T; 5237th; C → T, 5642th; G → T, 8023rd; G → A, 23255th; A → G 28733rd; G → A, 28734th; G → A and 28735th; G → C from the MT509958 whole genome (**Table 1**). However, the MT568643 whole genome showed no mutation against the reference sequence. From them, six (6) mutations namely 93rd; C → T, 2889th; C → T, 23255th; A→G, 28733rd; G → A, 28734th; G → A and 28735th; G → C were found in CNCB[3] resource where the available mutations of SARS-CoV-2 were enlisted (**Table 1**). These mutations have already been found in different countries where the SARS-Cov-2 has been sequenced. These mutations were mostly found in the United States of America (USA) and the United Kingdom (UK) (**Table 1**). The position of 93rd; C → T mutation is located in 5'UTR upstream region. And the position 2889th; C → T mutation has shown no change to its protein sequence. The other 23255th; A→G, 28733rd; G → A, 28734th; G → A and 28735th; G → C mutations can alter the amino acid sequence and can have the missense effect on the protein (**Table 1**). Apart from these 6 mutations, seven (7) mutations have shown as unique variants against the reference sequence (**Table 1 and Fig 2**). There is no report previously having found for these mutations. Besides, we analyzed our assemble genome to look for any insertion/deletions but these two genomes contain no deletions/insertions. However, we have identified one (1) insertion and seven (7) deletions of the eight (8) Bangladeshi strain EPI_ISL_466692, EPI_ISL_450343, EPI_ISL_450344, EPI_ISL_468074,

EPI_ISL_514614, EPI_ISL_445213, EPI_ISL_445217 and EPI_ISL_450842 (**Table 2 and Fig 3**). We additionally scrutinized these deleted regions but we didn't find any domain or motif on this region. Apart from our reported complete genomes, we have also identified the most common mutations occurred in Bangladesh from the complete genomes reported in GSAID database from Bangladesh. These genomes showed three mutations in the positions 14408, 23403 and 28878 compared to reference genome (**Table 1**).

Point mutations alter respective protein structures

We have analyzed the mutational effect of all the mutations. Therefore, the 3D structure was built to explore the mutational effect on the protein structure (**Figure 4**). In this case, two types of 3D structure was built i) the structure with wild type residue and ii) the structure with mutant residue. We have performed the energy minimization of both the wild type and mutants. We have found significant differences in the stability of the structure upon mutation (**Table 3**). Mutants 479th; T → A and 1015th; A → T showed higher energy minimization which predicted these proteins to be more stable than the wild type. All the other protein models based on mutation showed less energy minimization than the wild type protein model. Therefore, these protein structure could be more unstable upon mutation in the protein sequence. The highest difference was observed in the mutation in the 5642th position; G → T mutation (from -23276.78 kj/mol to -22377.976 kj/mol) (**Table 3**). Afterwards, the binding site was analyzed to determine whether the wild type and the mutant residues fell within the ligand binding site or not. The binding site residues confirmed that all the mutations from the complete genome belonged to the binding site region (**Figure 5**). Later, we have performed the drug binding analysis followed by virtual drug screening in DrugBank server. Ivermectin and Remdisivir drugs topped the list of potential drug candidates. We then prepared the protein structures and converted them to .pdbqt format for molecular docking experiment. We identified the binding site region for each protein and set the grid box to allow the drugs only to bind to that specific region. The binding affinity analysis showed that compared to the wild type, the drug Ivermectin bound more tightly to proteins which has mutation at positions 479 (T → A), 5642 (G → T) and 8023 (G → A) which bound more loosely to proteins with mutations at positions 481 (C→A), 1015 (A → T), 5098 (G → T) and 5237 (C → T). Remdisivir binds more tightly to proteins with mutations at 28733 (G → A), 28734 (G → A), and 28735 (G → C) and more loosely to 23255 (A → G) compared to the wild type (**Table 4**). It is to be noted

that the interaction of residues of wild type protein were found to be different than that of mutant model. For example, 479th; T → A mutant model which acquired the V→D amino acid interacted with Glu37, Glu41, LEU177, GLY180, LEU104, VAL108, HIS110, Glu87, Leu88, Lys141, Tyr154 residues whereas wild type interacted with Leu18, Val28, Glu37, Glu41, His 45, Leu53, Val54, Ile71, Arg73, Val86, , Val121, Leu122, Asp139.

Discussion

COVID-19 is highly contagious and the variation in its genome could be a leading reason for this feature. Besides, to understand the origin of the strains, the exploration of the whole-genome sequencing (WGS) data of SARS-CoV-2 strains is highly necessary. In this study, we investigated the variation, insertion, and deletion of the Bangladeshi SARS-CoV-2 strains. We collected the samples SARS-CoV-2/human/BGD/NIB_01/2020 and SARS-CoV-2/human/BGD/NIB-BCSIR_02/2020 from the patients who were tested as COVID-19 positive. We extracted the viral RNA from the samples and converted them to cDNA. We performed the Sanger sequencing of SARS-CoV-2/human/BGD/NIB_01/2020 and Next Generation Sequencing of SARS-CoV-2/human/BGD/NIB-BCSIR_02/2020. The total length of the genomes were 29724 and 29737 nucleotides respectively. These two genomes were submitted in both Global Initiative on Sharing All Influenza Data (GISAID) and National Center for Biotechnology Information (NCBI) databases. These two databases accepted the genomes and provided the accession number EPI_ISL_458133 from GISAID and MT509958 and MT568643 from NCBI. We investigated the possible variations of these two genome and found thirteen (13) mutations in SARS-CoV-2/human/BGD/NIB_01/2020 against the reference genome of SARS-CoV-2 (**Table 1 and Figure 2**). The mutations belonged different regions of the genome, but mostly found in the Orf11ab gene. Eight (8) mutations were found in the Orf1ab region. These mutations were 479th; T → A, 481th; C→A; 1015th; A → T, 2889th; C → T, 5098th; G → T; 5237th; C → T, 5642th; G → T, 8023rd; G → A. The mutation 93rd; C → T was located in the 5' upstream region of the gene. The mutations 28733rd; G → A, 28734th; G → A and 28735th; G → C were located in the N gene of the genome. 23255th; A → G mutation was located in the S gene of the genome (**Table 1**). Among the thirteen (13) mutations, six (6) mutations, namely 93rd; C → T, 2889th; C → T, 23255th; A→G, 28733rd; G → A, 28734th; G → A and 28735th; G → C were reported in different countries according to CNCB[3] database (**Table 1**). All the variations showed a missense effect upon structure except

2889th; C → T variant (**Table 1**). The other seven (7) mutations, 479th; T → A, 481th; C→A; 1015th; A → T, 5098th; G → T; 5237th; C → T, 5642th; G → T and 8023rd; G → A presented themselves as unique mutations in the MT509958 genome (**Table 1 and Figure 2**). Surprisingly, we didn't find any mutation in SARS-CoV-2/human/BGD/NIB-BCSIR_02/2020. We also looked for indel profile of our assembled genomes. However, we did not find any insertion/deletion occurred into the genome. We looked for it in the rest of the genomes of SARS-CoV-2 in Bangladesh. Seven (7) deletions were found in Bangladeshi strains EPI_ISL_450343, EPI_ISL_450344, EPI_ISL_468074, EPI_ISL_514614, EPI_ISL_445213, EPI_ISL_445217 and EPI_ISL_450842. Among them, EPI_ISL_450343, EPI_ISL_450344, EPI_ISL_468074 and EPI_ISL_514614 shared common deletions. These deletions belonged to the ORF8 gene and the length ranged from 27913 to 28254 (**Table 2 and Figure 3**). We have found another deletion in the Orf7a gene whose position in the genome ranged from 27476 to 27668. This was found in an isolate from Dhaka region of Bangladesh (EPI_ISL_450842) (**Figure 3**).

Three (3) dimensional structures were built for both wild type and the mutants in order to observe the mutational impact on their corresponding proteins (**Figure 4**). We have predicted the stability of the protein structure of the corresponding variant based on energy minimization analysis. The variants 479th; T → A and 1015th; A → T were the more stable compare to the other variants. These variant's structures consumed more energy than the wild type structure. The other variants exhibited a decrease in stability (**Table 3**). The binding site of the protein structures was analyzed to look for the location of the relevant amino acid variant. We have found that all the variants were located within the ligand binding site (**Figure 5**). Therefore, these residues could be considered very important in terms of ligand/drug binding. Ivermectin and Remdisivir were selected for the drug binding analysis. We performed molecular docking for each of the wild type and mutant structures with these two drugs and analyzed the interactions. We observed that the interacting wild type residues were replaced with different residues in after molecular docking with the drugs (**Table 4**). Here, the binding affinity was also found to be different from the wild type structure. It is to be clearly understood that only a single amino acid change from the wild type structure was responsible for these changes. From these analysis, it can be concluded that if any therapeutics are to be applied on these variants, the therapeutics might not work effectively due to the alteration of the residues in the mutant proteins.

Conclusions

To reiterate the core of our study, we have performed whole genome sequencing of SARS-CoV-2 to identify genetic variations and then analyzed their impact on the structures of their corresponding proteins. We have also identified the insertions/deletions among all the sequenced Bangladeshi SARS-CoV-2 strains. The energy minimization and the drug binding analysis suggested that the identified mutations might have significant impact on structure and function of their target proteins. Therefore, the present study might be of great interest to the researchers/companies working to develop therapeutics against SARS-CoV-2 as well as gaining fundamental insights into pathogenesis of the virus.

References

[1] C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, L. Zhang, G. Fan, J. Xu, X. Gu, Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China, The Lancet, 395 (2020) 497-506

[2] Q. Li, X. Guan, P. Wu, X. Wang, L. Zhou, Y. Tong, R. Ren, K.S. Leung, E.H. Lau, J.Y. Wong, Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia, New England Journal of Medicine, (2020).

[3] L. Zou, F. Ruan, M. Huang, L. Liang, H. Huang, Z. Hong, J. Yu, M. Kang, Y. Song, J. Xia, SARS-CoV-2 viral load in upper respiratory specimens of infected patients, New England Journal of Medicine, 382 (2020) 1177-1179.

[4] D. Wang, B. Hu, C. Hu, F. Zhu, X. Liu, J. Zhang, B. Wang, H. Xiang, Z. Cheng, Y. Xiong, Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus–infected pneumonia in Wuhan, China, JAMA, 323 (2020) 1061-1069.

[5] Y.S. Malik, S. Sircar, S. Bhat, K. Sharun, K. Dhama, M. Dadar, R. Tiwari, W. Chaicumpa, Emerging novel coronavirus (2019-nCoV)—current scenario, evolutionary perspective based on genome analysis and recent developments, Veterinary Quarterly, 40 (2020) 68-76.

[6] Z. Xu, L. Shi, Y. Wang, J. Zhang, L. Huang, C. Zhang, S. Liu, P. Zhao, H. Liu, L. Zhu, Pathological findings of COVID-19 associated with acute respiratory distress syndrome, The Lancet respiratory medicine, (2020).

[7] Y. Chen, Q. Liu, D. Guo, Emerging coronaviruses: genome structure, replication, and pathogenesis, Journal of medical virology, (2020).

[8] A. Wu, Y. Peng, B. Huang, X. Ding, X. Wang, P. Niu, J. Meng, Z. Zhu, Z. Zhang, J. Wang, Genome composition and divergence of the novel coronavirus (2019-nCoV) originating in China, Cell host & microbe, (2020).

[9] A.R. Fehr, S. Perlman, Coronaviruses: an overview of their replication and pathogenesis, in: Coronaviruses, Springer, 2015, pp. 1-23.

[10] J. Hellewell, S. Abbott, A. Gimma, N.I. Bosse, C.I. Jarvis, T.W. Russell, J.D. Munday, A.J. Kucharski, W.J. Edmunds, F. Sun, Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts, The Lancet Global Health, (2020).

[11] R. Lu, X. Zhao, J. Li, P. Niu, B. Yang, H. Wu, W. Wang, H. Song, B. Huang, N. Zhu, Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding, The Lancet, 395 (2020) 565-574.

[12] DNA Sequence Assembler v4 (2013), Heracle BioSoft, www.DnaBaser.com

[13] Burland TG. DNASTAR's Lasergene sequence analysis software. Methods Mol Biol. 2000;132:71-91. doi: 10.1385/1-59259-192-2:71. PMID: 10547832.

[14] Paijmans, Johanna LA, Sina Baleka, Kirstin Henneberger, Ulrike H. Taron, Alexandra Trinks, Michael V. Westbury, and Axel Barlow. "Sequencing single-stranded libraries on the Illumina NextSeq 500 platform." arXiv preprint arXiv:1711.11004 (2017).

[15] Johnson, Mark, Irena Zaretskaya, Yan Raytselis, Yuri Merezhuk, Scott McGinnis, and Thomas L. Madden. "NCBI BLAST: a better web interface." Nucleic acids research 36, no. suppl_2 (2008): W5-W9.

[16] Shu, Yuelong, and John McCauley. "GISAID: Global initiative on sharing all influenza data–from vision to reality." Eurosurveillance 22, no. 13 (2017): 30494.

[17] Carver, Tim J., Kim M. Rutherford, Matthew Berriman, Marie-Adele Rajandream, Barclay G. Barrell, and Julian Parkhill. "ACT: the Artemis comparison tool." Bioinformatics 21, no. 16 (2005): 3422-3423.

[18] Kim, David E., Dylan Chivian, and David Baker. "Protein structure prediction and analysis using the Robetta server." Nucleic acids research 32, no. suppl_2 (2004): W526-W531.

[19] Lindahl, Erik, Berk Hess, and David Van Der Spoel. "GROMACS 3.0: a package for molecular simulation and trajectory analysis." Molecular modeling annual 7, no. 8 (2001): 306-317.

[20] Trott, Oleg, and Arthur J. Olson. "AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading." Journal of computational chemistry 31, no. 2 (2010): 455-461.

[21] Bell, Trevor G., and Anna Kramvis. "Fragment merger: an online tool to merge overlapping long sequence fragments." Viruses 5, no. 3 (2013): 824-833.

Tables:

Table 1: NIB_01 polymorphisms against reference sequence and their mutational effect.

SL No.	**Query (Reference) Position**	**Position in subject(NIB-01) sequence**	**Amino acid**	***Query → Subject***	**Gene**	**Mutants Observed**	**Mutants observed in No. in Virus**	**Impact on Proteins**	**Protein ID of mutants**
1.	241th	93	-	**C** → T	5` UTR	US, UK	15042	Upstream variant	QHD43415.1
2.	627th	479	Val121	**T** → A	Orf1ab	-	-	-	-
3.	629th	481	Leu122	**C** → A	Orf1ab	-	-	-	-
4.	1163th	1015	Ile120	**A** → T	Orf1ab	-	-	Coding Variant	-
5.	3037th	2889	Phe924	**C** → T	Orf1ab	US, UK	15059	Coding Variant	QHD43

									415.1
6.	5246th	5098	Val843	**G** → T	Orf1ab	-	-	Coding Variant	QHD43415.1
7.	5385th	5237	Ala889	C → T	Orf1ab	-	-	Coding Variant	QHD43415.1
8.	7790th	5642	Gly1691	**G** → T	Orf1ab	-	-	Coding Variant	QHD43415.1
9.	8171th	8023	Ala1818	**G** → A	Orf1ab	-	-	Coding Variant	QHD43415.1
10.	23403th	23255	Asp614	**A** → G	"S"	US, UK	15118	Coding Variant	QHD43416.1
11.	28881th	28733	Ser202	**G** → A	"N"	UK	4551	Coding Variant	QHD43423.2
12.	28882th	28734	Arg203	**G** → A	"N"	UK	4542	Coding Variant	QHD43423.2
13.	28883th	28735	Gly204	**G** → C	"N"	UK	4537	Coding Variant	QHD43

									423.2

Table 2: Indel profile of Bangladeshi whole genome.

Serial No.	Sequence Identifier	Indel Type	Region
1)	hCoV-19/Bangladesh/BCSIR-NILMRC-071/2020|EPI_ISL_466692|2020-05-26	Insertion	ORF8 (27910-27985)
2)	hCoV-19/Bangladesh/BARJ-CVASU-CTG-511/2020|EPI_ISL_450343|2020-05-09	Deletion	ORF8 (27913-28254)
3)	hCoV-19/Bangladesh/BARJ-CVASU-CTG-517/2020|EPI_ISL_450344|2020-05-03	Deletion	ORF8 (27913-28254)
4)	hCoV-19/Bangladesh/CHRF-0006/2020|EPI_ISL_468074|2020-05-17	Deletion	ORF8 (27913-28254)
5)	hCoV-19/Bangladesh/NGRI-NSTU-31/2020|EPI_ISL_514614|2020-07-21	Deletion	ORF8 (27913-28254)
6)	hCoV-19/Bangladesh/DNAS-CPH-	Deletion	ORF7 (27476-27668)

	467/2020|EPI_ISL_445213|2020-04-28		
7)	hCoV-19/Bangladesh/DNAS-CPH-436/2020|EPI_ISL_445217|2020-04-28	Deletion	ORF7 (27476-27668)
8)	hCoV-19/Bangladesh/DU-50761/2020|EPI_ISL_450842|2020-05-06	Deletion	ORF7 (27476-27668)

Table 3: Energy minimization of the wild type and mutant model.

Sl no.	**Variant in NIB_01 Genome position**	**Energy Minimization score**	**Stability of protein**
1.	479 (Wild Type)	-4714.146 $^{kj}/_{mol}$	Stable
2.	479 (Mutant Type)	-4718.002 $^{kj}/_{mol}$	
3.	481(Wild Type)	-4710.914 $^{kj}/_{mol}$	De-stable
4.	481 (Mutant Type)	-4382.711 $^{kj}/_{mol}$	
5.	1015 (Wild Type)	-14247.945 $^{kj}/_{mol}$	Stable
6.	1015 (Mutant Type)	-14678.767 $^{kj}/_{mol}$	
7.	5098 (Wild Type)	-10989.878 $^{kj}/_{mol}$	De-stable
8.	5098 (Mutant)	-10847.174 $^{kj}/_{mol}$	
9.	5237 (Wild Type)	-10989.688 $^{kj}/_{mol}$	De-stable
10.	5237 (Mutant type)	-10731.945 $^{kj}/_{mol}$	
11.	5642 (Wild type)	-23276.789 $^{kj}/_{mol}$	De-stable

12.	5642 (Mutant type)	-22377.976 $^{kj}/_{mol}$	
13.	8023 (Wild Type)	-23276.789 $^{kj}/_{mol}$	De-stable
14.	8023 (Mutant Type)	-23051.338 $^{kj}/_{mol}$	
15.	23255 (Wild type)	-9876.789 $^{kj}/_{mol}$	De-stable
16.	23255 (Mutant type)	-9308.898 $^{kj}/_{mol}$	
17.	28733 (Wild Type)	-8789.976 $^{kj}/_{mol}$	De-stable
18.	28733 (Mutant Type)	-8567.712. $^{kj}/_{mol}$	
19.	28734 (Wild Type)	-8789.976 $^{kj}/_{mol}$	De-stable
20.	28734 (Mutant Type)	-8356.435 $^{kj}/_{mol}$	
21.	28735 (Wild Type)	-8789.976 $^{kj}/_{mol}$	De-stable
22.	28735 (Mutant Type)	-8089.172. $^{kj}/_{mol}$	

Table 4: Drug binding analysis of variant.

Variant Position in subject(NIB-01) sequence	**Drug Name**	**Binding Energy**	**Amino Acid Residue**
479 (Wild Type)	Ivermectin	-7.9 kcal/mol	Leu18, Val28, Glu37, Glu41, His 45, Leu53, Val54, Ile71, Arg73, Val86, , Val121, Leu122, Asp139
479 (Mutant Type)	Ivermectin	-8.2 kcal/mol	Glu37, Glu41, LEU177, GLY180, LEU104, VAL108, HIS110, Glu87, Leu88, Lys141, Tyr154
481(Wild Type)	Ivermectin	-8.7 kcal/mol	Leu18, Val28, Glu37, Glu41, His 45, Leu53, Val54, Ile71, Arg73, Val86, , Val121, Leu122, Asp139
481 (Mutant Type)	Ivermectin	-7.7 kcal/mol	Glu37, Glu41, LEU177, GLY180, LEU104, VAL108, HIS110, Glu87, Leu88, Lys141, Tyr154

1015 (Wild Type)	Ivermectin	-7.9 kcal/mol	Lys22, Met135, His202, Glu206, Lys214, Arg107,Ile120,Val126, Asn133, Phe633, Thr634, Ala889
1015 (Mutant Type)	Ivermectin	-6.9 kcal/mol	GLY235, ALA241, ILE251,GLY260, ASP268, SER303, LYS317, LYS335, GLU359, ALA360, THR371, PHE406, THR412
5098 (Wild Type)	Ivermectin	-8.8 kcal/mol	ASP137, THR149, TYR170, VAL173, CYS186, ALA206,Ala219, CYS230
5098 (Mutant)	Ivermectin	-8.0 kcal/mol	LEU133, PRO134, SER178, LEU195, ARG213, ALA214, LEU225
5237 (Wild Type)	Ivermectin	-7.6 kcal/mol	ASP137, THR149, TYR170, VAL173, CYS186, ALA206,Ala219, CYS230
5237 (Mutant type)	Ivermectin	-7.1 kcal/mol	LEU133, PRO134, SER178, LEU195, ARG213, ALA214, LEU225
5642 (Wild type))	Ivermectin	-6.7kcal/mol	HIS12, CYS23, VAL33, SER52, LYS61, TYR68, GLY73, VAL104, ILE133, SER147, THR196 ALA200, ASN235, ARG250
5642 (Mutant type)	Ivermectin	-8.5 kcal/mol	PHE69, ASN87, ILE105, LYS119, LEU135, VAL150, ALA151, GLU181, ASP251, LEU252, CYS258
8023 (Wild Type)	Ivermectin	-6.6 kcal/mol	HIS12, CYS23, VAL33, SER52, LYS61, TYR68, GLY73, VAL104, ILE133, SER147, THR196 ALA200, ASN235, ARG250
8023 (Mutant Type)	Ivermectin	-7.5 kcal/mol	PHE69, ASN87, ILE105, LYS119, LEU135, VAL150, ALA151, GLU181, ASP251, LEU252, CYS258

23255 (Wild type)	Remdisivir	-8.7 kcal/mol	Gln321, Glu324, Ile326, His519, Phe543, Lys557, Thr553, Lys 557, Ile569, Ala570, Asp614
23255 (Mutant type)	Remdisivir	-8.3 kcal/mol	ASN87, ILE128, ASN234, ILE235, ASP290, PHE329, PHE347, LEU368, ASN394, THR415, GLY416, ASN501, PHE515, GLY526
28733 (Wild Type)	Remdisivir	-7.7 kcal/mol	Ser21, Asp22, Ser202, Arg203, Gly204Leu224, Arg226, Glu280, Gln281, Ser318, Lys338, Val392
28733 (Mutant Type)	Remdisivir	-8.1 kcal/mol	ASN192, GLY295, GLN306, GLU323, SER327, GLY328, ASP343, LEU353, ASN354, THR366, THR379, VAL392, ASP415, SER416, THR417, GLN418, ALA419
28734 (Wild Type)	Remdisivir	-6.8 kcal/mol	Ser21, Asp22, Ser202, Arg203, Gly204Leu224, Arg226, Glu280, Gln281, Ser318, Lys338, Val392
28734 (Mutant Type)	Remdisivir	-7.8 kcal/mol	ASN192, GLY295, GLN306, GLU323, SER327, GLY328, ASP343, LEU353, ASN354, THR366, THR379, VAL392, ASP415, SER416, THR417, GLN418, ALA419
28735 (Wild Type)	Remdisivir	-7.1 kcal/mol	Ser21, Asp22, Ser202, Arg203, Gly204, Leu224, Arg226, Glu280, Gln281, Ser318, Lys338, Val392
28735 (Mutant Type)	Remdisivir	-8.4 kcal/mol	ASN192, GLY295, GLN306, GLU323, SER327, GLY328, ASP343, LEU353, ASN354, THR366, THR379, VAL392, ASP415, SER416, THR417, GLN418, ALA419

Red color denotes the wild mutation.

Figures:

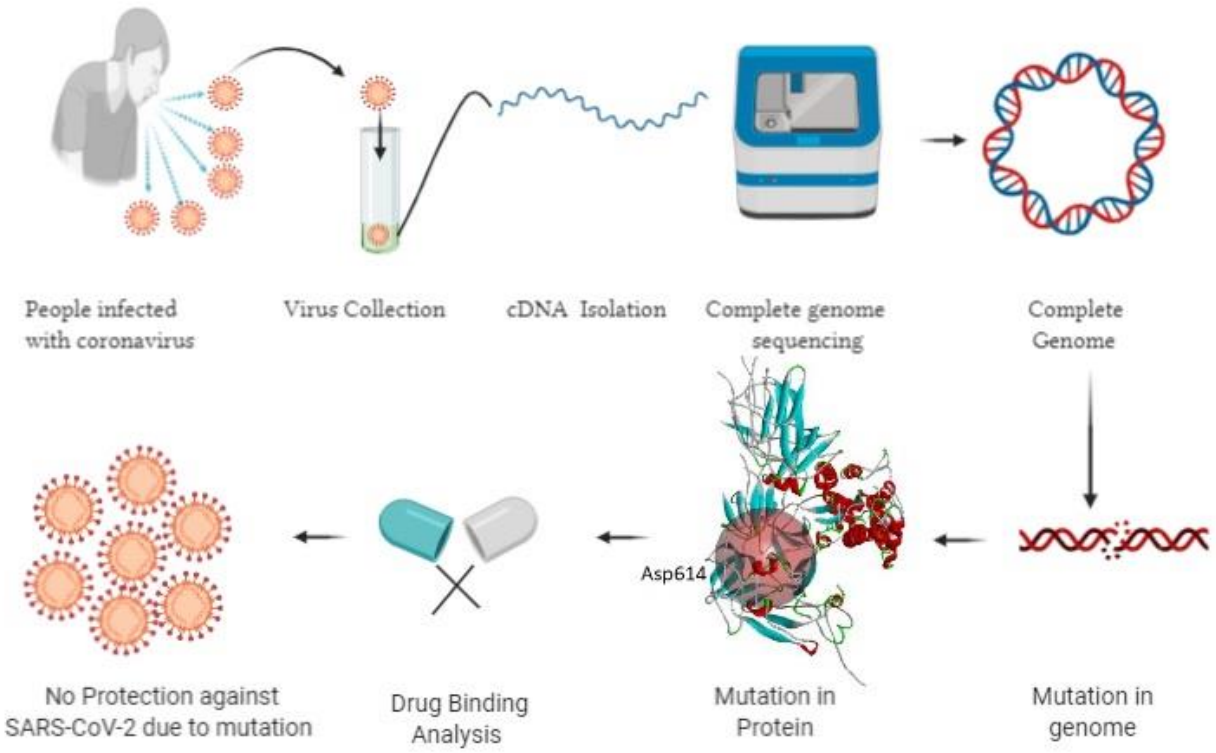
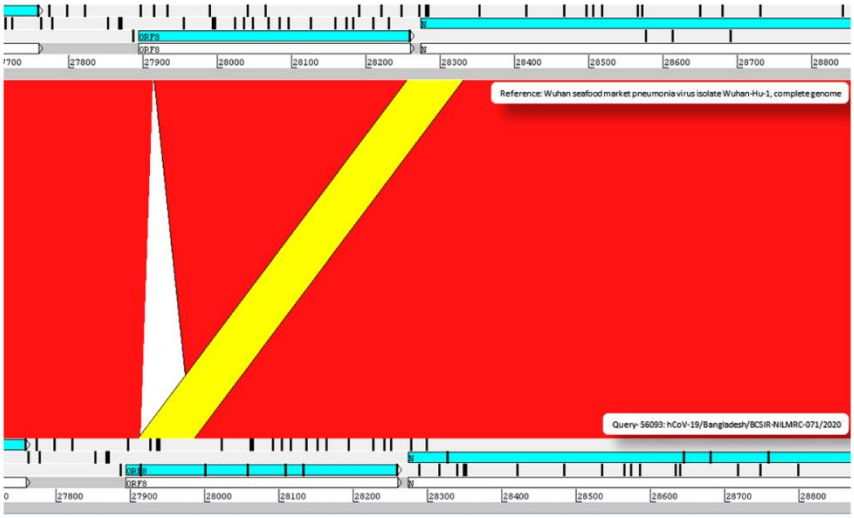


Fig 1: Complete workflow of the study

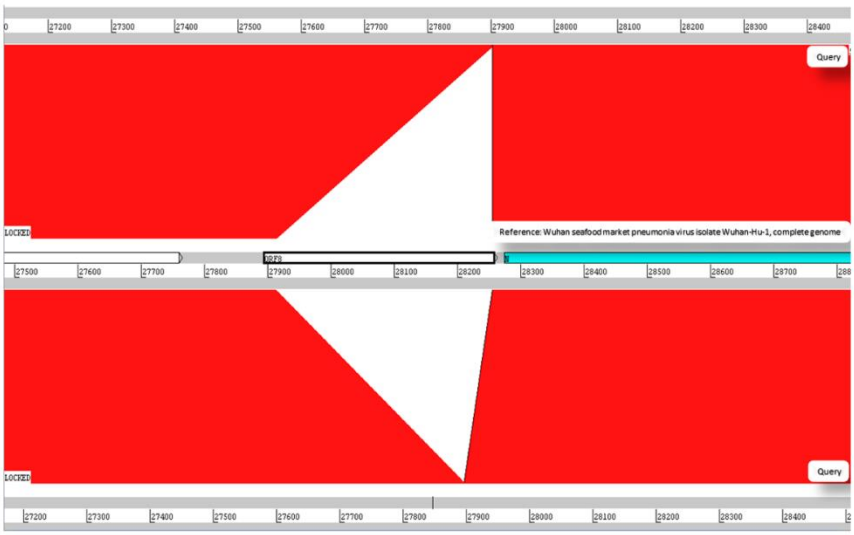
```
Query  121   VLLRKNGNKGAGGHSYGADLKSFDLGDELGTDPYEDFQENWNTKHSSGVTRELMRELNGG  180
Sbjct  121   DI............................................................  180

Query  61    YTERSEKSYELQTPFEIKLAKKFDTFNGECPNFVFPLNSIIKTIQPRVEKKKLDGFMGRI  120
Sbjct  61    ............................................................F  120

Query  841   PQVNGLTSIKWADNNCYLATALLTLQQIELKFNPPALQDAYYRARAGEAANFCALILAYC  900
Sbjct  841   ..F.............................................V...........  900

Query  1681  GSIHLYFDKAGQKTYERHSLSHFVNLDNLRANNTKGSLPINVIVFDGKSKCEESSAKSAS  1740
Sbjct  1681  ..........C...................................................  1740

Query  1801  ELAKNVSLDNVLSTFISAARQGFVDSDVETKDVVECLKLSHQSDIEVTGDSCNNYMLTYN  1860
Sbjct  1801  .................T........................................  1860
```

Fig 2: Novel Mutation in protein sequence in NIB_01 whole genome

(A)


Reference: Wuhan seafood market pneumonia virus isolate Wuhan-Hu-1, complete genome
Query- 56093: hCoV-19/Bangladesh/BCSIR-NILMRC-071/2020
ORF8
N

(B)


Query
Reference: Wuhan seafood market pneumonia virus isolate Wuhan-Hu-1, complete genome
ORF8
N

(C)

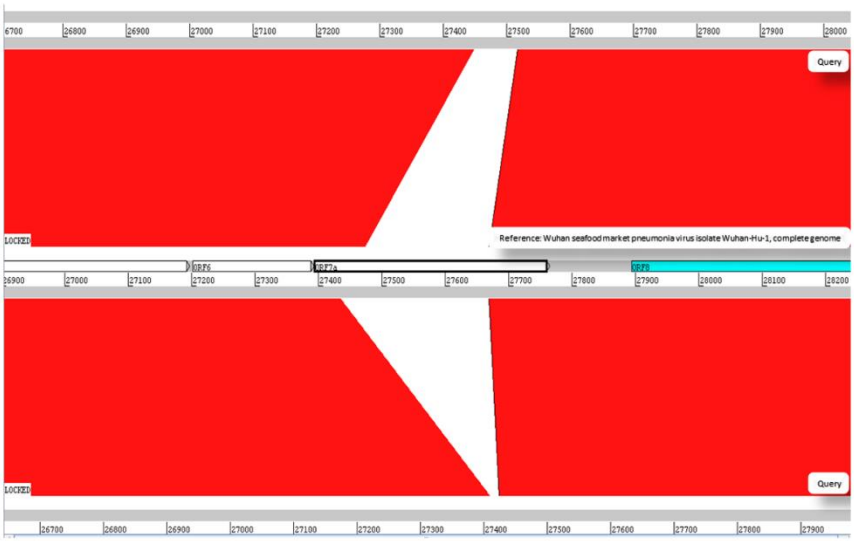
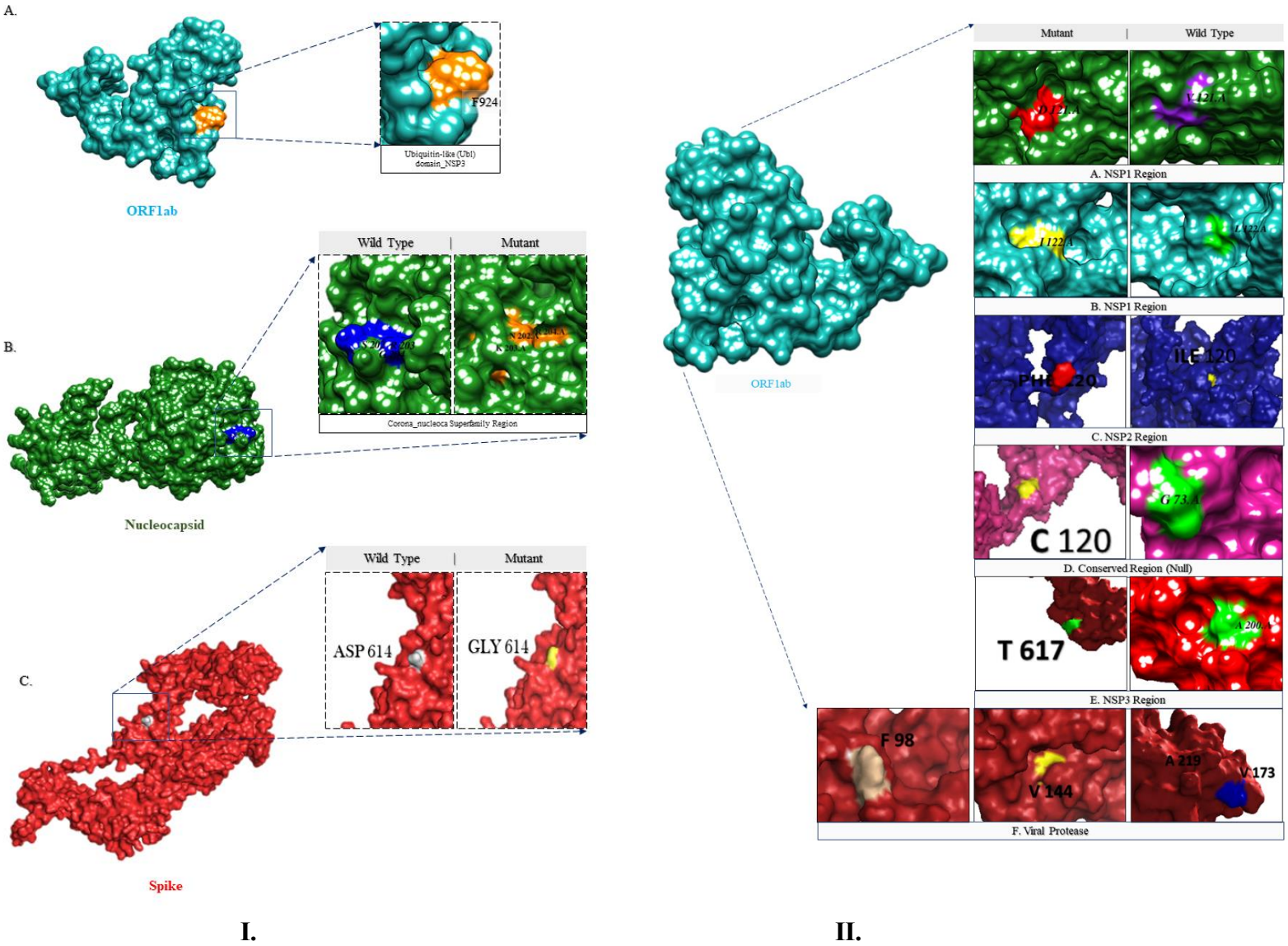

Query
Reference: Wuhan seafood market pneumonia virus isolate Wuhan-Hu-1, complete genome
ORF6
ORF7a
ORF8

Fig 3: Deletion of sequences in complete genome of the Bangladesh strain. Indel experimentation result from comparative genome browsing against reference genome in Artemis window. A) Mobile element insertion from **N** to **ORF8** in EPI_ISL_466692. Yellow bar representing homologous sequence moiety between these two regions. B) EPI_ISL_450343, EPI_ISL_450344, EPI_ISL_468074, EPI_ISL_514614 are sharing the same deletion pattern in **ORF8**. Majority portion of **ORF8,** lost in the deletion event. C) EPI_ISL_445213, EPI_ISL_445217 and EPI_ISL_450842 are also sharing the same deletion pattern in **ORF7a**. Majority portion of **ORF7a,** lost in the deletion event.



I. **II.**

Fig 4: Amino acid location in tertiary structure. I) Previously Found Variants. II) Unique Variants to Bangladesh.

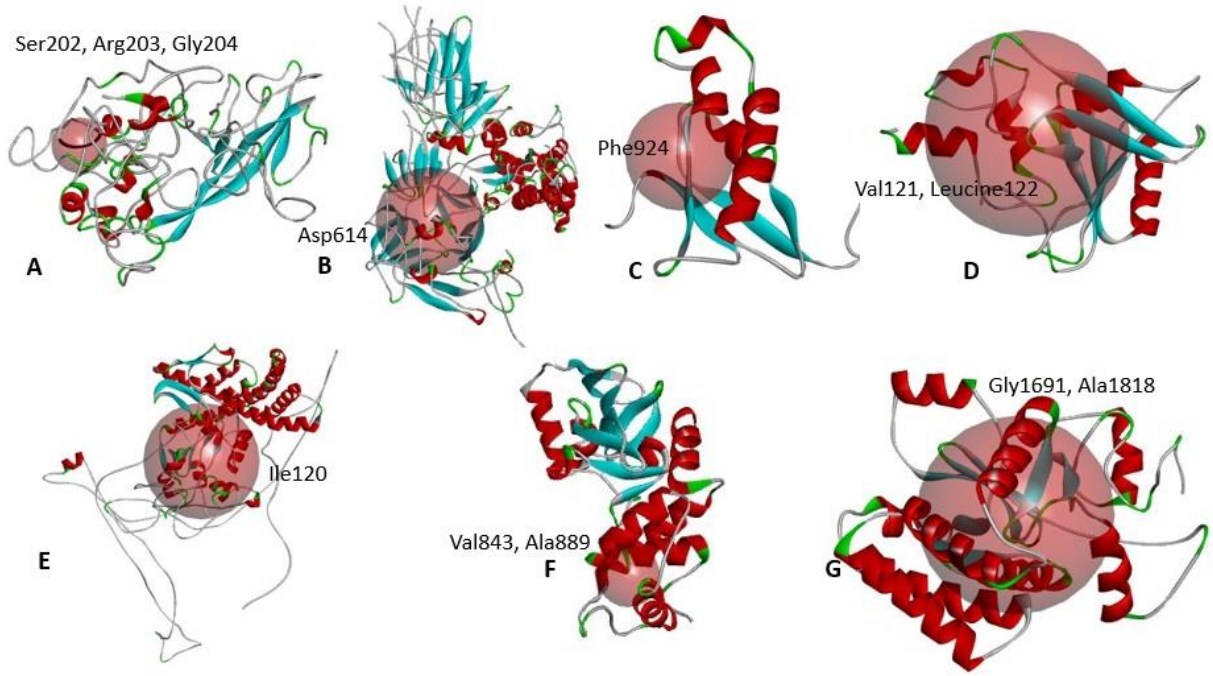


Fig 5: Mutation in Binding site of the protein.