Phylogenetic analysis of first Bangladeshi SARS-CoV-2 strain isolated in Bangladesh understanding the possible origin and novel mutations

Otun Saha[1], Md Miraj Kobad Chowdhury[1], Md. Shahadat Hossain[2], Afroza Sultana[2], and Md. Mizanur Rahaman[1,3]

[1]Dhaka University
[2]Noakhali Science and Technology University
[3]University of Dhaka

June 5, 2020

Abstract

Severe acute respiratory syndrome virus 2 (SARS-CoV-2), the causative agent of the ongoing pandemic of COVID-19 has already marched across 210 countries globally. This novel coronavirus is highly infectious and left us with no available cure and only with containment option. To understand the molecular epidemiology and vaccine design, genetic sequences from different regions are published and analyzed. Here, the first published whole genome sequence of Bangladesh was compared with Asian countries as well as representative strains from across the globe. Phylogenetic analysis of the first Bangladeshi genome, that was published on May 12, 2020 revealed at least 9 Single nucleotide polymorphisms from the Wuhan, China strains and 2 of these mutations are novel compared to the strains analyzed. Among the novel mutations nucleotide 1163 mutation is very novel when compared with all the genomes deposited at Global Initiative on Sharing Avian Influenza Data (GISAID). However, the other mutation at nucleotide 17019, mutation NSP13 E261D is occurred second time after a strain from Austrian patient showed the similar mutation. Structure and screening results of both novel mutation were discussed in detail. Further analysis of these identified novel mutations will boost the understanding of the behavior of SARS-CoV-2 in this region and vaccination and drug development programs would be beneficial.

Introduction

A disease with a cluster of pneumonia like symptoms arose in Wuhan City, China back in December, 2019 and broke out across the whole world later on. World Health Organization (WHO) recognized a new coronavirus strain behind that disease and designated the virus as 2019 novel coronavirus (nCoV-2019) [1, 2]. Known as Severe Acute Respiratory Syndrome (SARS CoV-2), this novel coronavirus actually roots back to the viral family of *Coronaviridae* and order of *Nidoviralae* as phylogenetic analysis suggests [3]. The coronavirus disease 2019 (COVID-19) was announced as a pandemic by WHO [4] as it has currently affected almost all the countries and territories on earth. As of 13th May at 16.28 GMT, over 2.4 million people are currently infected and over 295,000 have fallen to death due to the pandemic (https://www.worldometers.info/coronavirus/). Asian territory is suffering much as some 718,000 cases are on the notice at the moment and about 24,000 are dead (https://www.worldometers.info/coronavirus/). To understand further epidemiology of the outbreak in Asian countries especially highly densely populated South Asian regions, analysis of the genome sequencing of the viral strains from this region is very important. So far, almost 1400 whole genome sequencing data have been published from Asian regions and among South Asian countries (https://www.gisaid.org/), India, Pakistan, Sri-Lanka have already published their genome sequences, while Bangladesh has just published its genome sequencing data for the first time. In order to understand the molecular epidemiology of the outbreak and suitability to design universal vaccine, we analyzed the sequences of SARS-CoV-2 genome derived from Bangladeshi sample comparing with other available genomes deposited at Global Initiative on Sharing Avian Influenza Data (GISAID).

Materials and Methods

The Global Initiative on Sharing Avian Influenza Data (GISAID) was founded in 2006, and, since 2010, has been hosted by the German Federal Ministry of Food, Agriculture and Consumer Protection. GISAID has also become a coronavirus repository since December 2019. As of 13 May 2020, the cutoff point for our phylogenetic analysis, the GISAID database (https://www.gisaid.org/) had compiled 16,667 coronavirus full genomes, isolated from humans, Chinese pangolins, and bat*Rhinolophus affinis* . Among the all deposited genome sequences 1485 from all Asian countries. Although SARS-CoV-2 is an RNA virus, the deposited sequences, by convention, are in DNA format. We discarded partial sequences, and used only the most complete genomes that we aligned to the full reference genome (NC_045512.2) by Wu et al. (2020) [5] comprising 29,903 nucleotides which was retried from NCBI (https://www.ncbi.nlm.nih.gov/nuccore). Finally, to ensure comparability, we truncated the flanks of all sequences to the consensus range 56 to 29,797, with nucleotide position numbering according to the Wuhan 1 reference sequence [5]. To analyze the obtained 1st Bangladeshi SARS-CoV-2 genome derived from the infected female patients aged 22 (GISAID accession ID: EPI_ISL_437912) which was submitted by Child Health Research Lab, Bangladesh in a phylogenetic context, a dataset of 32 available SARS-Cov-2 complete genomes from different Asian countries followed by few other continent countries was retrieved from GISAID (https://www.gisaid.org/, last access 12 May 2020). At least one sequence from all Asian countries who has submitted SARS-Cov-2 genome in the GISAID database was taken to reveal the draft scenario of the circulating SARS-CoV-2 strain in this Asia zone in comparison of newly revealed Genome from Bangladesh. Sequence alignment was performed using Multiple Sequence Comparison by Log- Expectation (MUSCLE) software (http://www.clustal.org)[6]. Estimation of the best fitting substitution model (Hasegawa, Kishino, and Yano, HKY model) and inference of the phylogenetic tree were conducted by a neighbor-joining approach using Molecular Evolutionary Genetics Analysis across Computing Platforms (MEGA 7; https://www.megasoftware.net/) [7]. Support for the tree topology was estimated with 1,000 bootstrap replicates. Using an alignment, the single nucleotide polymorphisms (SNPs) composition and the potentially resulting variable amino-acids in derived protein sequences compared with the Wuhan reference sequences (NC_045512), were further investigated with six other genome sequences (EPI_ISL_430111, EPI_ISL_437762, EPI_ISL_412974, EPI_ISL_417444, EPI_ISL_427813, EPI_ISL_437438) that clustered or non-clustered from Asia and Europe with the sequence of the patient in Bangladesh. For mutation type analysis MEGA7 and Datamonkey.org web server was used [8]. For analysis of the novel mutation NCBI Blast was used (https://blast.ncbi.nlm.nih.gov/Blast.cgi). Protein structures were predicted using Phyre2 (Protein homology/analogy recognition engine v2.0) [9] and I-TASSER (Iterative threading assembly refinement) [10]. Templates with the highest confidence were used to generate the model in each case. For Phyre2, intense model was used. Generated PDB files were analysed and aligned using PyMOL v2.3.2. Images were processed in adobe illustrator vCS6. Secondary structures were predicted using PSIPRED included in Phyre2.

Results

Phylogenetic analysis

The neighbor-joining phylogenetic tree in the Figure 1 shows a main clade containing 5 clusters. The viral genome sequence of the female 22 year aged Bangladeshi patients (GISAID accession ID: EPI_ISL_437912) was closely related to that of another 4 genome sample taken from the four different Asian countries (GISAID accession ID: EPI_ISL_430111; EPI_ISL_438966; EPI_ISL_425214; EPI_ISL_422227). These five genome sequences were located in a cluster A.1 with genomes mainly from Asia (Russia, Japan, Kuwait), but also one from Spain (Figure, highlighted in blue). In cluster A, six other six genomes mainly from Asian countries like Saudi Arabia (EPI_ISL_437762), Israel (EPI_ISL_435292), Turkey (EPI_ISL_437335), Myanmar (EPI_ISL_434709) and one from United Kingdom (GISAID accession ID: EPI_ISL_442502). Based on these comparisons, it can be predicted that the circulating Bangladeshi SARS-CoV-2 isolate published has phylogenetic relevance to some of the Asian countries as well as Spain. In the tree, some sequences from other SARS-CoV-2 collected in Asian countries segregated in separate clusters from the one cluster containing the respective Bangladeshi sequences characterized in this study. There was for example a cluster B formed by 8 sequences dominated by all Asian countries (Cambodia, Nepal, Malaysia, Pakistan, and Indonesia) except one from USA. This cluster B also contain the reference sequence from Wuhan, china (NC_045512.2). Cluster C contain 6 genome sequences followed by Cluster D (2 genome sequences). The two sequences (EPI_ISL_438138 and EPI_ISL_437438) retrieved from the neighboring country of Bangladesh, India, which covers almost all the direction except south (Bay of Bengal is in south) are in the cluster E (Figure 1).

Analysis of SNPs

The genome-wide SNPs are reported in Table 1 (positions referred respect to the reference sequence; GenBank accession number: NC_045512.2). The corresponding amino-acid positions and variations inside the proteins are shown in Table 2. The genome sequence from the Bangladesh differed in nine nucleotide positions from that of the COVID-19 patient in Wuhan (NC_045512.2), while with the genome sequence isolated from the Indian patient showed 15 nucleotide variations followed by Italian patient 11 nucleotide variations (Table 1). Among all of the 11 nt mutation 2 mutation at position 1162 and 17019 reveled the novel mutation in comparison of the all submitted sequences in the NCBI database (Figure 2). Table 2 that depicts overall five Nonsynonymous mutations that was observed compared to the reference Wuhan sequence. The sequence of the Bangladeshi female patient (EPI_ISL_437912) presented a mutation 388F with respect to the reference Wuhan genome (L) (Table 3). In addition, this EPI_ISL_437912 also presented EP mutation 4803L, 5673D, 6508G, 9627K and 9628R with respect to the reference Wuhan genome P,E, D, R,G respectively with 2 amino acid changes (Figure 2 & Table 3). In comparison with the most close neighboring country India sequences, EPI_ISL_437912 presented 11 mutation at 388F,607G,2104T,3694L,4808L, 5673I,6508G,9503P,9516R,9627K,9628R with respect to the Indian patient genome I, S, K, F, S, D, L, S, R, G respectively (Table 3). Meanwhile, the sequence of the Russia patient (EPI_ISL_430111) presented I and S at amino acidic position 388 and 5673 with respect to the Bangladeshi sequence (F and I)(Table 3).Furthermore, predicted 3D structures of NSP13-261E and NSP13-261D were >90% confident. These structures were identical and completely overlapped with each other when aligned using PyMOL. The predicted secondary structures were similar, both 261E and 261D were localized in α -helix (Figure 3a). Based on these findings, it can be concluded that the E261D mutation in NSP13 has no effect on protein structure and hence the function. Nonetheless, amino acid 261 is not a part of NSP13 active site. However, the predicted structures of both NSP2-120I and NSP-120F were around 46% confident. These structures failed to align, but the phenylalanine reside of the NSP-120F was in a α -helix (Figure 3b) whereas the isoleucine reside of the NSP2-120I was not. Despite the reliability of Phyre2 structures, predicted secondary structures also showed the same thing with high confidence. Such findings suggest that the I120F mutation in NSP2 could affect the structure and function of NSP2. The structure and function of NSP2 is yet to explore, but current prediction suggests that there is no known functional domain spanning this mutation site (Figure 3a & Figure 3b).

Discussion

In this study, the full length genome of SARS-CoV-2 strain Bangladesh is completely analyzed and compared with the viral genome sequence of the COVID-19 patient in Wuhan [5] and some neighboring countries as well as Western countries (https://www.gisaid.org/). Phylogenetic analysis consistently placed the Bangladeshi patient's strain in a distinct cluster from the neighboring country India and also the Wuhan patient's strain. The strain of Bangladeshi patient grouped in this small study with 33 sequences within mostly Asian countries like Japan, Kuwait, Russia and also European country, Spain. In comparison with the Wuhan patient's strain a total of 9 SNPs were observed [5]. This is consistent with the other reports published recently where India reported their isolated strains with 15 mutations while Italy reported 11 mutations compare to the Wuhan strains. Such variations reflects the various kind of mutations that are common to the most RNA viruses [11, 12, 13, 14]. This strain of Bangladeshi female patient also showed at least SARS-CoV-2 strain 10 SNPs when compared with the Indian strains. Again it clearly is very much diverse as the viral outbreak might be initiated by the people who visited the hotspots or migrated from those areas. And Bangladeshi and Indian people are working across the globe in Europe, America and Middle Eastern countries being the being the dominant countries. While trading relationship is also very high for these countries to diverse overseas countries.

The sequence data of Bangladesh revealed 2 novel SNPs. Among them, Mutation NSP13 E261D already occurred 2 times (0.01% of all samples with NSP13 sequence) in 2 countries including Bangladesh. The first strain with this mutation, collected in March 2020, was hCoV-19/Austria/CeMM0004/2020. The mutation most recently occurred in strain hCoV-19/Bangladesh/CHRF_nCOV19_0001/2020, collected in April 2020. This finding is very important as it only occurred second time. Even the NCBI database revealed this as first time for Bangladesh, but it matched with Austrian strain when we analyzed with the GISAID database. There is at least 1 SNP that was observed in Bangladeshi Patient's sequence that is completely novel. It is the mutation in NSP3 region nucleotide 1162 which is T for Bangladeshi strain, but A for Wuhan strain. This is for the amino acid Iso leucine (I) is replaced by Phenyl alanine (F). It would be very interesting to analyze the protein function more details of NSP3. However, initial analysis performed by i-tesser didn't show significant differences. The mutations should be further investigated to understand whether they may affect virus characteristics. During this analysis, the lack of epidemiological information available with most sequences deposited in the database and the number of incomplete genomes available made difficulties in the strain selections. Nevertheless, these data provides valuable insights about SARS-CoV-2 strain in Bangladesh and may be useful to understand the dynamics of the local transmission of SARS-CoV-2 in the coming days.

References

1. Islam, M. M., Rakhi, N. N., Islam, O. K., Saha, O., & Rahaman, M. M. Challenges to be considered to evaluate the COVID-19 preparedness and outcome in Bangladesh. *International Journal of Healthcare Management* 1-2(2020).
2. Ahamed, M. M., Naznin, R. N., Saha, O., & Rahaman, M. M. Recommendation of fecal specimen for routine molecular detection of SARS-CoV-2 and for COVID-19 discharge criteria. Pathogens and global. Health (2020).
3. Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., . . . & Cheng, Z. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. The Lancet 395(10223): 497-506 (2020).
4. World Health Organization. Novel Coronavirus. Available online: https://www.who.int/ csr/don/19-march-2020-novel-coronavirus-japan-ex-china/en/ (accessed on 19 march 2020).
5. F. Wu et al. A new coronavirus associated with human respiratory disease in China. Nature 579, 265–269 (2020).
6. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5(1):113(2004).
7. Kumar, S., Stecher, G., and Tamura, K . MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol. Biol. Evol 33, 1870–1874(2016). doi: 10.1093/molbev/msw054.
8. Rahman, A., Islam, M. S., Saha, O., & Saha, T. C. Molecular Evolutionary Analysis of a-Defensin Peptides in Vertebrates. *Rajshahi University* Journal of Science and Engineering, *44* , 85-93 (2018).
9. Kelley LA et al. The Phyre2 web portal for protein modeling, prediction and analysis. Nature Protocols 10: 845-858(2015).
10. J Yang, Y Zhang. I-TASSER server: new development for protein structure and function predictions. Nucleic Acids Research 43: W174-W181(2015).
11. Yadav PD, Potdar VA, Choudhary ML, Nyayanit DA, Agrawal M, Jadhav SM, Majumdar TD, Shete-Aich A, Basu A, Abraham P, Cherian SS. Full-genome sequences of the first two SARS-CoV-2 viruses from India. Indian J Med Res 151:200-9(2020).
12. Pachetti, M., Marini, B., Benedetti, F., Giudici, F., Mauro, E., Storici, P., Masciovecchio, C., Angeletti, S., Ciccozzi, M., Gallo, R. C., Zella, D., & Ippodrino, R. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. *Journal of translational medicine* 18(1), 179(2020). https://doi.org/10.1186/s12967-020-02344-6
13. Sardar, R., Satish, D., Birla, S., & Gupta, D. Comparative analyses of SAR-CoV2 genomes from dif-

ferent geographical locations and other coronavirus family genomes reveals unique features potentially consequential to host-virus interaction and pathogenesis. *bioRxiv* , (2020).

14. Sefanelli, P., Faggioni, G., Lo Presti, A., Fiore, S., Marchi, A., Benedetti, E., et al. Whole genome and phylogenetic analysis of two SARS-CoV-2 strains isolated in Italy in January and February 2020: additional clues on multiple introductions and further circulation in Europe. *Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin 25* (13), 2000305(2020).

Author's contributions

O.S., M.K.C and N.N.R carried out the studies (Data collection and data analysis). O.S., M.S.H and N.N.R drafted the manuscript. M.M.R. developed the hypothesis, supervised the whole work and critically review the drafted manuscript. All authors read and approved the final manuscript.

Acknowledgments

The authors would like to acknowledge Bangabandhu Science & Technology Fellowship Trust for supporting Otun Saha with PhD fellowship.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Funding source

No Funding

Table 1: Single nt polymorphisms (SNPs)a deduced by comparison of 1st Bangladesh whole genome sequences of SARS-CoV-2 focused in this study with selected SARS-CoV-2 sequences (n=7 compared sequences)

SARS-CoV-2 sequence ID (country from which the sequence originated)	**Nucleotide position of SARS-CoV-2 Genome**	**Nucleotide position of SARS-CoV-2 Genome**	**Nucleotide position of SARS-CoV-2 Genome**	**Nucleotide position of SARS-CoV-2 Genome**	**Nucleotide position of SARS-CoV-2 Genome**	**Nucleotide position of SARS-CoV-2 Genome**	**Nucleotide position of SARS-CoV-2 Genome**	**Nucleotide position of SARS-CoV-2 Genome**	**Nucleotide position of SARS-CoV-2 Genome**	**Nucleotide position of SARS-CoV-2 Genome**	**Nucleotide position of SARS-CoV-2 Genome**	**Nucleotide position of SARS-CoV-2 Genome**	**Nucleotide position of SARS-CoV-2 Genome**	**Nucleotide position of SARS-CoV-2 Genome**	**Nucleotide position of SARS-CoV-2 Genome**	**Nucleotide position of SARS-CoV-2 Genome**	**Nucleotide position of SARS-CoV-2 Genome**	**Nucleotide position of SARS-CoV-2 Genome**	**Nucleotide position of SARS-CoV-2 Genome**	**Nucleotide position of SARS-CoV-2 Genome**
	241	**884**	**1163**	**1348**	**1820**	**3037**	**5572**	**6310**	**6702**	**9159**	**11083**	**14408**	**14805**	**17019**	**18744**	**18877**	**23403**	**23929**	**25613**	**2614**
NC_045512.2 (China)	C	C	A	C	G	C	G	C	T	C	G	C	C	G	C	C	A	C	G	G
EPI_ISL_437912 (Bangladesh)	T	C	T	C	G	T	G	C	T	C	G	T	C	T	C	C	G	C	G	G

SARS-CoV-2 sequence ID (country from which the sequence originated)	Nucleotide position of SARS-CoV-2 Genome	Nucleotide position of SARS-CoV-2 Genome	Nucleotide position of SARS-CoV-2 Genome	Nucleotide position of SARS-CoV-2 Genome	Nucleotide position of SARS-CoV-2 Genome	Nucleotide position of SARS-CoV-2 Genome	Nucleotide position of SARS-CoV-2 Genome	Nucleotide position of SARS-CoV-2 Genome	Nucleotide position of SARS-CoV-2 Genome	Nucleotide position of SARS-CoV-2 Genome	Nucleotide position of SARS-CoV-2 Genome	Nucleotide position of SARS-CoV-2 Genome	Nucleotide position of SARS-CoV-2 Genome	Nucleotide position of SARS-CoV-2 Genome	Nucleotide position of SARS-CoV-2 Genome	Nucleotide position of SARS-CoV-2 Genome	Nucleotide position of SARS-CoV-2 Genome	Nucleotide position of SARS-CoV-2 Genome	Nucleotide position of SARS-CoV-2 Genome	Nucleotide position of SARS-CoV-2 Genome
EPI_ISL_430111 (Russia)	T	C	A	C	G	T	G	C	T	C	G	T	C	G	T	C	G	C	G	G
EPI_ISL_437762 (Saudi Arabia)	T	C	A	C	G	T	G	C	T	C	G	T	C	G	C	T	G	C	T	G
EPI_ISL_412974 (Italy)	C	C	A	C	G	C	G	C	T	C	T	C	C	G	C	C	A	C	G	T
EPI_ISL_417444 (Pakistan)	C	C	A	A	G	C	G	C	T	T	T	C	C	G	C	C	A	C	G	G
EPI_ISL_427813 (Korea)	T	C	A	C	G	C	T	C	T	C	T	C	T	G	C	C	A	C	G	T
EPI_ISL_437438 (India)	C	T	A	C	A	C	G	A	C	C	T	C	C	G	C	C	A	T	G	G

N: nucleocapsid protein; SARS-CoV-2: severe acute respiratory syndrome coronavirus; SNP: single nt polymorphism; UTR: untranslated region. SNPs are shown according to nt positions in the genome sequence and gene location. The targeted sequences focused in this study is the ones from Bangladesh (EPI_ISL_437912).

NC_045512.2 (China) sequence used as reference genome

Table 2: Single nt polymorphisms (SNPs) a deduced by comparison of 1st Bangladesh whole genome sequences of SARS-CoV-2 focused in this study with reference genome (NC_045512.2)

Position	**NC_045512.2**	**EPI_ISL_437912**	**Mutation type**	**Protein change**
241	C	T	Upstream gene variation	Non Coding
1163	A	T	Nonsynonymous	I F
3037	C	T	*Synonymous*	No Change
14408	C	T	Nonsynonymous	P L
17019	G	T	*Nonsynonymous*	E D
23403	A	G	*Nonsynonymous*	D G
28881	G	A	*Nonsynonymous*	R K
28882	G	A	*Synonymous*	No Change
28883	G	C	*Nonsynonymous*	G R

Here, NC_045512.2= References sequence; EPI_ISL_437912=1st Bangladeshi whole genome sequences; synonymous: mutation is a change in the DNA sequence that codes for amino acids in a protein sequence, but does not change the encoded amino acid; *nonsynonymous* : substitution is a nucleotide mutation that alters the amino acid sequence of a protein.

Table 3: Amino acid variations deduced by comparing translations of targeted (EPI_ISL_437912) whole genome sequences of SARS-CoV-2 focused in this study with those of selected SARS-CoV-2 sequences (n=7 compared sequences)

SARS-CoV-2 sequence ID (country from which the sequence originated)	**Amino Acids of SARS-CoV-2 Genome**	**Amino Acids of SARS-CoV-2 Genome**	**Amino Acids of SARS-CoV-2 Genome**	**Amino Acids of SARS-CoV-2 Genome**	**Amino Acids of SARS-CoV-2 Genome**	**Amino Acids of SARS-CoV-2 Genome**	**Amino Acids of SARS-CoV-2 Genome**	**Amino Acids of SARS-CoV-2 Genome**	**Amino Acids of SARS-CoV-2 Genome**	**Amino Acids of SARS-CoV-2 Genome**	**Amino Acids of SARS-CoV-2 Genome**	**Amino Acids of SARS-CoV-2 Genome**	**Amino Acids of SARS-CoV-2 Genome**	**Amino Acids of SARS-CoV-2 Genome**	**A Ac of SA Co 2 G**
	295	**388**	**466**	**607**	**1857**	**2103**	**3053**	**2104**	**3694**	**4803**	**5673**	**6248**	**6508**	**7801**	**85**
NC_045512.2(China)	**R**	**I**	**V**	**G**	**M**	**G**	**P**	**T**	**L**	**P**	**E**	**T**	**D**	**C**	**R**
EPI_ISL_437912 (Bangladesh)	**R**	**F**	**V**	**G**	**M**	**G**	**P**	**T**	**L**	**L**	**D**	**T**	**G**	**C**	**R**

SARS-CoV-2 sequence ID (country from which the sequence originated)	Amino Acids of SARS-CoV-2 Genome	Amino Acids of SARS-CoV-2 Genome	Amino Acids of SARS-CoV-2 Genome	Amino Acids of SARS-CoV-2 Genome	Amino Acids of SARS-CoV-2 Genome	Amino Acids of SARS-CoV-2 Genome	Amino Acids of SARS-CoV-2 Genome	Amino Acids of SARS-CoV-2 Genome	Amino Acids of SARS-CoV-2 Genome	Amino Acids of SARS-CoV-2 Genome	Amino Acids of SARS-CoV-2 Genome	Amino Acids of SARS-CoV-2 Genome	Amino Acids of SARS-CoV-2 Genome	Amino Acids of SARS-CoV-2 Genome	Am Ac of SA Co 2 Ge
EPI_ISL_430111 (Russia)	R	I	V	G	M	G	P	T	L	L	E	M	G	T	R
EPI_ISL_437762 (Saudi Arabia)	R	I	V	G	M	G	P	T	L	L	E	T	G	C	I
EPI_ISL_412974 (Italy)	R	I	V	G	M	G	P	T	F	P	E	T	D	C	R
EPI_ISL_417444 (Pakistan)	C	I	I	G	M	G	L	T	F	P	E	T	D	C	R
EPI_ISL_427813 (Korea)	R	I	V	G	I	T	P	T	F	P	E	T	D	C	R
EPI_ISL_437438 (India)	R	I	V	S	M	G	P	K	F	P	E	T	D	C	R

SARS-CoV-2: severe acute respiratory syndrome coronavirus. The amino acid positions refer to those in each respective protein sequence of the Wuhan reference (GenBank accession number: NC_045512.2), starting from the first Serine. The only sequences targeted in this study is the ones from Bangladesh (EPI_ISL_-

437912).

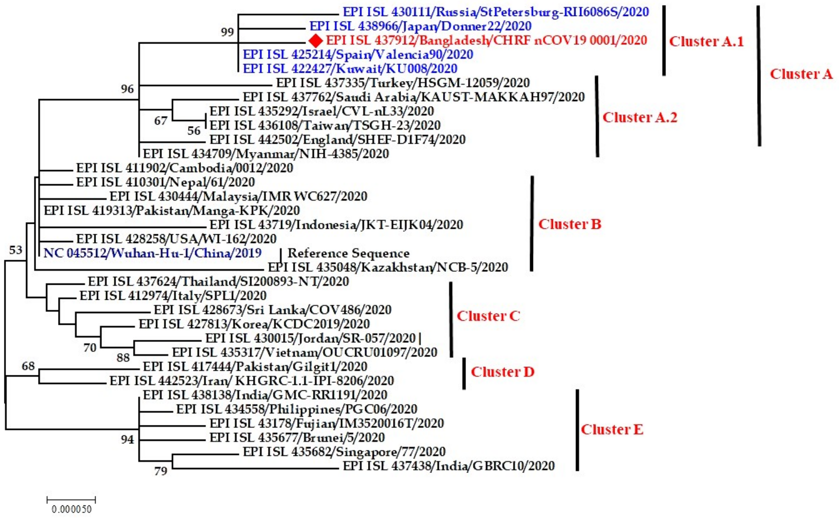
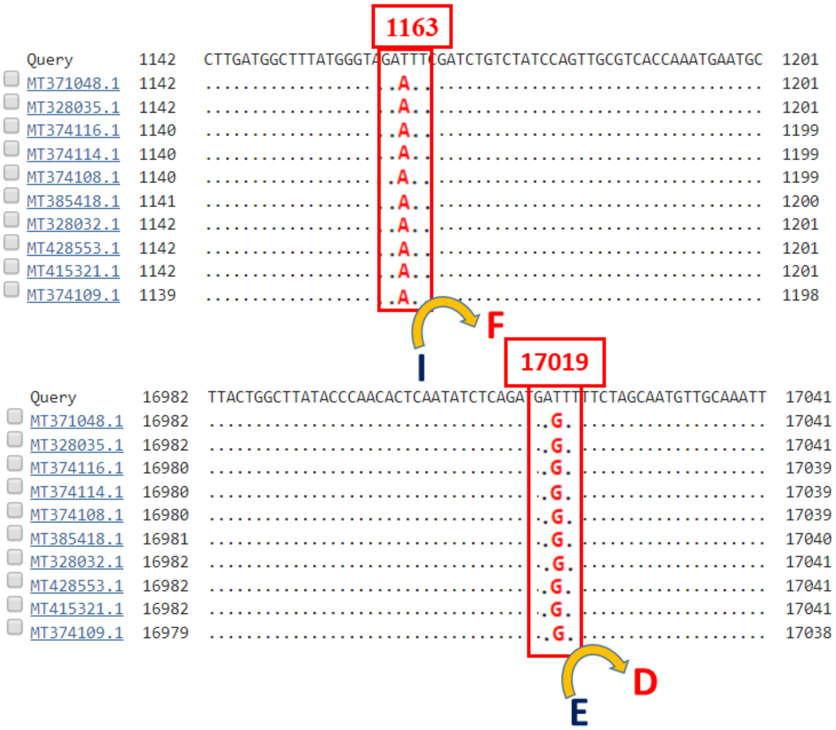

EPI ISL 430111/Russia/StPetersburg-RII6086S/2020
EPI ISL 438966/Japan/Donner22/2020
◆ EPI ISL 437912/Bangladesh/CHRF nCOV19 0001/2020
EPI ISL 425214/Spain/Valencia90/2020
EPI ISL 422427/Kuwait/KU008/2020
Cluster A.1
Cluster A
99
96
EPI ISL 437335/Turkey/HSGM-12059/2020
EPI ISL 437762/Saudi Arabia/KAUST-MAKKAH97/2020
EPI ISL 435292/Israel/CVL-nL33/2020
EPI ISL 436108/Taiwan/TSGH-23/2020
EPI ISL 442502/England/SHEF-D1F74/2020
EPI ISL 434709/Myanmar/NIH-4385/2020
67
56
Cluster A.2
EPI ISL 411902/Cambodia/0012/2020
EPI ISL 410301/Nepal/61/2020
EPI ISL 430444/Malaysia/IMR WC627/2020
EPI ISL 419313/Pakistan/Manga-KPK/2020
EPI ISL 43719/Indonesia/JKT-EIJK04/2020
EPI ISL 428258/USA/WI-162/2020
NC 045512/Wuhan-Hu-1/China/2019
Reference Sequence
EPI ISL 435048/Kazakhstan/NCB-5/2020
Cluster B
53
EPI ISL 437624/Thailand/SI200893-NT/2020
EPI ISL 412974/Italy/SPL1/2020
EPI ISL 428673/Sri Lanka/COV486/2020
EPI ISL 427813/Korea/KCDC2019/2020
EPI ISL 430015/Jordan/SR-057/2020
EPI ISL 435317/Vietnam/OUCRU01097/2020
70
88
Cluster C
68
EPI ISL 417444/Pakistan/Gilgit1/2020
EPI ISL 442523/Iran/ KHGRC-1.1-IPI-8206/2020
Cluster D
EPI ISL 438138/India/GMC-RR1191/2020
EPI ISL 434558/Philippines/PGC06/2020
EPI ISL 43178/Fujian/IM3520016T/2020
EPI ISL 435677/Brunei/5/2020
EPI ISL 435682/Singapore/77/2020
EPI ISL 437438/India/GBRC10/2020
94
79
Cluster E
0.000050

Figure 1. Phylogenetic analysis of 1st Bangladeshi SARS-CoV-2 complete genome sequences retrieved in this study, with available complete sequences from different countries (n=33 genome sequences). GISAID: Global Initiative on Sharing All Influenza Data; HKY: Hasegawa, Kishino, and Yano; MEGA 7: Molecular Evolutionary Genetics Analysis across Computing Platforms; SARS-CoV-2: severe acute respiratory syndrome coronavirus. Main clusters are highlighted in blue colors. The Wuhan reference genome is in navy color marked as reference sequence (GenBank accession number: NC_045512.2). The scale bar at the bottom of the tree represents 0.000050 nt substitutions per site. The cluster containing the viral sequence of the 1stBangladeshi SARS-Cov-2 genome sequence (GISAID accession ID: EPI_ISL_437912) is in red. The viral genome sequence from all other countries are divided in to fide clade marked as Cluster A to Cluster E. tree was built by using the best fitting substitution model (HKY) through MEGA 7 software.



```
  Query        1142   CTTGATGGCTTTATGGGTAGATTTCGATCTGTCTATCCAGTTGCGTCACCAAATGAATGC  1201
□ MT371048.1   1142   .....................A......................................  1201
□ MT328035.1   1142   .....................A......................................  1201
□ MT374116.1   1140   .....................A......................................  1199
□ MT374114.1   1140   .....................A......................................  1199
□ MT374108.1   1140   .....................A......................................  1199
□ MT385418.1   1141   .....................A......................................  1200
□ MT328032.1   1142   .....................A......................................  1201
□ MT428553.1   1142   .....................A......................................  1201
□ MT415321.1   1142   .....................A......................................  1201
□ MT374109.1   1139   .....................A......................................  1198

  Query        16982  TTACTGGCTTATACCCAACACTCAATATCTCAGATGATTTTTCTAGCAATGTTGCAAATT  17041
□ MT371048.1   16982  .....................................G......................  17041
□ MT328035.1   16982  .....................................G......................  17041
□ MT374116.1   16980  .....................................G......................  17039
□ MT374114.1   16980  .....................................G......................  17039
□ MT374108.1   16980  .....................................G......................  17039
□ MT385418.1   16981  .....................................G......................  17040
□ MT328032.1   16982  .....................................G......................  17041
□ MT428553.1   16982  .....................................G......................  17041
□ MT415321.1   16982  .....................................G......................  17041
□ MT374109.1   16979  .....................................G......................  17038
```

Figure 2: Genomic novelty analysis of Bangladeshi SARS-CoV-2 (EPI_ISL_437912) in comparison with the other NCBI submitted SARS-CoV-2 sequences. Genomic novelty analysis of SARS-CoV-2 Bangladeshi strains identified (a) One nt A has mutated to T at the genomic portion 1163 that causes nonsynonymous changes of amino acid I to F (b) One nt G has mutated to T at the genomic portion 17019 that causes nonsynonymous changes of amino acid E to D.

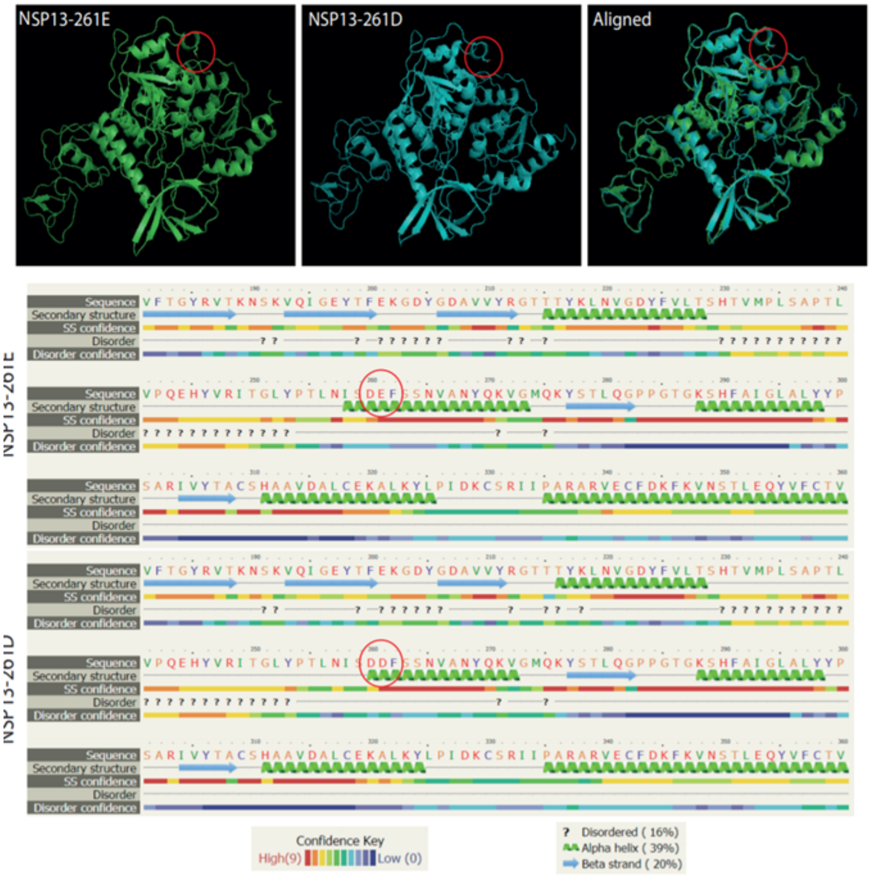


Figure 3a: The computational structure of SARS-CoV-2 E261D (shown as green cartoon). The computational model of the of SARS-CoV-2 E261D showed an alpha (Green) and beta helix(Blue).

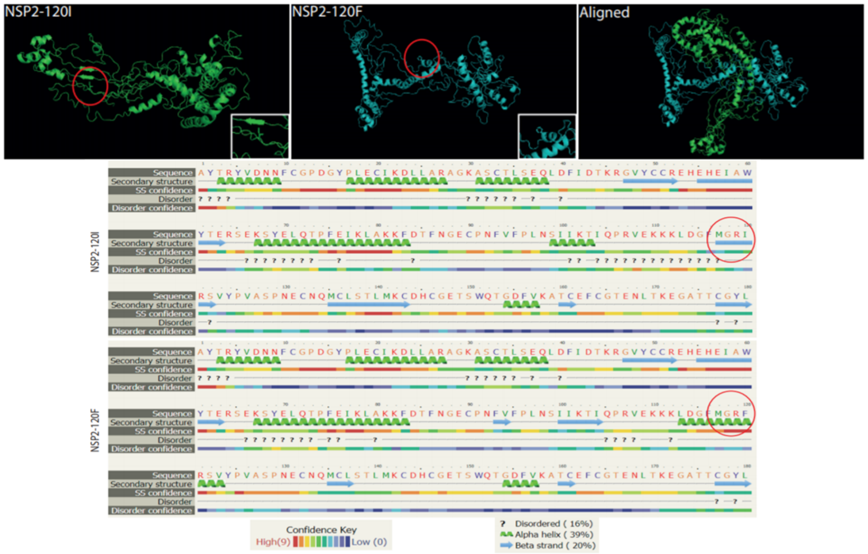


Figure 3b: The computational structure of SARS-CoV-2 I20F (shown as green cartoon). The computational model of the of SARS-CoV-2 E261D showed an alpha (Green) and beta helix (Blue).

Hosted file

`figure.docx` available at https://authorea.com/users/325047/articles/457169-phylogenetic-analysis-of-first-bangladeshi-sars-cov-2-strain-isolated-in-bangladesh-understanding-the-possible-origin-and-novel-mutations