# Global and Local Mutations in Bangladeshi SARS-CoV-2 Genomes

Md. Mahbub Hasan, Rasel Das, Md. Rasheduzzaman, Md. Hamed Hussain, Nazmul Hasan Muzahid, Asma Salauddin, Meheadi Hasan Rumi, S.M. Mahbubur Rashid, Zonaed Siddiki AMAM, Adnan Mannan

PII: S0168-1702(21)00097-6

DOI: https://doi.org/10.1016/j.virusres.2021.198390

Reference: VIRUS 198390

To appear in: *Virus Research*

Received Date: 29 August 2020

Revised Date: 7 March 2021

Accepted Date: 9 March 2021

Please cite this article as: Hasan MM, Das R, Rasheduzzaman M, Hussain MH, Muzahid NH, Salauddin A, Rumi MH, Mahbubur Rashid SM, AMAM ZS, Mannan A, Global and Local Mutations in Bangladeshi SARS-CoV-2 Genomes, *Virus Research* (2021), doi: https://doi.org/10.1016/j.virusres.2021.198390

# Global and Local Mutations in Bangladeshi SARS-CoV-2 Genomes

Md. Mahbub Hasan[1,2#], Rasel Das[3#], Md. Rasheduzzaman[1], Md Hamed Hussain[4], Nazmul Hasan Muzahid[4], Asma Salauddin[1], Meheadi Hasan Rumi[1], S M Mahbubur Rashid[5], AMAM Zonaed Siddiki[6], Adnan Mannan[1*]

[1] Department of Genetic Engineering and Biotechnology, Faculty of Biological Sciences, University of Chittagong, Chattogram-4331, Bangladesh.

[2] Institute of Pharmaceutical Science, School of Cancer and Pharmaceutical Sciences, King's College London, Franklin-Wilkins Building, 150 Stamford Street, London SE1 9NH, UK.

[3] Global Innovation Center, Kyushu University, Fukuoka 816-8580, Japan.

[4] School of Science, Monash University Malaysia, 47500 Selangor, Malaysia.

[5] Department of Genetic Engineering and Biotechnology University of Dhaka, Ramna, Dhaka-1000, Bangladesh.

[6] Genomics Research Group, Faculty of Veterinary Medicine, Chittagong Veterinary and Animal Sciences University, Chattogram-4202, Bangladesh.

#These authors contributed equally.

*Corresponding Author: Adnan Mannan, E-mail: adnan.mannan@cu.ac.bd

## Highlights

- Comprehensive genomic analysis of 371 SARS-CoV-2 genomes from Bangladeshi patients.
- Extensive analyses showed 4604 mutations among Bangladeshi SARS-CoV-2 genomes.
- D614G mutation in spike glycoprotein was foremost (98%) in Bangladeshi isolates.
- The average mutation number was higher in genomes with mutation at G614 than D614.
- A total of 34 unique amino acid changes were identified in this study.

## Abstract:

Coronavirus Disease 2019 (COVID-19) warrants comprehensive investigations of publicly available Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) genomes to gain new insight about their epidemiology, mutations, and pathogenesis. Nearly 0.4 million mutations are identified so far among the ~60,000 SARS-CoV-2 genomic sequences. In this study, we compared a total of 371 SARS-CoV-2 published whole genomes reported from different parts of Bangladesh with 467 sequences reported globally to understand the origin of viruses, possible patterns of mutations, and availability of unique mutations. Phylogenetic analyses indicated that SARS-CoV-2 viruses might have transmitted through infected travelers from European countries, and the GR clade was found as predominant in Bangladesh. Our analyses revealed 4604 mutations at the RNA level including 2862 missense mutations, 1192 synonymous mutations, 25 insertions and deletions and 525 other types of mutation. In line with the global trend, D614G mutation in spike glycoprotein was predominantly high (98%) in Bangladeshi isolates. Interestingly, we found the average number of mutations in ORF1ab, S, ORF3a, M, and N were significantly higher ($p$<0.001) for sequences containing the G614 variant compared to those having D614. Previously reported frequent mutations, such as R203K, D614G, G204R, P4715L and I300F at protein levels were also prevalent in Bangladeshi isolates. Additionally, 34 unique amino acid changes were revealed and categorized as originating from different cities.

These analyses may increase our understanding of variations in SARS-CoV-2 virus genomes, circulating in Bangladesh and elsewhere.

**Keywords:** SARS-CoV-2; Whole-genome sequence; COVID-19; Mutation; Bangladesh

## 1. Introduction:

Severe Acute Respiratory Syndrome-Coronavirus-2 (SARS-CoV-2) has been identified as the etiological agent of the disease called Coronavirus Disease-2019 (COVID-19). To elucidate the viral pathogenesis, modern genomic tools are highly crucial and have been employed by researchers around the world. An increasing number of whole-genome datasets of the SARS-CoV-2 virus are now being submitted in publicly accessible databases from different parts of the globe every day. It is high time to analyze the variations among those sequences which will help future strategic efforts for its preventive measures, such as vaccine design and development of potential novel therapeutics. SARS-CoV-2 consists of positive-sense single-stranded RNA with a genome size ranging from ~29.8 to 29.9 Kb (Khailany et al., 2020). It contains a variable number (9-11) of open-reading frames (ORFs) and the first ORF covers almost two-thirds of the whole genome (Khailany et al., 2020; Kim et al., 2020). The genome encodes 4 structural, 16 non-structural (NS), and 6 accessory proteins (Astuti, 2020; Chen et al., 2020; O'Meara et al., 2020; Yoshimoto, 2020).

According to a recent study on 48,635 SARS-CoV-2 genome sequences, a total of 353,341 mutations have been observed globally compared with that of the reference genome from Wuhan (Accession ID MN908947). Of them, sequences from India, Congo, Bangladesh, and Kazakhstan were reported to show significantly high numbers of mutations per sample

compared with the global average (Mercatelli and Giorgi, 2020). Out of these mutations, D614G mutation (causing aspartate to glycine at 614 in S protein ) is reported to be the most prevalent mutations in Europe, Oceania, South America, Africa, Middle East, and India (Gong et al., 2020; Raghav et al., 2020; Zhang et al., 2020). Several laboratory experiment-based investigations have reported that the level of angiotensin-converting enzyme 2 (ACE2) expression was distinctly higher by the retroviruses pseudo-typed with G614 compared with that of D614 in cell culture experiments (Dumonteil and Herrera, 2020; Zhang et al., 2020). Another reported mutation of ORF1ab is P4715L linked with D614G, which is linked with higher fatality rates in 28 countries and 17 states of the United States (Mercatelli and Giorgi, 2020; Toyoshima et al., 2020). Three other mutations namely C14408T (ORF1b), C241T (5' UTR), C3037T (ORF1a) are reported to be common and coexisting in the same genome, while G11083T has been found mostly in Asian countries (Toyoshima et al., 2020).

In Bangladesh, until November, 2020, nearly 465,000 people have been infected and 6644 people have died due to COVID-19 (https://iedcr.gov.bd/). During this period, a total of 371 complete and high coverage SARS-CoV-2 whole-genome sequences were submitted in the GISAID database (https://www.gisaid.org). Analyses of these sequences are sparsely reported in the literature. As for instance, one study reported analyses of 60 SARS-CoV-2 genome sequences from Bangladesh and compared them with 6 Southeast Asian countries (Islam et al., 2020). Another group analyzed only 14 isolates and found 42 mutations (Parvez et al., 2020). A separate study identified only 9 variants where unique mutations (UMs) were found prevalent mostly in ORF1ab (Hasan et al., 2020). Another study involved 64 SARS-CoV-2 whole-genome sequences and identified the presence of 180 mutations in the coding regions of the viruses, and mutations at nsp2 were the most prevalent (Ahmed et al., 2020). Due to the small numbers of

genome sequences analyzed, most of these findings were not conclusive and truly representative of the whole country. Further comprehensive analyses are therefore necessary to better understand the circulating virus in this highly populated country.

The present study was based on SARS-CoV-2 genome sequences submitted in the global publicly accessible GISAID database (GISAID, 2020). According to the GISAID database, clades are characterized by genomic specificity and classified by calculating genome distances of the variants through phylogenetic clusters analysis. Currently, all of the SARS-CoV-2 variants are classified into 7 clades- S, L, V, G, GH, GR, and GV. We used this clade-based classification in this study. We compared a total of 371 whole-genome sequences obtained from isolates collected from Bangladeshi COVID-19 patients with that of 467 global sequences that are from Asia, Africa, Europe, Australia and North American countries using time-resolved phylogenetic analysis. We studied the frequency of mutations in different ORFs of SARS-CoV-2 genomes, having the mutation of D614G. In addition, we explored the origin and distribution of SARS-CoV-2 in Bangladesh along with the circulating variants present in different parts of the country. This was achieved through identifying the pattern of mutations including the unique mutations (UMs). These pave the way to increase our understanding of the distribution of different variants of SARS-CoV-2 virus in different regions and associated mutation events.

## 2. Materials and methods:

### 2.1 Dataset

A total of 35,723 complete high-coverage genomic RNA sequences of SARS-CoV-2 were submitted to GISAID until June 30, 2020. From these downloaded sequences, a custom python script was used to retrieve unique sequences. The same script also removed any sequence,

containing "N" and other ambiguous IUPAC codes (Korber et al., 2020a). A total of 8723 complete genomic sequences were returned upon running the script as per the aforementioned logic. To select representative sequences from curated 8723 global sequences and make a comparison against sequences from Bangladesh, priorities were given to those countries that had a higher number of infections in each continent (https://www.worldometers.info/coronavirus/). We selected these sequences considering that each continent must be represented by at least one sequence from each GISAID clade. The number of sequences selected from a country was based on the total number of unique sequences retrieved. This resulted in a total of 839 unique representative sequences from 42 countries (see Supplementary File Table S1).

**2.2 Phylogenetic analysis**

From Bangladesh, 518 whole-genome RNA sequences of SARS-CoV-2 were uploaded to GISAID until November 30, 2020. Only high coverage complete sequences (n=371) were considered for analysis in this study. All of these 371 sequences were retrieved and aligned with the previously selected 467 representative sequences along with that of Wuhan-1 (Accession ID MN908947) as a reference sequence (Wu et al., 2020). A list of accession numbers of all sequences, used in this study, is provided in Supplementary File Table S2. To ensure comparability, the flanks of all the sequences were truncated to the consensus range from 56 to 29,797 (Forster et al., 2020), with nucleotide position numbering according to the Wuhan-1 reference sequence, prior to alignment. Multiple Sequence Alignment (MSA) and phylogenetic tree construction were carried out using Molecular Evolutionary Genetic Analysis X (MEGA X) software (version 10.1) (Kumar et al., 2018). The selected sequences were aligned using the

MUSCLE software tool (Edgar, 2004). Later, a NJ (Neighbor-Joining (NJ)) phylogenetic tree was constructed using the Tamura-Nei model (Saitou and Nei, 1987). Tree topology was assessed using a fast bootstrapping function with 1000 replicates. Tree visualization and annotations were performed in the Interactive Tree of Life (iTOL) v5 (Letunic and Bork, 2019).

### 2.3 Mutation analysis

The Genome Detective Coronavirus Typing Tool Version 1.13 was used for variant analyses at the RNA level of SARS-CoV-2 which is specially designed for this virus analysis (https://www.genomedetective.com/app/typingtool/cov/) (Cleemput et al., 2020). For investigating the UM at protein level among 371 genomic sequences from Bangladesh, we used a CoV server hosted by the GISAID server (https://www.gisaid.org/epiflu-applications/covsurver-mutations-app/). The server analyzed our dataset against all available genomic sequences of SARS-CoV-2 including the Wuhan reference sequence deposited on GISAID until November 30, 2020. The spatial map was created using layers downloaded from GeoDASH (The Bangladesh Geospatial Data Sharing Platform) website on ArcGIS Desktop (Esri Inc.,109 Redlands, California, United States) licensed to King's College London.

### 2.4 Statistical analysis

Descriptive and inferential statistics were used to analyze different mutations and their correlation with different categorical variables. For correlation, we used one-way ANOVA

analysis of variance using SPSS Statistics 25 (IBM, Armonk, New York) licensed to King's College London.

## 3. Results and discussion

### 3.1 Phylogenetic analysis of SARS-CoV-2 circulating in Bangladesh

To understand the SARS-CoV-2 viral transmission in Bangladesh, phylogenetic analysis was performed based on the selected 371 complete viral genomes from different regions of Bangladesh along with selected 467 globally submitted sequences from 42 countries of 6 continents as shown in Figure 1. This represents an overall clade distribution of all global sequences along with sequences from Bangladeshi isolates. Analysis of Figure 1 depicts the GR clade, which is found predominant in Bangladesh. About 86% of the sequences are grouped to this clade followed by G and GH with ~6%, respectively. Similar clade distribution has been reported in SARS-CoV-2 isolates, originating from European countries (Hamed et al., 2020). We also compared the sequence data among different regions of Bangladesh, looking at the district-wise distribution of clades as shown in Figure 2. As can be seen in Figure 2, the sequences from all districts largely clustered with that of GR clade except sequences from Chittagong.

We also observed that the major clade distribution of SARS-CoV-2 isolates from Chittagong is consists of both GR and GH (Figure 2). The GH clade was predominantly found in Saudi Arabia

(Mercatelli and Giorgi, 2020), which is clustered together with Chittagong as shown in Figure 1. Based on this evidence, it is highly likely that the introduction of GH clade in Bangladesh could be of Middle-Eastern origin.

### 3.2 Mutations analysis in Bangladeshi SARS-CoV-2 genomic sequences

#### 3.2.1 Prevalence of global common mutations

Our analyses revealed a total of 4604 mutations observed among 371 Bangladeshi SARS-CoV-2 genomic sequences that include 2862 missense, 1192 synonymous, 25 insertion/deletions and 525 other mutations (Ambiguous mutation and nucleotide mutation in UTR region) (See Table 1). Among the identified mutations, 28881G>A & 28882G>A (R203K; N protein), 23403A>G (D614G; S glycoprotein), 28883G>C (G204R; N protein), 14408C>T (P4715L; nsp12), and 1163A>T (I300F; nsp2) were the most frequently occurring common mutations found in Bangladesh with a frequency of 650, 365, 325, 304 and 243, respectively (Figure 3). Nucleotide mutations of 28881G>A and 28882G>A resulted in R203K due to codon degeneracy. Notably, no particular mutations occurred at any specific time rather they have been observed over the whole period of disease incidence. However, this explanation is predicted and needs to be investigated in detail in the future. The RNA-dependent RNA polymerase (i.e. nsp12) is essential for replication and transcription of the viral RNA genome. The observed mutation P4715L in nsp12 was also found in most of the US states (28 out of 31 states from where the sequences were deposited). The same mutation was prevalent in European countries like Spain, France etc. This alteration could affect the pathogenesis triggered by antibody escape variants with epitope loss (Banerjee et al., 2020; Gupta and Mandal, 2020). In a separate study, it has been reported

that the G614 type might have originated either in Europe or China (Korber et al., 2020b). They also reported that the original Wuhan D614 form was also predominant in Asian samples. Meanwhile, the G614 form had clearly been established and started expanding in countries outside of China. We noticed that 98% of genomes from Bangladesh have D614G mutation, which is also dominant in the world. However, the average number of mutations per ORF is varied among genomes containing D614 and G614 that we have studied (n=838) as revealed by Table 2. The average number of mutations per genome between D614 and G614 is 6.12 and 10.68, respectively. The average number of mutations in ORF1ab, S, ORF3a, M and N of genomes having mutated G614 (n=626) are significantly higher ($p \leq 0.001$) than those having the original D614 (n=212) in S glycoprotein (See Table 2 and Figure 1). Interestingly, the average mutation number is declined in ORF8, having G614 mutation ($p<0.001$). This correlation indicates that the genomes containing D614G mutation also bear more mutations which is aligned with the various reports on the link between the transmission and pathogenesis of SARS-CoV-2 (Grubaugh et al., 2020; Korber et al., 2020a; Zhang et al., 2020), and this mutation increases cell entry and transduction due to resistance to proteolytic cleavage (Daniloski et al., 2021; Ozono et al., 2021). However, without further experimental evidence this is not conclusive. Meanwhile, R203K and G204R mutations in N protein were previously reported in Indian, Spanish, Italian, and French samples (Koyama et al., 2020; Maitra et al., 2020). These mutations are located in the site of the SR-rich region which has been reported to be intrinsically disordered (Chang et al., 2014). This region further incorporates a few phosphorylation sites (Surjit et al., 2005), including the GSK3 (Glycogen synthase kinase 3) phosphorylation at Ser202 and a CDK (Cyclin-dependent kinase) phosphorylation site at Ser206 which are located close to the position of this mutation. The 'SRGTS' (202-206) and 'SPAR' (206-209) sequence motifs

are dependent on GSK3 and CDK phosphorylation motifs, respectively. Other variations 28881G>A and 28882G>A together convert polar to non-polar amino acid (R203K) and 28883G>C variation converts nonpolar to polar amino acid (G204R) (See Figure 3).

### 3.2.2 Unique mutations (UM) existing in Bangladeshi SARS-CoV-2 isolates

We observed 34 unique shifts from the different proteins of SARS-CoV-2 isolated from Bangladesh (Table 3). Surprisingly, most of these UMs were found to be in patients in a specific area with a few exceptions. For example, some UMs as shown in Table 3 are only found in Barisal, Chandpur, Chittagong, Dhaka, Jessore, Moulvibazar, Mymensingh, and Rangpur districts. Interestingly, some UMs are concurrently present in multiple cities. These changes in amino acids might have occurred due to rapid mutation events and/or recombination with existing CoVs in the human body (Hassan et al., 2020). The circulation of a high number of these UMs in different cities indicates the possible emergence of community transmission in the Bangladeshi population (Samyuktha and Kumar, 2020).

We observed 2 UMs in nsp1 (See Table 3), but the location of these amino acids was not in the KH domain (K164 and H165 of nsp1), which binds with the 40s subunit of ribosome (Thoms et al., 2020). However, nsp1 acts as a primary virulence factor in SARS-CoV-2 infection, and mutation in this protein could affect the structure and functional properties, thereby altering its virulence properties. Similar to nsp1, 2 UMs were seen in nsp2, but their effects on host cells were merely reported in the literature. Since nsp2 interacts with the host proteins and disrupts the host cell survival signaling pathway (Bianchi et al., 2020), any mutation in nsp2 may play a crucial role in SARS-CoV infections. In general, mutations of nsp3 are

responsible for affecting the virus assembly and hence their replication. One would assume that this is due to the disruption of the replicase polyprotein processing into nsps. These nsps assemble with cellular membranes and facilitate virus replication. Among 34 amino acid mutations, 9 mutations were located in nsp-3. The UMs found in nsp3 might have some significant impacts in the viral pathogenesis. Firstly, the UMs (A889V and V843F) were found in the main domain of nsp3 that is important for processing endopeptidases from coronaviruses. Secondly, some UMs (e.g., G1691C, A602S and L373M) are found in the topological (cytoplasmic) domain. Thirdly, we observed one UM (L373M) in the ADP-ribose-1′-phosphatase (ADRP) or (Macro) domain of nsp3. It has been shown that mutation of the ADRP domains does not diminish virus replication in mice, but reduces the production of the cytokine IL-6, which is an important proinflammatory molecule (Mielech et al., 2015). However, we could not find any mutation in the active sites and zinc finger motif, attributing normal catalytic activity of nsp3. Overall, our opinion is that the PL-PRO domain is important for the development of antiviral drugs and the actual role of this enzyme is to cleave of several nsps from the polyprotein pp1a and also the biogenesis of the SARS-CoV replicase complex is yet to be explored. It can be postulated that the proteins like nsp3, nsp4 and nsp6 through their transmembrane domains, are involved in the replicative and transcription complex (Coppée et al., 2020). In this study, we observed only two UMs in nsp4 and nsp6, respectively. Meanwhile, nsp5 encodes 3C-like proteinase that cleaves the C-terminus of the replicase polyprotein at 11 sites (Lokhande et al., 2020; Roe et al., 2021). One UM that we found in nsp5 did not fall in its active sites (3304 and 3408 in ORF1ab) warrants further investigation with a large number of sequence datasets.

## 4. Conclusion

The present global outbreak of COVID-19, caused by SARS-CoV-2, has already taken more than 2.5 million lives. Bangladeshi citizens are highly vulnerable to COVID-19 as evident by a number of circulating variants in different regions of this country. To combat this deadly disease, we need a better understanding of the pathobiology of this deadly virus. Hence, it is essential to minimize the translational gap between viral genomic information and its clinical consequences for developing effective therapeutic strategies. In this study, we have attempted to explore genomic variations of Bangladeshi SARS-CoV-2 viral isolates while comparing them with a large cohort of global isolates. Our analyses will facilitate the understanding of the origin, mutation patterns, and their possible effect on viral pathogenicity. We have addressed the importance of the variations in the viral genomes and their necessity for therapeutic interventions. The unique insights from this study will undoubtedly be supportive for a better understanding of the molecular mechanism of SARS-CoV-2 which would help to understand pathophysiology of the virus in highly populated countries like Bangladesh.

## Author statement

MMH, RD, SMM and AMN designed the study. MMH, RD, ZS and AMN reviewed the literature, extracted the data and checked the extracted data. MMH, RD, MR, MHH, NMH, AS, and MHR planned and conducted the analysis. MMH, RD, SMM, ZS and AMN wrote the first draft and all authors reviewed and approved the final manuscript paper; approved the final version and agreed to be accountable for the work.

**Declaration of interests**

Not applicable.

## References:

Ahmed, T.S., Naser, I.B., Faruque, S.M., 2020. In silico comparative genomics of SARS-CoV-2 to determine the source and diversity of the pathogen in Bangladesh. bioRxiv, https://doi.org/10.1101/2020.1107.1120.212563.

Astuti, I., 2020. Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2): An overview of viral structure and host response. Diabetes & Metabolic Syndrome: Clinical Research & Reviews.

Banerjee, S., Seal, S., Dey, R., Mondal, K.K., Bhattacharjee, P., 2020. Mutational spectra of SARS- CoV- 2 orf1ab polyprotein and signature mutations in the United States of America. Journal of Medical Virology.

Bianchi, M., Benvenuto, D., Ciccozzi, M., Pascarella, S., 2020. Mutational screening of the proteome of Sars-Cov-2 isolates: mutability of ORF3a, Nucleocapsid and Nsp2 proteins. Preprints 10.20944/preprints202007.200049.v202001.

Chang, C.-k., Hou, M.-H., Chang, C.-F., Hsiao, C.-D., Huang, T.-h., 2014. The SARS coronavirus nucleocapsid protein–forms and functions. Antiviral research 103, 39-50.

Chen, Y., Liu, Q., Guo, D., 2020. Emerging coronaviruses: genome structure, replication, and pathogenesis. Journal of medical virology 92(4), 418-423.

Cleemput, S., Dumon, W., Fonseca, V., Abdool Karim, W., Giovanetti, M., Alcantara, L.C., Deforche, K., De Oliveira, T., 2020. Genome Detective Coronavirus Typing Tool for rapid identification and characterization of novel coronavirus genomes. Bioinformatics 36(11), 3552-3555.

Coppée, F., Lechien, J., Declèves, A.-E., Tafforeau, L., Saussez, S., 2020. SARS-CoV-2: virus mutations in specific European populations. New Microbes and New Infections, 100696.

Daniloski, Z., Jordan, T.X., Ilmain, J.K., Guo, X., Bhabha, G., Sanjana, N.E., 2021. The Spike D614G mutation increases SARS-CoV-2 infection of multiple human cell types. Elife 10, e65365.

Dumonteil, E., Herrera, C., 2020. Polymorphism and selection pressure of SARS-CoV-2 vaccine and diagnostic antigens: implications for immune evasion and serologic diagnostic performance. Pathogens 9(7), 584.

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research 32(5), 1792-1797.

Forster, P., Forster, L., Renfrew, C., Forster, M., 2020. Phylogenetic network analysis of SARS-CoV-2 genomes. Proceedings of the National Academy of Sciences 117(17), 9241-9243.

GISAID, 2020. Clade and lineage nomenclature aids in genomic epidemiology studies of active hCoV-19 viruses. Vol. 2020.

Gong, Y.-N., Tsao, K.-C., Hsiao, M.-J., Huang, C.-G., Huang, P.-N., Huang, P.-W., Lee, K.-M., Liu, Y.-C., Yang, S.-L., Kuo, R.-L., 2020. SARS-CoV-2 genomic surveillance in Taiwan revealed novel ORF8-deletion mutant and clade possibly associated with infections in Middle East. Emerging microbes & infections 9(1), 1457-1466.

Grubaugh, N.D., Hanage, W.P., Rasmussen, A.L., 2020. Making sense of mutation: what D614G means for the COVID-19 pandemic remains unclear. Cell 102(4), 794-795.

Gupta, A.M., Mandal, S., 2020. Non-synonymous Mutations of SARS-Cov-2 Leads Epitope Loss and Segregates its Varaints. Research Square, 10.21203/rs.21203.rs-29581/v21201.

Hamed, S.M., Elkhatib, W.F., Khairallah, A.S., Noreddin, A.M., 2020. Global dynamics of SARS-CoV-2 clades and their relation to COVID-19 epidemiology.

Hasan, S., Khan, S., Ahsan, G.U., Hossain, M.M., 2020. Genome Analysis of SARS-CoV-2 Isolate from Bangladesh. BioRxiv, https://doi.org/10.1101/2020.1105.1113.094441.

Hassan, S.S., Choudhury, P.P., Basu, P., Jana, S.S., 2020. Molecular conservation and Differential mutation on ORF3a gene in Indian SARS-CoV2 genomes. Genomics 112, 3226-3237.

Islam, O., Al-emran, H., Hasan, M., Anwar, A., Jahid, M., Hossain, M., 2020. Emergence of European and North American mutant variants of SARS-CoV-2 in South-East Asia. Transboundary and Emerging Diseases 00(Jul), 1-9.

Khailany, R.A., Safdar, M., Ozaslan, M., 2020. Genomic characterization of a novel SARS-CoV-2. Gene reports, 100682.

Kim, D., Lee, J.-Y., Yang, J.-S., Kim, J.W., Kim, V.N., Chang, H., 2020. The architecture of SARS-CoV-2 transcriptome. Cell.

Korber, B., Fischer, W., Gnanakaran, S.G., Yoon, H., Theiler, J., Abfalterer, W., Foley, B., Giorgi, E.E., Bhattacharya, T., Parker, M.D., 2020a. Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. bioRxiv, https://doi.org/10.1101/2020.1104.1129.069054.

Korber, B., Fischer, W.M., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., Hengartner, N., Giorgi, E.E., Bhattacharya, T., Foley, B., 2020b. Tracking changes in SARS-CoV-2 Spike: evidence that D614G increases infectivity of the COVID-19 virus. Cell 182(4), 812-827.e819.

Koyama, T., Platt, D., Parida, L., 2020. Variant analysis of SARS-CoV-2 genomes. Bulletin of the World Health Organization 98(7), 495.

Kumar, S., Stecher, G., Li, M., Knyaz, C., Tamura, K., 2018. MEGA X: molecular evolutionary genetics analysis across computing platforms. Molecular biology and evolution 35(6), 1547-1549.

Letunic, I., Bork, P., 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic acids research 47(W1), W256-W259.

Lokhande, K.B., Doiphode, S., Vyas, R., Swamy, K.V., 2020. Molecular docking and simulation studies on SARS-CoV-2 Mpro reveals Mitoxantrone, Leucovorin, Birinapant, and Dynasore as potent drugs against COVID-19. Journal of Biomolecular Structure and Dynamics, 1-12.

Maitra, A., Sarkar, M.C., Raheja, H., Biswas, N.K., Chakraborti, S., Singh, A.K., Ghosh, S., Sarkar, S., Patra, S., Mondal, R.K., 2020. Mutations in SARS-CoV-2 viral RNA identified in Eastern India: Possible implications for the ongoing outbreak in India and impact on viral structure and host susceptibility. Journal of Biosciences 45(1), 76.

Mercatelli, D., Giorgi, F.M., 2020. Geographic and Genomic Distribution of SARS-CoV-2 Mutations. Frontiers in Microbiology 11, 1800.

Mielech, A.M., Deng, X., Chen, Y., Kindler, E., Wheeler, D.L., Mesecar, A.D., Thiel, V., Perlman, S., Baker, S.C., 2015. Murine coronavirus ubiquitin-like domain is important for papain-like protease stability and viral pathogenesis. Journal of virology 89(9), 4907-4917.

O'Meara, M.J., Guo, J.Z., Swaney, D.L., Tummino, T.A., Hüttenhain, R., 2020. A SARS-CoV-2-Human Protein-Protein Interaction Map Reveals Drug Targets and Potential Drug-Repurposing. bioRxiv, https://doi.org/10.1101/2020.1103.1122.002386.

Ozono, S., Zhang, Y., Ode, H., Sano, K., Tan, T.S., Imai, K., Miyoshi, K., Kishigami, S., Ueno, T., Iwatani, Y., 2021. SARS-CoV-2 D614G spike mutation increases entry efficiency with enhanced ACE2-binding affinity. Nature communications 12(1), 1-9.

Parvez, M.S.A., Rahman, M.M., Morshed, M.N., Rahman, D., Anwar, S., Hosen, M.J., 2020. Genetic analysis of SARS-CoV-2 isolates collected from Bangladesh: insights into the origin, mutation spectrum, and possible pathomechanism. bioRxiv, https://doi.org/10.1101/2020.1106.1107.138800.

Raghav, S., Ghosh, A., Turuk, J., Kumar, S., Jha, A., Madhulika, S., Priyadarshini, M., Biswas, V.K., Shyamli, P.S., Singh, B., 2020. Analysis of Indian SARS-CoV-2 Genomes Reveals Prevalence of D614G Mutation in Spike Protein Predicting an Increase in Interaction With TMPRSS2 and Virus Infectivity. Frontiers in Microbiology 11.

Roe, M.K., Junod, N.A., Young, A.R., Beachboard, D.C., Stobart, C.C., 2021. Targeting novel structural and functional features of coronavirus protease nsp5 (3CLpro, Mpro) in the age of COVID-19. Journal of General Virology, 001558.

Saitou, N., Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular biology and evolution 4(4), 406-425.

Samyuktha, V., Kumar, V.N., 2020. Emergence of RBD and D614G Mutations in Spike Protein: An Insight from Indian SARS-CoV-2 Genome Analysis. Preprints, doi: 10.20944/preprints202006.200032.v202001.

Surjit, M., Kumar, R., Mishra, R.N., Reddy, M.K., Chow, V.T., Lal, S.K., 2005. The severe acute respiratory syndrome coronavirus nucleocapsid protein is phosphorylated and localizes in the cytoplasm by 14-3-3-mediated translocation. Journal of virology 79(17), 11476-11486.

Thoms, M., Buschauer, R., Ameismeier, M., Koepke, L., Denk, T., Hirschenberger, M., Kratzat, H., Hayn, M., Mackens-Kiani, T., Cheng, J., 2020. Structural basis for translational shutdown and immune evasion by the Nsp1 protein of SARS-CoV-2. bioRxiv, https://doi.org/10.1101/2020.1105.1118.102467.

Toyoshima, Y., Nemoto, K., Matsumoto, S., Nakamura, Y., Kiyotani, K., 2020. SARS-CoV-2 genomic variations associated with mortality rate of COVID-19. Journal of human genetics, 1-8.

Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Song, Z.-G., Hu, Y., Tao, Z.-W., Tian, J.-H., Pei, Y.-Y., 2020. A new coronavirus associated with human respiratory disease in China. Nature 579(7798), 265-269.

Yoshimoto, F.K., 2020. The Proteins of Severe Acute Respiratory Syndrome Coronavirus-2 (SARS CoV-2 or n-COV19), the Cause of COVID-19. The Protein Journal, 1.

Zhang, L., Jackson, C.B., Mou, H., Ojha, A., Rangarajan, E.S., Izard, T., Farzan, M., Choe, H., 2020. The D614G mutation in the SARS-CoV-2 spike protein reduces S1 shedding and increases infectivity. bioRxiv, https://doi.org/10.1101/2020.1106.1112.148726.



## Figure legends:

Figure 1: Phylogenetic analysis of 371 Bangladeshi SARS-CoV-2 genomes with 467 representative sequences from 42 different countries worldwide. Sequence having the original D614 in Spike glycoprotein mutation is 'shaded cyan' while G614 in yellow. Outside the main tree and labels, there are three different annotation panels sequentially to show the sequences from Bangladesh (Orange Star), followed by the burden of mutation in different ORFs of SARS-CoV-2 in the outermost panel as 'heatmap' and finally, GISAID clades as a 'color strip'. The reference sequence, Wuhan-Hu-1 (Accession number MN908947), is indicated with ➞□.

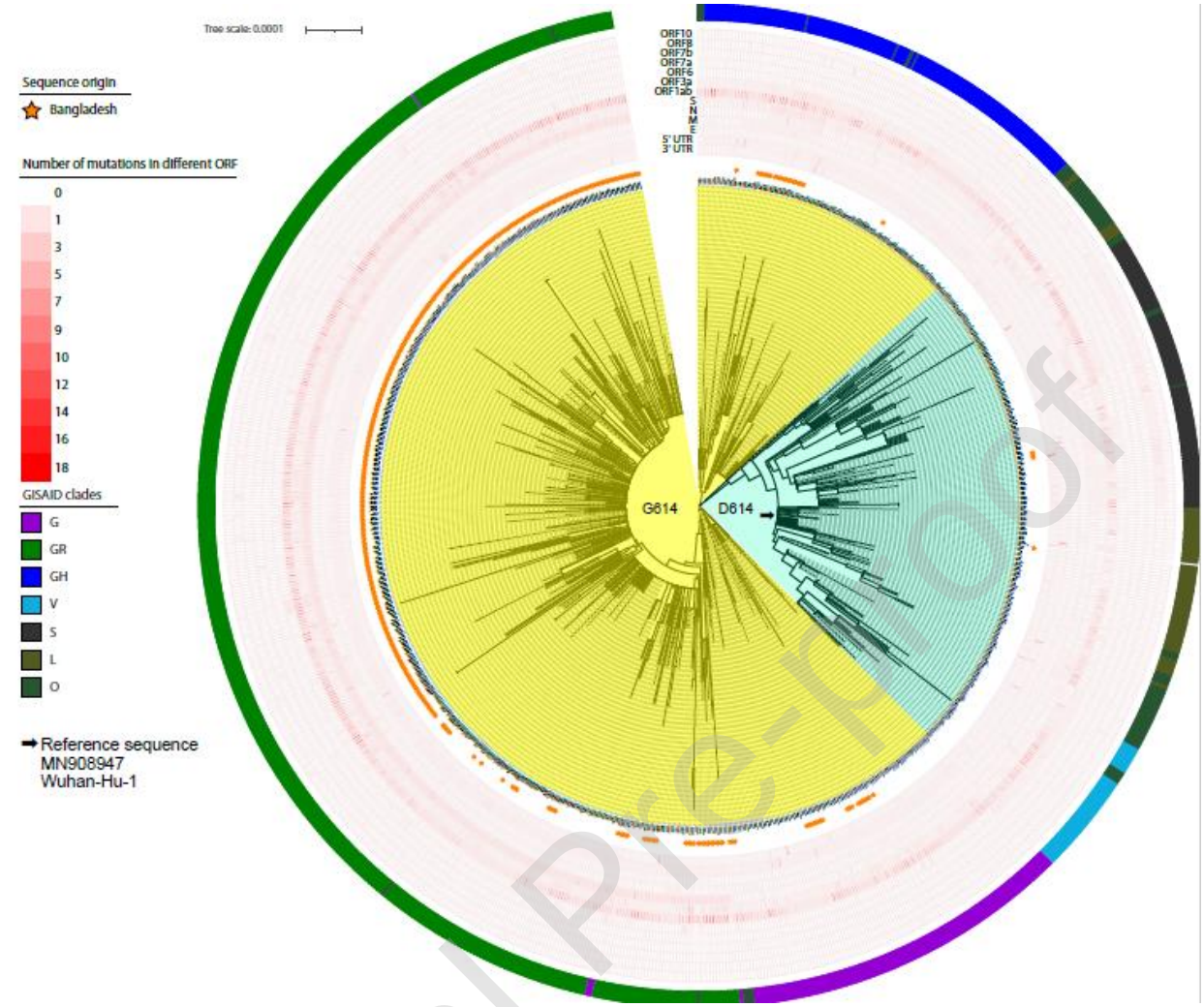



Figure 2: District-wise distribution of GISAID clades among circulating SARS-CoV-2 strains in Bangladesh. Please be noted that district information of 72 sequences out of 371 deposited from Bangladesh is not available from respective metadata. Districts from where samples were collected for sequencing are shaded pink and the number of the sequences from each district is proportional to the circumference of the piecharts. The spatial map was created using layers downloaded from GeoDASH (The Bangladesh Geospatial Data Sharing Platform) website on ArcGIS Desktop (Esri Inc.,109 Redlands, California, United States) licensed to King's College London.

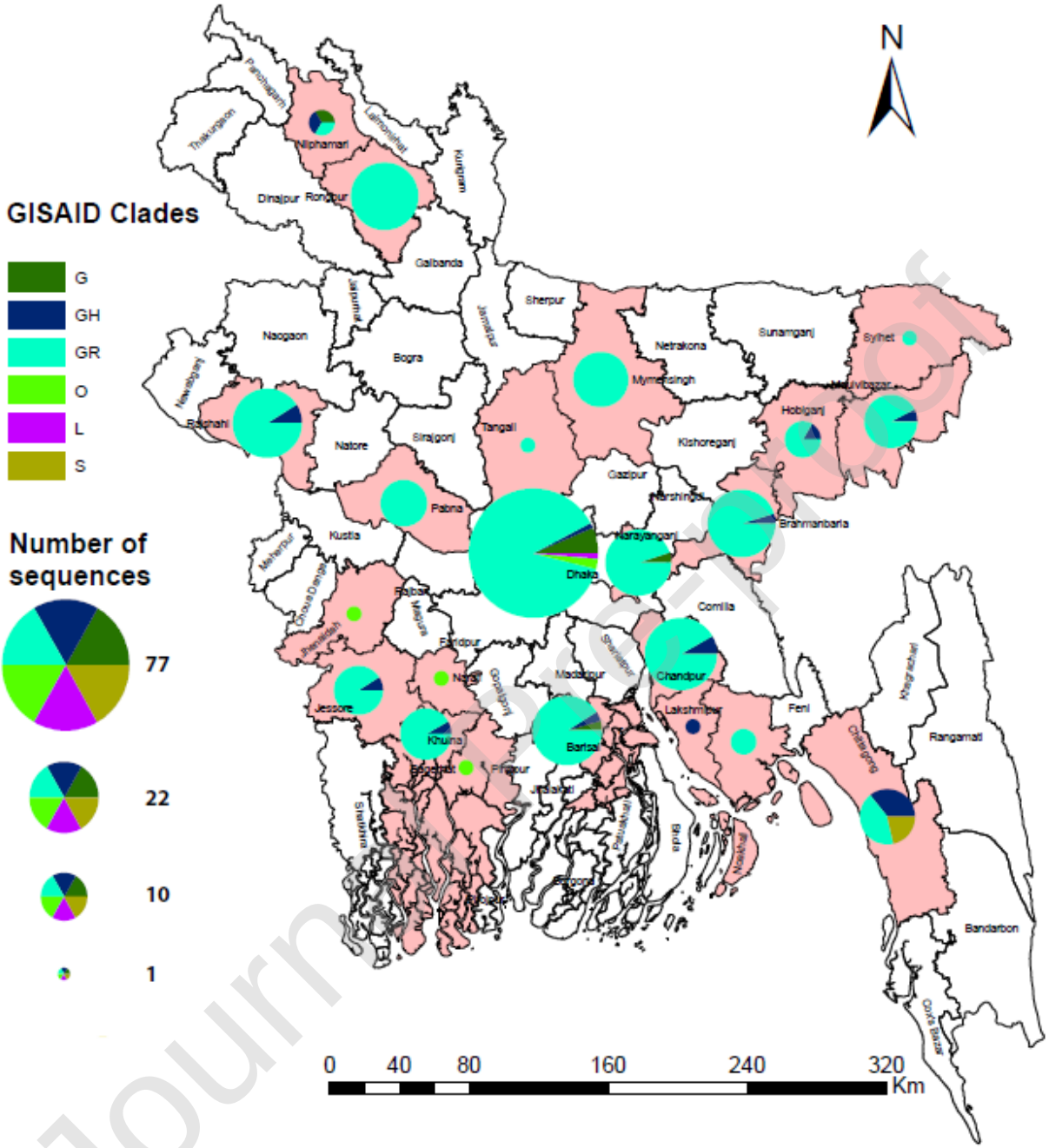
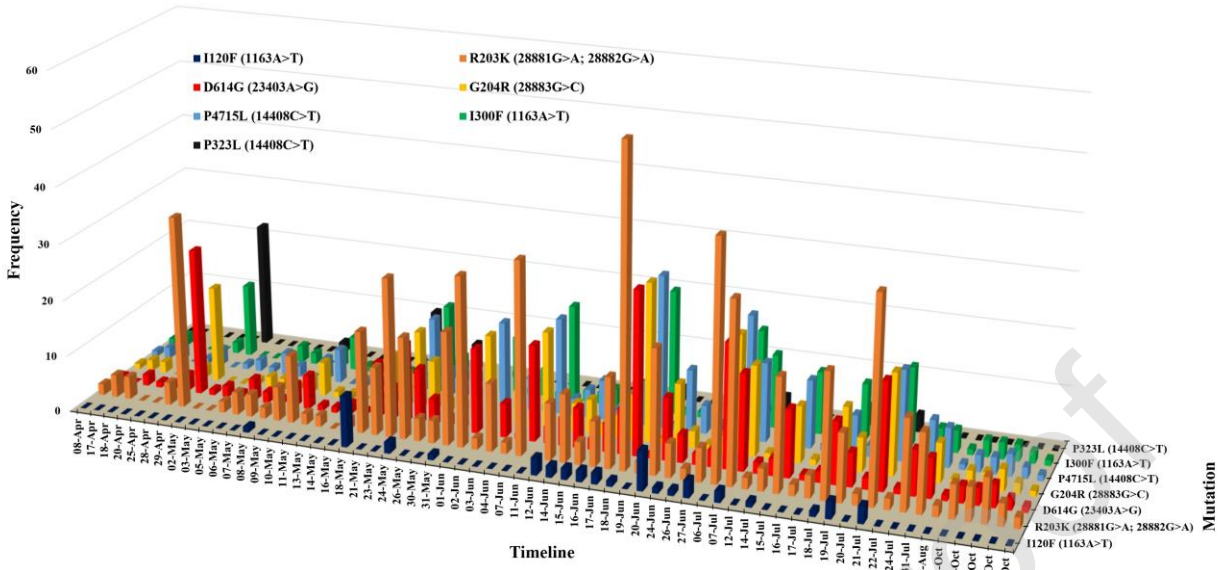



Figure 3: Frequency of common mutations observed in Bangladesh over time. The x-axis of the plot shows the timeline between April 2020 and October 2020 whereas the Y-axis depicts the number of mutations reported from the sequences of the same collection date.



**Tables**

**Table 1:** Number of common gene variants among 371 SARS-CoV-2 genomes from Bangladesh.

| Genome segment | Missense mutation | Synonymous mutation | Insertion Deletion | Others | Total |
|---|---|---|---|---|---|
| ORF1ab | 1141 | 898 | 1 | 4 | 2044 |
| S | 493 | 118 | 3 | 1 | 615 |
| ORF3a | 89 | 49 | 7 | 0 | 145 |
| E | 3 | 1 | 0 | 0 | 4 |
| M | 20 | 55 | 0 | 1 | 76 |
| ORF6 | 7 | 2 | 0 | 0 | 9 |
| ORF7a | 11 | 7 | 0 | 0 | 18 |
| ORF7b | 7 | 1 | 0 | 0 | 8 |
| ORF8 | 30 | 16 | 2 | 0 | 48 |
| ORF10 | 10 | 5 | 0 | 0 | 15 |
| N | 1051 | 40 | 2 | 1 | 1094 |
| Intergenic | 0 | 0 | 2 | 13 | 15 |
| 5' UTR | 0 | 0 | 4 | 414 | 418 |
| 3' UTR | 0 | 0 | 4 | 91 | 95 |
| Total | 2862 | 1192 | 25 | 525 | 4604 |

**Table 2:** Correlation of the average number of mutations per genome among different genomic segments with D614G mutation.

| Genome segment | Average number of mutations per genome | | Change (%) | *p* value |
|---|---|---|---|---|
| | D614 (wild; n=212) | G614 (n=626) | | |
| 5' UTR | 0.34 | 1.12 | 70 | <0.001 |
| ORF1ab | 3.20 | 4.75 | 33 | <0.001 |
| S | 0.51 | 1.50 | 66 | <0.001 |
| ORF3a | 0.40 | 0.41 | 2 | 0.794 |
| E | 0.00 | 0.01 | 100 | 0.501 |
| M | 0.07 | 0.17 | 59 | <0.001 |
| ORF6 | 0.01 | 0.03 | 67 | 0.134 |
| ORF7a | 0.02 | 0.04 | 50 | 0.175 |
| ORF7b | 0.01 | 0.01 | 0 | 0.585 |
| ORF8 | 0.45 | 0.10 | -350 | <0.001 |
| N | 0.58 | 2.24 | 74 | <0.001 |
| ORF10 | 0.01 | 0.03 | 67 | 0.092 |
| 3' UTR | 0.52 | 0.22 | -136 | <0.001 |
| Total | 6.12 | 10.68 | 43 | |

**Table 3:** Unique mutations in 371 SARS-CoV-2 genomes from Bangladesh.

| UM | Gene/ORF | District | Remarks |
|---|---|---|---|
| ***UMs localized to specific district*** | | | |
| E194Q | NS3 | Barisal | Negatively charged to Glu neutral Gln |
| Q62E | NS7a | Barisal | Neutral Gln to negatively charged Glu |
| N377D | NSP2 | Chandpur | Polar Asn to acidic Asp |
| G773A | NSP3 | Chandpur | Same type of change |
| Y660F | Spike | Chandpur | Aromatic Tyr to to aromatic Phe |
| N11D | N | Chittagong | Polar Asn to acidic Asp |
| N39Y | NS6 | Chittagong | Polar aliphatic Asn to aromatic Tyr |
| A602S | NSP3 | Chittagong | Hydrophobic, non-polar Ala to hydrophilic Ser |
| V56A | NSP1 | Dhaka | Hydrophobic Val to simple, non-polar Ala |
| D85E | NSP4 | Dhaka | Both are same type of amino acid |
| N133B | NSP5 | Dhaka | Ambigious |
| L22I | NSP6 | Jessore | Both are same type of amino acids |
| K84T | NSP9 | Jessore | Positively charged ε-amino group containing Lys to hydroxyl group containing Thr |
| T80I | NS8 | Moulvibazar | Hydroxyl group containing Thr to hydrophobic Ile |
| E242A | NS3 | Mymensingh | Acidic Glu to hydrophobic Ala |
| G42V | NS7a | Rangpur | Glycine convert to hydrophobic valine |
| ***UMs shared between districts*** | | | |
| P38R | NS8 | Pabna, Chandpur, Brahmanbaria | Phenylalanine convert to Arginine which contains positively charged group. |
| D179N | NSP16 | Khulna, Rajshahi | Acidic Asp to polar aliphatic Asn |
| V469A | NSP2 | Mymensingh, Barisal, Chandpur | Hydrophobic Val to another hydrophobic, non-polar Ala |
| L373M | NSP3 | Brahmanbaria, Moulvibazar, Rajshahi | Hydrophobic side chain containing Leu to sulfur containing Met |
| V120L | NSP6 | Habiganj, Pabna | Valine to hydrophobic side chain containing Leu |
| ***District level location is not given (NG) in the metadata/dataset*** | | | |
| H145N | N | NG | Polar, basic His to polar Asn |
| P104R | NS3 | NG | Phenylalanine convert to positively charged Arg |
| V121D | NSP1 | NG | Hydrophobic Val to acidic Asp |

| | | | |
|---|---|---|---|
| V31E | NSP12 | NG | Hydrophobic Val to negatively charged Glu |
| F217I | NSP14 | NG | Aromatic Phe to hydrophobic Ile |
| D1774H | NSP3 | NG | Acidic Asp to polar, basic His |
| E120K | NSP3 | NG | Negatively charged Glu to positively charged ε-amino group containing Lys |
| D1108N | NSP3 | NG | Acidic Asp to polar Asn |
| V843F | NSP3 | NG | Hydrophobic Val to aromatic Phe |
| A889V | NSP3 | NG | Hydrophobic, non-polar Ala to another hydrophobic Val |
| G1691C | NSP3 | NG | Simple Gly to sulfur-containing, acidic Cys |
| E42A | NSP4 | NG | Negatively charged Glu to hydrophobic, non-polar Ala |
| I569S | Spike | NG | Hydrophobic side chain containing amino acid Ile to polar Ser |