Study of Transmission, Mutation Rate, Genetic Variants, Non-synonymous Mutation, And Genomic Phylogeny of 263 SARS-CoV-2 Sequenced by Genomic Research Lab, BCSIR

Submitted by-

Genomic Research Laboratory
Bangladesh Council of Scientific and Industrial Research
Ministry of Science and Technology
Government of the People's Republic of Bangladesh
Phone: 0088029665546
Cell: +880-1712201504
E-mail: k2salim@bcsir.gov.bd

Study of Transmission, Mutation Rate, Genetic Variants, Non-synonymous Mutation, And Genomic Phylogeny of 263 SARS-CoV-2 Sequenced By Genomic Research Lab, BCSIR.

Highlights

1. We have sequenced 263 whole-genomes of SARS-CoV-2 strains collected from different divisions of Bangladesh.
2. The only dominated variant "G614" (due to replacement of aspartic acid at 614 number by glycine in spike protein) observed in 100% of cases
3. Although no distinct variants based on geography were detected in this study, a total of 3 distinct clusters with 9 subclusters of SARS-CoV-2 were detected
4. 243 out of 263 SARS-CoV-2 are belonging to GR Clade, 12 out of 262 are GH clade, 3 out of 263 G clade, and only 1 in O clade were examined.
5. A total of 737 mutation sites across the SARS-CoV-2 genome and 358 non-synonymous amino acid substitutions were detected.
6. The mutation rate of SARS-CoV-2 is estimated to be 24.64 nucleotide substitutions per year, indicating SARS-CoV-2 accumulate around two nucleotide substitutions per month**.**
7. Global mutation event per sample is 7.23 while, it is 12.6 in Bangladesh, indicating SARS-CoV changing rapidly in Bangladesh than the rest of the world
8. A total of 53 non-synonymous amino acids substitution and 103 nucleotide mutations in spike protein were detected
9. We detected five unique variants based on non-synonymous amino acid substitutions in spike protein relative to the global SARS-CoV-2 strains.
10. We examined 4 recurrent mutations in 100% of cases, these include 241C>T, 3037C>T, 14408C>T, and 23403A>G

Summary

Genomic mutation of the SARS-CoV-2 may impact the viral adaptation to the local environment, their transmission, disease manifestation, and the effectiveness of existing treatment and vaccination. The objectives of this study were to characterize genomic variations due to nucleotide substitution, non-synonymous amino acid substitutions, mutation events per samples, mutation rate, and overall characteristic SARS-CoV-2 collected from eight divisions of Bangladesh. To investigate the genetic diversity, a total of 263 genomes of SARS-CoV-2 strains were sequenced by genomic Research Lab, BCSIR, with sampling dates between the 7th of May 2020 and the 31st of July 2020 were analyzed. To date, a total of 737 nucleotide mutations located along the entire genome resulting in non-synonymous 358 amino acid substitutions in 25 different proteins were detected. The nucleotide mutation rate is

estimated to be 24.6 substitutions per year. The highest nucleotide mutations were observed at 112 positions of non-structural protein papain-like protease (nsp3), which led to the 62 non-synonymous amino acid substitutions. Among the structural proteins, the highest mutations were observed at 101 positions of spike proteins resulting in 53 non-synonymous amino acid substitution. The only dominated variant "G614" (due to the change of aspartic acid at 614 number to glycine in spike protein) found in 100% of cases is circulating across the country with co-evolving other variants including L323 (100%) in RNA dependent RNA polymerase (RdRp), K203 and R204 (93.5%) in nucleocapsid, and F120 (83%) in NSP2.

Background

Bats naturally host most of the coronaviruses, and it is postulated that coronaviruses jumped to humans from the bat reservoir [1, 2]. Only six coronaviruses have been known to cause infection in humans until late 2019; these include SARS-CoV, HCoV-OC43, CoV-HKU1, HCoV-NL63, HCoV-229E, and MERS-CoV. A seventh human coronavirus, named SARS-CoV-2, which is closely related to Bat coronavirus (96%) major cause of pandemic in the modern history of humans. Coronaviruses were mostly associated with mild symptoms [3], with two marked exceptions that caused a major epidemic. The first coronavirus mediated outbreak that caused Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV) occurred in 2002, with infected over 8,000 cases and killed 800 [4] and not known to circulate since 2003. The other coronavirus that caused Middle East Respiratory Syndrome (MERS-CoV) causing sporadic infections, mostly in the Arabian peninsula in 2012, a less infective but highly lethal virus, which infected 2,294 laboratory-confirmed cases and killed 858 [4, 5] including 38 deaths in South Korea [6].

As of the 24th of August 2020, the SARS-CoV-2 virus causative agent of COVID-19 has spread to over 216 countries or regions and had caused over 23,420,418 cases with 808,676 deaths worldwide [7]. The first COVID-19 case was reported in Bangladesh on the 8th of March 2020, and from there on, more than 294,598 confirmed cases had been declared with 3,941 death [8]. As COVID-19 is responsible for enormous human casualties and economic loss posing a serious threat to Bangladesh and globally, an understanding of the ongoing situation and the development of strategies to contain the virus's spread are urgently needed. The worldwide sequencing of the SARS-CoV-2 genome has continued unabated to track the geographic movement and evolution of the virus. To date, there are over 84,000 sequences publicly shared from half the countries in the world.

From Bangladesh 325 whole-genome sequencing data have already submitted to GISAID, among them Genomic Research Lab, BCSIR has contributed 263 SARS-CoV-2 whole-genome sequencing data, which comprises 81% of all sequenced data generated in Bangladesh. In this study, SARS-CoV-2 isolates were collected from across Bangladesh based on the demography of eight representative divisions (**Fig.1a**). The age of the COVID-19 patients ranging from 8 to 95 years old with age interval

87. Most of the samples, 48%, belong to the age group 20–40, while 35% were between 41-60 and 9 % was between 61 to 80 (**Fig. 1b**). Viral isolates were collected from 69% of male patients and 31% of females.

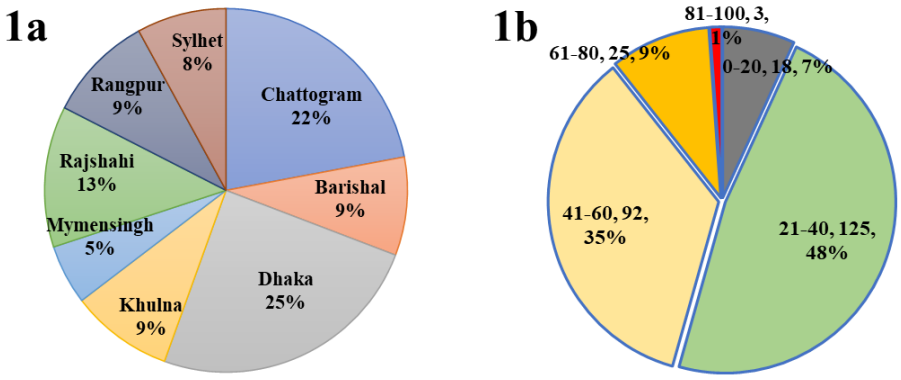

Figure 1: Geographic and age distribution of SARS-CoV-2 host

Transmission and mutation rate

We have shown evidence of multiple introductions of SARS-CoV-2 into Bangladesh from several countries, including Germany, Brazil, Vietnam, India, and Saudi Arabia (**Fig.2**). When a different variant of viral populations infecting patients, they may contribute to differences in clinical outcomes among the patients. It is, therefore, demand the constant surveillance of the newly arisen viral variant by continuously monitoring the sequences of SARS-CoV-2 from different patients. Usually, a higher rate of mutation occurred in typical RNA viruses that result in a different version of the viral population with diverse genomes. It has been estimated for coronavirus that the average evolutionary rate is approximately 10^{-4} nucleotide substitutions per site per year [3]. Based on the 263 sequencing data, we have calculated the mutation rate of SARS-CoV-2 in Bangladesh is 24.6 substitution per year (**Fig.3**).

Figure 2: Transmission of SARS-CoV-2 in Bangladesh (GISAID). The figure shows the multiple introductions of SARS-CoV-2 into Bangladesh.

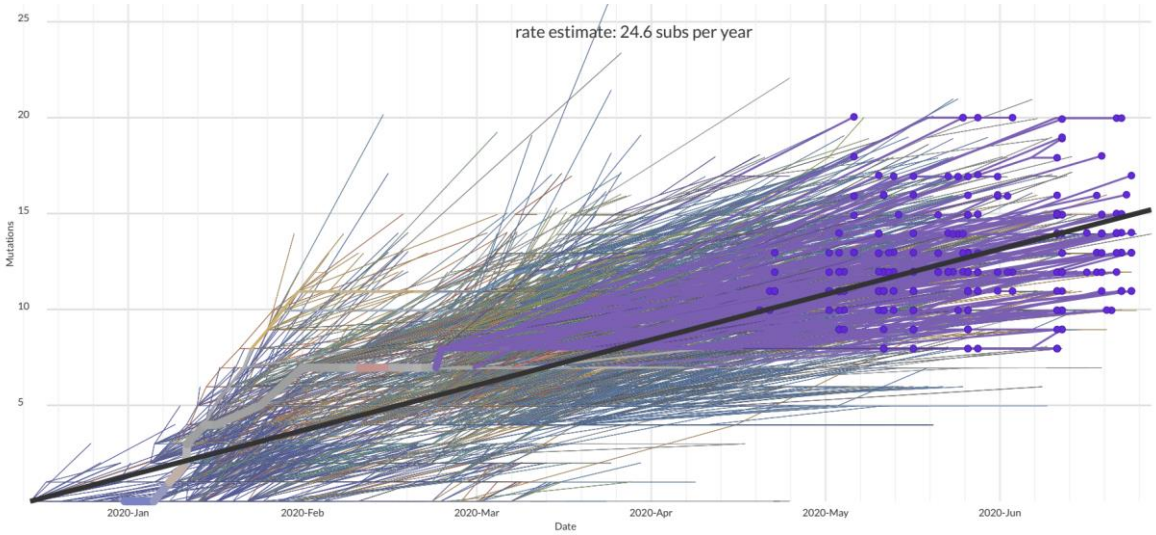

Figure 3: Mutation rate of SARS-CoV-2 of Bangladeshi isolates (GISAID). The nucleotide mutation rate is 24.6 substitution per year, indicating around 2 mutations per month.

Genomic epidemiology

In Bangladesh, SARS-CoV-2 in 99% of cases have characteristic 4 common genomic variations, including 241C>T, 3037C>T, 14408C>T, 23403A>G compare to the first isolate of SARS-CoV-2 in Wuhan, China. These mutations led to the 2 non-synonymous amino acid substitutions at RdRp: 314P>L, and spike protein (S): 614D>G. In 93.5% of cases, a further mutation was observed due to changes at the 28881-28883 GGG>AAC with non-synonymous amino acid substitution at nucleocapsid (N): 203 RG >KR. These variants are mostly seen as linked mutations and are part of a haplotype observed in Europe. It was found that 243 out of 263 isolates of SARS-CoV-2 in Bangladesh belongs to GR clade, while 16, 3, and 1 belongs to GH, G, and O clade (EPI_ISL_466692), respectively (**Fig.4a**). Further analysis of the sequencing data reveals that 51% belong to pangolin lineage B.1.1.25, followed by 29% to B.1.1 (**Fig. 4b**). For analyzing genomic epidemiology, sequence alignment was performed using MUSCLE while MEGA-X conducted an assessment of the best fitting substitution model, and inference of the phylogenetic tree. Support for the tree topology was estimated with 500 bootstrap replicates. One of the sequencing data was excluded due to unusual clustered SNP mutation from both genomic variant and phylogenetic analysis. Based on the sequencing data, we found three central variants distinguished by amino acid changes with strain EPI_ISL_466631 and EPI_ISL_465164) being the ancestral type according to the Wuhan first isolated SARS-CoV-2 outgroup reference coronavirus (EPI_ISL_402124). The maximum likelihood phylogenetic tree in the Figure shows three major clusters A, B, C, and each of the clusters containing several subclusters (**Fig. 5**). In this study, we have shown two genomic epidemiologic analysis using unrooted (**Fig. 5a**) and redial (**Fig. 5b**) phylogenetic tree. To date, most of the mutations that have occurred are only moderately genetically diverse, with an average pairwise difference of 9.6 SNPs between any two genomes, which shows that the common

ancestor is recent and has a mutation rate of around $6x10^{-4}$ nucleotides/genome/year. In this study, a timetree has inferred by applying the RelTime method [9, 10] to the phylogenetic tree constructed by MEGAX, whose branch lengths were calculated using the Maximum Likelihood (ML) method and the General Time Reversible substitution model [11]. Based on analysis using EPI_ISL_402124 as an outgroup, the relative time from Wuhan SARS-CoV-2 to Bangladeshi SARS-CoV-2 is calculated to be $3.5X10^{-4,}$ which is higher than the global mutation rate. It is inevitable that as time progresses, the virus accumulates more independent mutations in different locations. However, no specified clusters related to different divisions were observed in different divisions of Bangladesh (**Fig.6**).

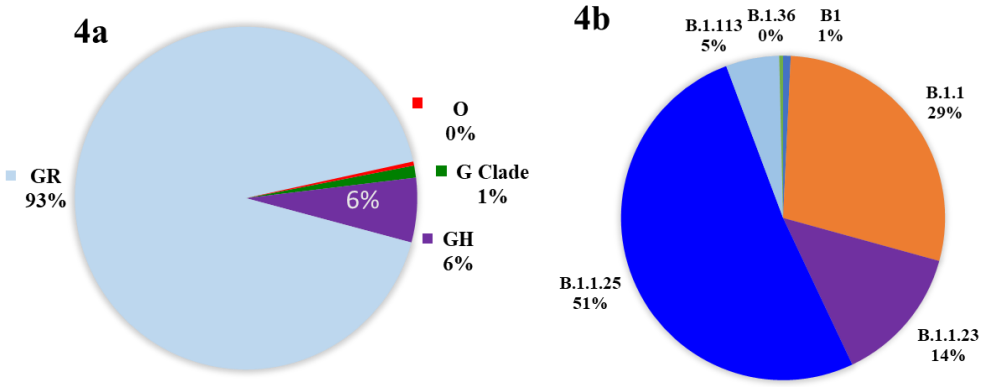

Figure 4: GISAID Clade and Pangolin lineage distribution of SARS-CoV-2 isolates

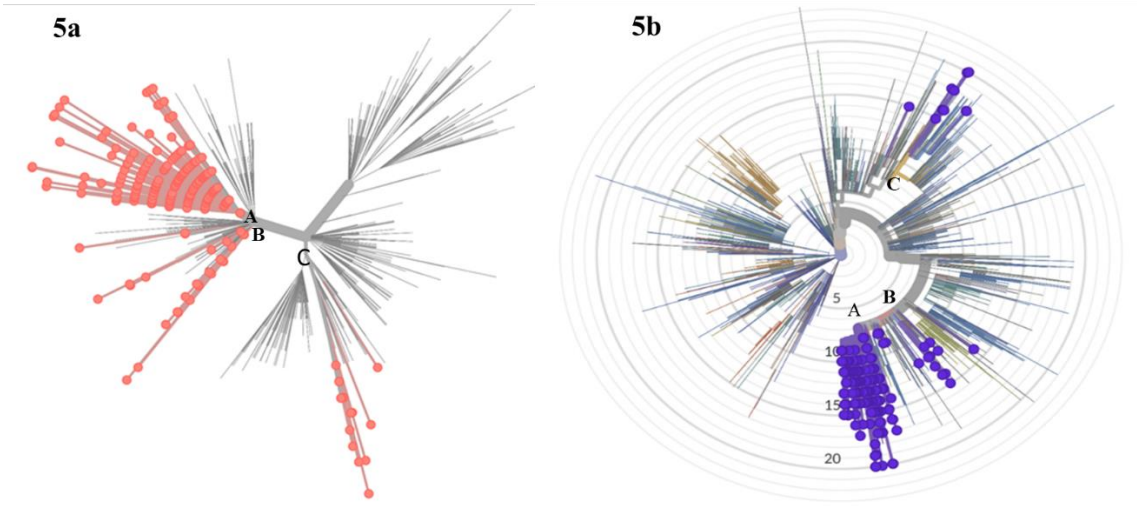

Figure 5: Phylogenetic analysis of SARS-CoV-2 isolates; unrooted (5a) and radial with divergence (5b). A, B, and C indicate significant clusters. In 5bm numbers 5, 10, 15, and 20 indicate the divergence.

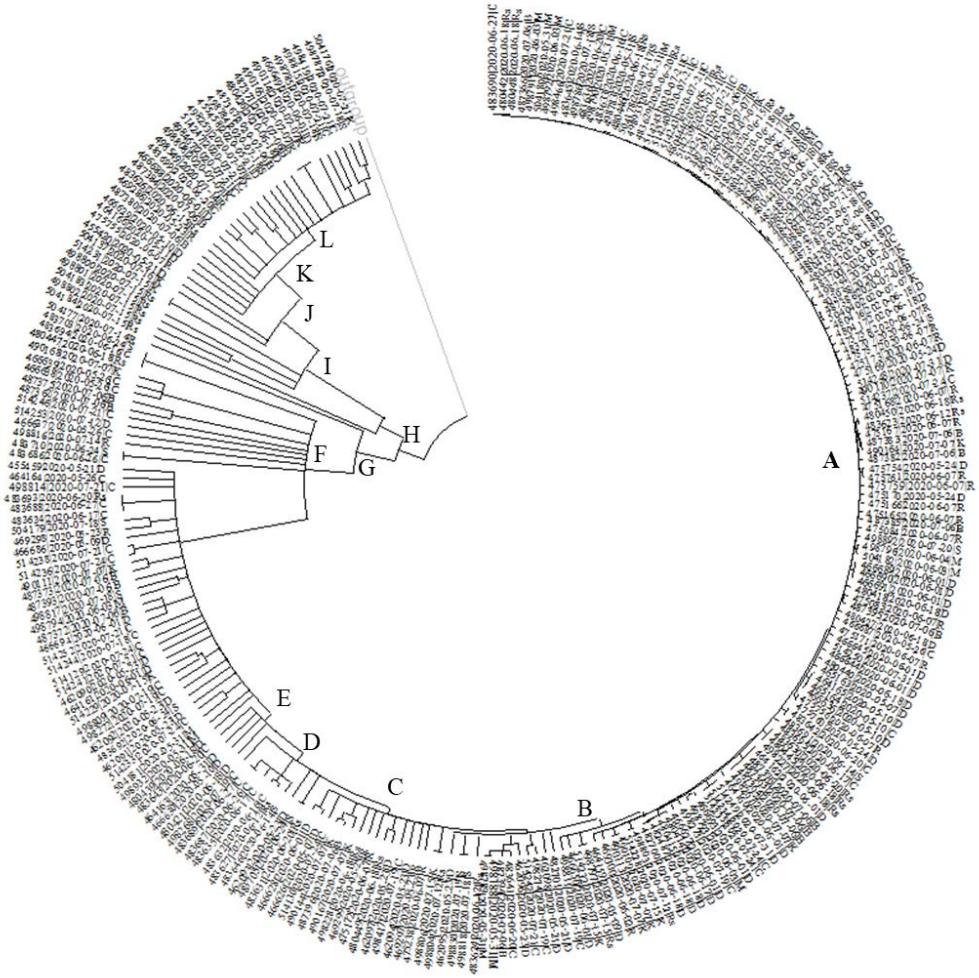

Figure 6: Phylogenetic analysis of 262 SARS-CoV-2 Whole-genome sequencing data. A-L indicate different subclusters.

Genomic variant analysis

Analysis of 262 SARS-CoV-2 highlights a total of 3308 nucleotide mutation events compared to the reference genome (NC_045512.3). This viral journey from Wuhan to Bangladesh, lasting 6 months, is documented by 12.6 mutations events per sample, whereas a common mutation per sample worldwide is 7.23 [12]. This result suggests significantly higher mutation events in Bangladesh compared to the rest of the world. Previously, we have shown that 1258 mutation events at the amino acid level are based on 265 sequencing data published in GISAID from Bangladesh, with an average of 6.8 per sample in Bangladesh [13]. In this study, a total of 3308 mutation events were examined in 262 samples; however, these mutations have occurred in a total of 737 different positions within 25 different proteins (**Fig. 7a &7b**). SNP mutations contributed 719, followed by 15 deletions, and 3 indels. Of the 719 SNP mutations, thymine replaces cytosine, guanine, and adenine by 304, 175, and 13 times, respectively. At the same time, thymine was replaced by cytosine (74), adenine (7), and guanines (12) altogether by 93 times. Nucleotide guanine replaced by adenine and cytosine in 48 and 9 events, respectively. Also,

adenine replaced by guanine, and cytosine in 55 and 8 events, respectively. Moreover, cytosine replaced by adenine 11 and guanine 3 times.

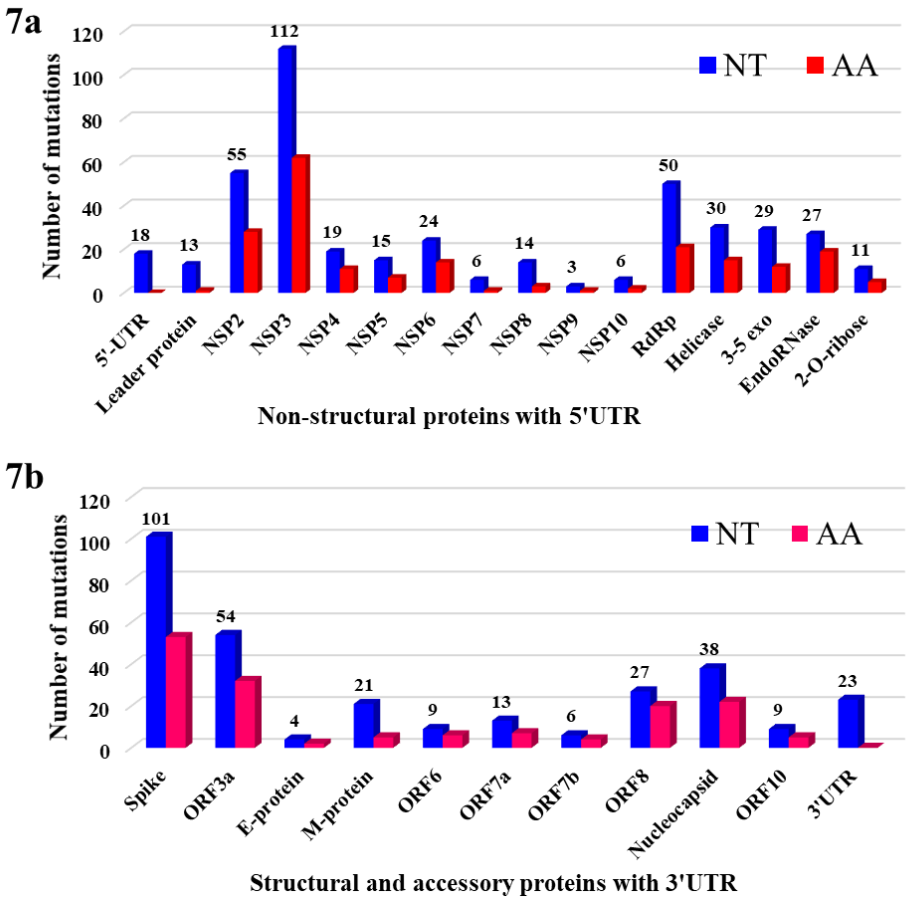

Figure 7: Genomic variant analysis of sequenced SARS-CoV-2 in Bangladesh

Although orf1ab covered 71.3 % of the entire coronavirus genome, it accounts for only 56.2% of active mutations. Papain like protease (PL^{pro}) is 1945 amino acid long, one of two proteases of SARS-CoV-2, which generates the first three mature non-structural proteins. The highest number of mutations occurred at 112 positions of PL^{pro} encoding genome, rendering non-synonymous 62 amino acid change. In $PL^{pro,}$ the most recurrent mutations observed in 100% of cases were 3037, in which cytosine was replaced by thymine. However, this mutation did not change the amino acid. Whereas the main virulence factors 3 like cysteine ($3CL^{pro}$), which is responsible for maturation of 11 viral functional proteins, undergoes less mutational changes. We examined nucleotides mutations at 15 positions of $3CL^{pro}$, resulting in an alteration of 7 amino acids. Another vital enzyme RNA dependent RNA polymerase (RdRp) has also shown nucleotide mutations at 50 positions, which led to the 21 amino acid substitution. The most recurrent variant based on variations in the RdRp was 14408C>T found in 100% of cases. In which nucleotide cytosine at 14408 changed to thymine (uracil), which has altered amino acid proline to Leucine at 323 positions of RNA dependent RNA polymerase.

The primary determinant of the host range and pathogenicity of SARS-CoV-2 is its surface-associated spike or S protein. Spike is a trimeric glycoprotein composed of three chains, each chain consisting of two subunits. The S1 subunit is located at N-terminal, while the S2 subunit is located at the C-terminal. The S1 subunit is responsible for recognizing and binding to the host cell receptor ACE2 while the S2 subunit is specializing for membrane fusion. Compared with the S1, the S2 subunit shows much lower variability [14]. Although 1937 mutation sites are identified so far in S protein, only 1061 lead to amino acid changes. Amino acids 336 to 516 of 1274 amino acid long spike protein form the binding domain, which mediates coronavirus attachment with ACE2 receptors. It has been found that 129 changed amino acids located in the Receptor Binding Domain (RBD). Among them, 11 out of 17 critical amino acids are responsible for protein interactions. In this study, the second-highest mutations due to nucleotide change at 101 different positions of spike protein were detected; 53 of those mutations (52.5%) have rendered the amino acid changes. The most dominant variant G614 was found in 100% of cases due to nucleotide change at position 23403A>G. In a study, it has been shown that retroviruses pseudotyped with SG614 can infect angiotensin-converting enzyme expressing cells significantly more effective way than the SD614 [15]. G614 variant of SARS-CoV-2 isolates predominates over time in locales indicating that this change in the spike protein enhances viral transmission. The second highest variants are based on a mutation in spike protein due to nucleotide changes at position 22444C>T covered by 5% of cases. To date, 5 unique variants of SARS-CoV-2 based on a mutation in spike proteins have been found in Bangladesh. These variants include 14Q>H, 26P>L, 140F del, 144Y del, and 248Y>H. In spike protein, a total of 13 mutations at nucleotide level occurred within the ACE2 receptor binding domain resulting in 5 amino acid substitution at 382V>L, 444K>N, 450N>K, 518L>I, and 520A>S. All of the spike protein mutations were highlighted in figure (**Fig. 7**).

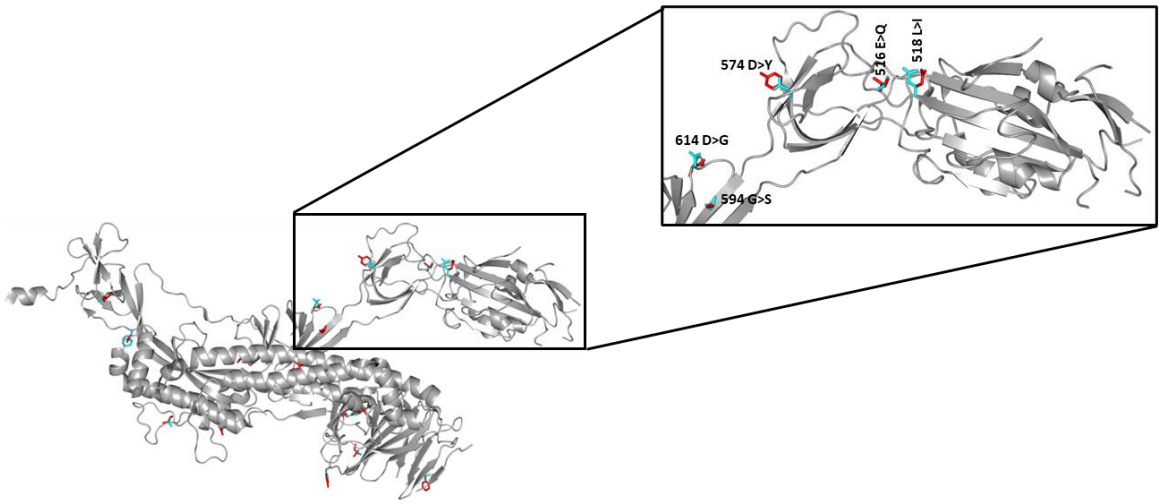

Figure 8: Mutated amino acids are modelled in spike protein structure (PDB ID: 6vxx) using COOT software (Emsley et al., 2010). The structure figures were produced using PYMOL (2).

It has been suggested based on selection analysis that as the virus is transmitted between humans, the genes of both spike and nucleocapsid undergo episodic selection [16]. The positive selection for parts

of any emerging virus is typical [17]. Mutations in the spike or nucleocapsids genes, and subsequently, their adaptation could affect virus stability and pathogenicity [18]. Nucleocapsid (N) protein is encoded by a highly variable region of the N gene, responsible for the formation of the helical nucleocapsid. N protein can elicit both cell-mediated and humoral immune response and thus has potential value in vaccine development [3]. In the N gene, one nucleotide change was detected at position G28881A found in 246 samples. The Arg residue at 203 and Gly residue at 204 positions of nucleocapsid has changed into Lys and Arg- residues, respectively, in 93.5% of cases. However, the mutations, as mentioned above, did not yet to be associated with changes in viral transmissibility or pathogenicity.

Two silent mutations were also detected at two different positions, including 5'UTR and NSP2 in 100% and 83% of cases, respectively. The nucleotides thymine replaces cytosine at 241 of 5'UTR, and the amino acid at 1163 position of NSP2 has changed from Isoleucine to Phenylalanine. SARS-CoV-2 has eight accessory proteins including orf3a, orf3b, orf6, orf7a, orf7b, orf8, orf9b, and orf14 [19]. Although orf14 was not detected in 263 sequencing data analyzed in this study. As more genomes from Bangladesh are made publicly available, analysis of genetic variants might have revealed the highest diversity occurring in nsp3 along with spike proteins, nucleocapsids, ORF3a, and ORF8 across the samples. Among the accessory proteins, ORF3a contains six functional domains (I to VI), which may link to viral infectivity, virulence, and ion channel formation [20]. In a previous study based on 2782 whole-genome sequencing data, it has shown that 51 different non-synonymous amino acid substitutions in the 3a proteins [20]. In this study, we observed 54 mutations at the nucleotide level, which lead to 32 non-synonymous amino acid substitutions in orf3a. Besides, 20 non-synonymous mutations were detected in orf8.

In conclusion, our analysis sheds light on the different variants of SARS-CoV-2 currently circulating in Bangladesh. Most recurrent synonymous mutations were detected at positions 241 in 5'UTR, and 3037 in nsp3. Whereas non-synonymous recurrent mutations were detected at positions 1063, 14408, 23403, 29881-29883 in nsp2, nsp12, Spike protein, and nucleocapsid, respectively. These variants indicate the underlying link of Bangladeshi SARS-CoV-2 isolates with part of a haplotype observed high in Europe. Also, we have shown the transmission and clades of circulating SARS-CoV-2 in Bangladesh. Further research should be done to monitor the genetic and non-anonymous variants circulating in Bangladesh for understanding the infectivity and transmission of SARS-CoV-2.

Data availability statement

Whole-genome sequences of SARS-CoV-2 strains are available in the GISAID database. The GISAID ID is provided separately in supplementary data-1.

Acknowledgements and Funding

We acknowledge the Bangladesh Council of Scientific and Industrial Research authority for supporting the study, recognize the National Institute of Laboratory Medicine and Referral Center (NILMRC), Illumina channel partner Invent Technologies Limited, ScITech Consulting Ltd, developers, scientists associated with GISAID and specially Ministry of Science and Technology, Bangladesh.

Ethics statement

Ethical approval is not required as samples from suspected COVID-19 patients were collected and tested at the NILMRC but not in BCSIR. However, the whole-genome sequencing of SARS-CoV-2 was approved by the human research ethics committee of the National Institute of Laboratory Medicine and Referral Center (NILMRC).

References

1. Li, W., et al., Bats are natural reservoirs of SARS-like coronaviruses. Science, 2005. **310**(5748): p. 676-679.
2. Dominguez, S.R., et al., Detection of group 1 coronaviruses in bats in North America. Emerging Infectious Diseases, 2007. **13**(9): p. 1295.
3. Su, S., et al., Epidemiology, Genetic Recombination, and Pathogenesis of Coronaviruses. Trends in Microbiology, 2016. **24**(6): p. 490-502.
4. Graham, R.L. and RS Baric, Recombination, Reservoirs, and the Modular Spike: Mechanisms of Coronavirus Cross-Species Transmission. Journal of Virology, 2010. **84**(7): p. 3134-3146.
5. Cui, J., F. Li, and Z.-L. Shi, Origin and evolution of pathogenic coronaviruses. Nature Reviews Microbiology, 2019. **17**(3): p. 181-192.
6. Lee, J., G. Chowell, and E. Jung, A dynamic compartmental model for the Middle East respiratory syndrome outbreak in the Republic of Korea: A retrospective analysis on control interventions and superspreading events. Journal of Theoretical Biology, 2016. **408**: p. 118-126.
7. Zhou, P., et al., A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature, 2020. **579**(7798): p. 270-273.
8. IEDCR, Updates on the coronavirus disease 2019 (covid-19) situation in Bangladesh. 2020: https://www.iedcr.gov.bd/.
9. Tamura, K., et al., Estimating divergence times in large molecular phylogenies. Proceedings of the National Academy of Sciences, 2012. **109**(47): p. 19333-19338.
10. Tamura, K., Q. Tao, and S. Kumar, Theoretical Foundation of the RelTime Method for Estimating Divergence Times from Variable Evolutionary Rates. Molecular Biology and Evolution, 2018. **35**(7): p. 1770-1782.
11. Thomas, R.H., Molecular Evolution and Phylogenetics. Heredity, 2001. **86**(3): p. 385-385.
12. Mercatelli, D. and F.M. Giorgi, Geographic and Genomic Distribution of SARS-CoV-2 Mutations. Frontiers in Microbiology, 2020. **11**(1800).
13. Mahmud, A.S.M., et al., In-vivo evaluation of the anti-diarrheal effect of Lactococcus lactis subspecies lactis and Lactococcus piscium isolated from yogurt. bioRxiv, 2020: p. 2020.07.30.226688.
14. Masters, P.S., The molecular biology of coronaviruses. Advances in Virus Research, 2006. **66**: p. 193-292.
15. Zhu, N., et al., A novel coronavirus from patients with pneumonia in China, 2019. New England Journal of Medicine, 2020.

16. Benvenuto, D., et al., The 2019-new coronavirus epidemic: Evidence for virus evolution. Journal of Medical Virology, 2020. **92**(4): p. 455-459.
17. Sironi, M., et al., Evolutionary insights into host-pathogen interactions from mammalian sequence data. Nature Reviews Genetics, 2015. **16**(4): p. 224-36.
18. Baric, RS, et al., Episodic evolution mediates interspecies transfer of a murine coronavirus. Journal of Virology, 1997. **71**(3): p. 1946-55.
19. Ceraolo, C. and F.M. Giorgi, Genomic variance of the 2019-nCoV coronavirus. Journal of Medical Virology, 2020. **92**(5): p. 522-528.
20. Issa, E., et al., SARS-CoV-2 and ORF3a: Non-synonymous Mutations, Functional Domains, and Viral Pathogenesis. mSystems, 2020. **5**(3): p. e00266-20.