ORIGINAL ARTICLE

Extensive genetic diversity with novel mutations in spike glycoprotein of severe acute respiratory syndrome coronavirus 2, Bangladesh in late 2020

S. Z. Afrin[1], S. K. Paul[2], J. A. Begum[3], S. A. Nasreen[1], S. Ahmed[1], F. U. Ahmad[4], M. A. Aziz[5], R. Parvin[3], M. S. Aung[6] and N. Kobayashi[6]

1) *Department of Microbiology, Mymensingh Medical College, Mymensingh,* 2) *Department of Microbiology, Netrokona Medical College, Netrokona,* 3) *Department of Pathology, Faculty of Veterinary Science, Bangladesh Agricultural University, Mymensingh,* 4) *Department of Microbiology, TMSS Medical College, Bogura,* 5) *Department of Microbiology, Rangpur Medical College, Rangpur, Bangladesh and* 6) *Department of Hygiene, Sapporo Medical University School of Medicine, Sapporo, Japan*

Abstract

In Bangladesh, coronavirus disease 2019 (COVID-19) has been highly prevalent during late 2020, with nearly 500 000 confirmed cases. In the present study, the spike (S) protein of severe acute respiratory coronavirus 2 (SARS-CoV-2) circulating in Bangladesh was genetically investigated to elucidate the diversity of mutations and their prevalence. The nucleotide sequence of the S protein gene was determined for 15 SARS-CoV-2 samples collected from eight divisions in Bangladesh, and analysed for mutations compared with the reference strain (hCoV-19/Wuhan/WIV04/2019). All the SARS-CoV-2 S genes were assigned to B.1 lineage in G clade, and individual S proteins had 1–25 mutations causing amino acid substitution/deletion. A total of 133 mutations were detected in 15 samples, with D614G being present in all the samples; 53 were novel mutations as of January 2021. On the receptor-binding domain, 21 substitutions including ten novel mutations were identified. Other novel mutations were located on the N-terminal domain (S1 subunit) and dispersed sites in the S2 subunit, including two substitutions that remove potential N-glycosylation sites. A P681R substitution adjacent to the furin cleavage site was detected in one sample. All the mutations detected were located on positions that are functionally linked to host transition, antigenic drift, host surface receptor binding or antibody recognition sites, and viral oligomerization interfaces, which presumably related to viral transmission and pathogenic capacity.

Keywords: Bangladesh, mutation, severe acute respiratory syndrome coronavirus 2, spike glycoprotein
Original Submission: 29 March 2021; **Accepted:** 20 April 2021
Article published online: 24 April 2021

Corresponding author: N. Kobayashi, Department of Hygiene, Sapporo Medical University School of Medicine, S-1 W-17, Chuo-ku, Sapporo, 060-8556, Japan.
Corresponding author: R. Parvin, Department of Pathology, Faculty of Veterinary Science, Bangladesh Agricultural University, Mymensingh, 2202, Bangladesh.
E-mails: **rokshana.parvin@bau.edu.bd (R. Parvin), nkobayas@sapmed.ac.jp (N. Kobayashi)**

Introduction

Coronavirus disease 2019 (COVID-19) pandemic is caused by a novel species of coronavirus, i.e. severe acute respiratory syndrome coronavirus type 2 (SARS-CoV-2) [1]. The virus was first recognized in Wuhan, China, in December 2019 and since then its worldwide spread has had a detrimental effect on global health and economy [2]. In Bangladesh, the first case of COVID-19 was identified on 8 March 2020 (https://corona.gov.bd/). Then, to prevent further human-to-human transmission, measures were scheduled quickly by the Government of Bangladesh including initial lockdown, social distancing, movement restrictions, closure of public and private facilities, and reduction of domestic and international air transport. Nevertheless, despite these strict measures, Bangladesh has been vulnerable to the current COVID-19 pandemic because it is one of the world's most densely populated countries. As a consequence, until the end of January 2021, a total of 535 230 cases with 8102 death have been recorded (https://corona.gov.bd/?gclid) officially in the country.

https://doi.org/10.1016/j.nmni.2021.100889

Given the evolving nature of the SARS-CoV-2 genome, drug and vaccine developers continue to be vigilant for the emergence of new variants or sub-strains of the virus [3]. Knowledge of the molecular background of circulating strains is vital for the appropriate selection of protective vaccine targets and to maintain molecular diagnostic tools for Bangladesh. Although structural proteins of SARS-CoV-2 show extremely high conservation, mutations have been identified, most frequently in the spike (S) protein as well as in the nucleocapsid protein, among global genomic data of this virus [4]. The S protein of coronaviruses is a trimeric glycoprotein that contains two subunits (S1 and S2) that mediate attachment and fusion of viral and cellular membranes, respectively [5,6]. In the process of viral entry into a cell, the receptor-binding domain (RBD) of the S1 subunit attaches to the host cell's receptor angiotensin-converting enzyme 2, while the S2 subunit mediates fusion of viral and cellular membranes [7]. For the process of virus–cell fusion, the S protein requires priming by host proteases in cleavage sites with a polybasic (furin) cleavage motif (RRAR) at the S1/S2 boundary [7,8]. Mutations of the S protein have been considered important because they are associated with alteration of infectivity and antigenicity of SARS-CoV-2 [9,10].

In Bangladesh, the SARS-CoV-2 genome has been analysed in several reports, revealing variation and frequency of mutations in all the viral proteins [11–17]. However, mutations on the S protein have not yet been well investigated and sufficient information is not available. In the present study, we analysed variation and frequency of missense mutations in the S protein of SARS-CoV-2 from geographically wide areas in Bangladesh. The results revealed extensive diversification of the S protein associated with a wide spectrum of mutations that may affect viral transmission and pathogenicity.

Materials and methods

Viral RNA samples from COVID-19 patients were collected from laboratories in eight divisions of Bangladesh (Barisal, Chittagong, Dhaka, Khulna, Mymensingh, Rajshahi, Rangpur, Sylhet) between July and December 2020. Initial detection of SARS-CoV-2 in nasopharyngeal swab samples was performed by quantitative RT-PCR targeting the ORF1ab and N genes using a commercial kit (Novel Coronavirus Nucleic Acid Diagnostic Kit, Sansure Biotech, Changsha, China). Among 229 samples collected, 15 samples with higher viral load, presumed by the Cq value of the ORF1ab and N genes, were selected for further analysis (see Supplementary material, Table S1).

Full-length of S gene was amplified by conventional RT-PCR with newly designed primer sets (see Supplementary material, Table S2). PCR products were purified and were subjected to Sanger-sequencing. The obtained sequences were assembled, edited and aligned using BioEdit software. Sequence data obtained in this study were deposited in the GISAID database, from which accession numbers were assigned (see Supplementary material, Table S3). A phylogenetic tree was generated using MEGA X software [18] applying the maximum likelihood method and the Tamura–Nei model [19]. Deduced amino acid sequences of S proteins were aligned and compared with the reference strain (hCoV-19/Wuhan/WIV04/2019). The major amino acid changes at the receptor binding site, antigenic site and cleavage site of the S protein were represented in a three-dimensional model using Swiss-Model (https://swissmodel.expasy.org/) software. This study was approved by the institutional review board of Mymensingh Medical College, Bangladesh (MMC/IRB/2020/290).

Results

Nucleotide sequences of 15 SARS-CoV-2 S genes were assigned to clade G (GISAID nomenclature) and further clustered with the B.1 lineage [20] with 98.1%–99.9% identity to the reference strain hCoV-19/Wuhan/WIV04/2019. Phylogenetically, these 15 S protein genes were scattered over several sub-branches together with viruses of this group distributed globally (see Supplementary material, Fig. S1), showing the highest nucleotide sequence identity (99.21%–100%) to strains from the USA, India, Bangladesh, the Netherlands, Colombia, Egypt, Australia, South Africa and Canada (see Supplementary material, Table S4).

A total of 189 single nucleotide polymorphisms by comparison with the S gene sequence of the reference strain were detected in 15 SARS-CoV-2 in Bangladesh (Table 1). Deduced amino acid sequences of 15 samples exhibited a total of 133 replacements (substitution, deletion) with 110 different types of amino acid substitution (see Supplementary material, Table S5). Individual S proteins had 1–25 amino acid replacements (average 8.9 mutations). Among the amino acid replacements identified in the present study, 80 were previously known and 53 were novel mutations as of January 2021.

The S1 subunit contained more amino acid substitutions than S2, almost equally distributed in the N-terminal domain and RBD, and others in the subdomains and furin cleavage site (FCS). Among the 15 Bangladeshi SARS-CoV-2 samples, previously reported substitutions in the S1 subunit were located mostly in the N-terminal domain and in the vicinity of the FCS (48 among 58 substitutions). In contrast, novel mutations in the S1 subunit (total 26 substitutions) were identified mainly in the RBD (ten substitutions) and subdomain–FCS (nine substitutions). On the S protein RBD, 21 mutations were

TABLE 1. Mutations identified in spike protein of 15 SARS-CoV-2 samples in Bangladesh compared with hCoV-19/Wuhan/WIV04/2019

Sample ID	No. of single nucleotide polymorphisms	No.of amino acid replacements	Amino acid replacement (substitution/deletion)	
			Known	Novel[a]
D3	32	7	D614G, P681R, I1081V	P491H, P499A, S514C, N1098D
D8	10	25	R158I, S162T, Q173H, V213L, Q218H, L244V, G257C, A260T, R273T, D614G, I664K, T676N, A845S, Q1201R	Y144del, R237T, N282D, K300T, A668G, I670K, Y1215F, G1219R, I1221K, A1222T, K1266N
C2	21	12	D614G, Q1180H, E1182D, N1187Y	C662W, I670K, C671G, V976G, D979E, V987L, V991M, K1266N
C3	30	20	D614G, Q779H, V785F, T883A, F888S, G889A, G891V, A892V, Q895R, G908C, S974L, D1260N	V781N, A893T, F898Y, M900K, Q901P, M902K, N907T, H1058Q
R2	12	9	L517P, A522S, P527L, T531N, D614G	P521H, S530Y, V991M, H1058Q
R5	1	1	D614G	None
B4	27	18	V3D, L18S, P272S, K378R, Y380F, V382L, S477I, T500I, L517I, D614G, K1073N	L5T, P295T, G381A, G496W, V512G, L518V, Q920P
B9	5	4	V11I, D614G, V620I	K1266N
S3	12	5	K304R, T307S, E309Q, D614G	H1058Q
S7	9	4	R158K, E298K, D614G, V620I	None
T16	2	1	D614G	None
T17	6	6	P561L, D614G, S758I, N764S, A771S	S359I
M1	17	2	D614G	K1266N
M77	3	17	L141I, M153I, A163V, N544Y, D574Y, R577S, P579L, D614G, Q675H, P809S	Y144del, K537T N556H, K557T, R567I, D568Y, G1035A
K6	2	2	D614G, Q675H	None
Total	189	133	80	53

del, deletion.
[a]Analysis includes the sequences available up to 15 January 2021.

identified. Among the ten novel substitutions on RBD, three substitutions (P491H, G496W, P499A) were located on the receptor-binding motif along with two known substitutions. Substitution D614G was detected in all samples, among which two samples had mutation Q675H. Substitution P681R, located adjacent to the FCS motif, was detected in one sample. Two novel substitutions, N282D and N1098D, were presumed to remove a potential N-glycosylation site. As depicted in Fig. 1, the positions of representative novel substitutions in RBD are mapped near the outermost edge of S protein, whereas N1098D is located in the root of the S2 subunit close to the viral envelope. The missense mutations identified in S proteins were predicted to affect antigenicity, receptor binding ability, and viral oligomerization depending on the functional domains in S protein (see Supplementary material, Table S6).

Discussion

Mutation rate of SARS-CoV-2 has been estimated to be 6×10^{-4} nucleotide/genome/year, which is lower than other RNA viruses [21,22]. However, minimal mutations in the genome, represented by a single D614G mutation in the S protein, have the capacity to alter the traits of a protein, affecting virus infectivity, clinical outcomes, and also epidemiology [23,24]. Accordingly, it is important to monitor genetic diversity and mutations of SARS-CoV-2 on a global scale during the ongoing pandemic.

In Bangladesh, almost solely the D614G substitution in SARS-CoV-2 S protein was noted and its predominance was revealed by genomic analysis [11,13–15], as observed in the present study. However, other mutations in the S protein have not yet been analysed in detail, and only a few have been described [11,16,17]. In a recent study by Hasan *et al.* [17], a missense mutation in the S protein was reported to be the third most frequent among viral proteins, following ORF1ab and N protein, by analysis of 371 viral genomes. In contrast, lower mutation rates of the S protein have been described elsewhere [14]. Although mutations in other viral proteins were not analysed in the present study, we identified high genetic diversity in the S protein, with missense mutation ranging from only one to 25 per sample. Therefore, associated with a nationwide pandemic in Bangladesh since late 2020, SARS-CoV-2 is suggested to have undergone rapid genetic evolution, acquiring various mutations. The presence of functional domains in the S protein underscores the importance of analysing mutations in this protein.

The D614G mutation in the S protein has been globally increasing since the middle of 2020 [4]. In Bangladesh, it was already dominant early in the pandemic (April–June 2020), accounting for almost 90% of the sequences analysed [14]. In late 2020, this mutation appeared to have overwhelmed the country, as observed in the present study. The D614G substitution changes the physicochemical characteristics of the S protein to cause more efficient binding to angiotensin-converting enzyme 2 receptors, leading to increased

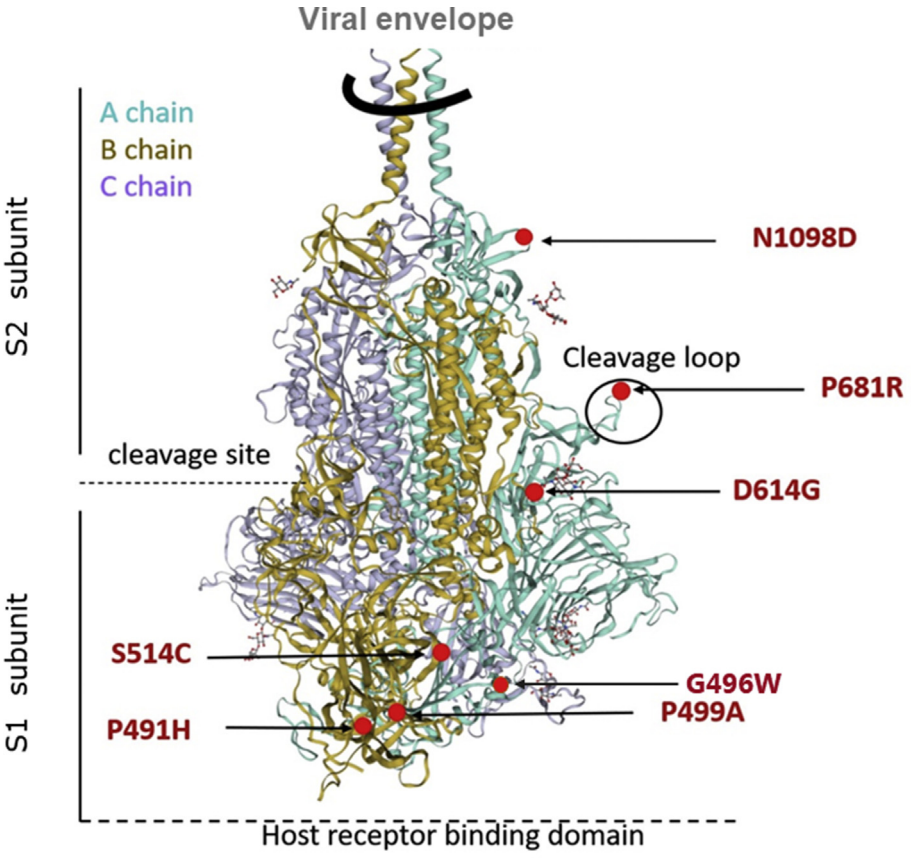

FIG. 1. Structural representation of spike glycoprotein of SARS-CoV-2. Positions of some novel and established substitutions are indicated with red points. The crystal structure shown in chain view was created using the online based modelling tool Swiss-Model (https://swissmodel.expasy.org).

infectivity of this variant [24,25]. A number of characteristic variants combined with the D614G substitution in the S glycoprotein have been identified previously and described as modifying the pattern of infection [9]. Among them, a D614G+Q675H variant was found in two samples in the present study.

In the present study, three novel substitutions were identified on the receptor-binding motif in RBD, suggesting their effect on alteration of viral infectivity. Although more evidence remains to be determined, the novel S protein mutations in SARS-CoV-2 in Bangladesh give rise to the possibility of a fitness advantage to target cells, which is likely to allow more successful person-to-person transmission. Another notable substitution was P681R, which is located adjacent to the FCS motif, which is highly conserved and important for pathogenesis of the virus [26,27]. On the other hand, loss of the FCS motif is related to attenuation of replication in cells and disease *in vivo* [26]. The P681R mutation was reported in only one study in Bangladesh [16], so appears to be still rare. Further surveillance may be necessary to monitor this mutation regarding its epidemiological and clinical impact. Mutations at N-glycosylation sites (amino acid 282 and 1098) were identified in the present study. Such glycosylation mutants had been reported previously and considered to be implicated in alteration of antigenic epitopes [9].

In conclusion, the present study revealed diverse nucleotide polymorphisms and amino acid substitutions in the S glycoprotein of SARS-CoV-2 in Bangladesh, suggesting rapid genetic evolution during the pandemic phase in late 2020. The epidemic of COVID-19 is still ongoing in Bangladesh, and the UK variant B.1.1.7 of SARS-CoV-2 was first identified in January 2021 [28], so continuous monitoring of genetic variation is necessary.

Funding

This research was supported by the project funded by The World Academy of Science (TWAS), grant number 20-284 RG/BIO/AS_G–FR3240314166 and NST (National Science and Technology) fellowship from the Ministry of Science and Technology, Bangladesh (no. 39.00.0000.012.002.05.20-04).

Conflict of interest

The authors declare no conflicts of interest.

Acknowledgements

We are grateful to Dr Safia Sultana (Syed Najrul Islam Medical College), Dr Md. Abdul Kalam (Chattogram Medical College), Dr S. M. Masudur Rahman (Khulna Medical college), Dr A.K.M. Akbar Kabir (Sher-e-Bangla Medical College) and Dr Shantanu Das (Sylhet MAG Osman Medical College) for cooperation with sample collection.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.nmni.2021.100889.

References

[1] Peeri NC, Shrestha N, Rahman MS, Zaki R, Tan Z, Bibi S, et al. The SARS, MERS and novel coronavirus (COVID-19) epidemics, the newest and biggest global health threats: what lessons have we learned? Int J Epidemiol 2020;49:717–26.

[2] Wang C, Horby PW, Hayden FG, Gao GF. A novel coronavirus outbreak of global health concern. Lancet 2020;395:470–3.

[3] Nakagawa S, Miyazawa T. Genome evolution of SARS-CoV-2 and its virological characteristics. Inflamm Regen 2020;40:17.

[4] Troyano-Hernáez P, Reinosa R, Holguín Á. Evolution of SARS-CoV-2 envelope, membrane, nucleocapsid, and spike structural proteins from the beginning of the pandemic to September 2020: a global and regional approach by epidemiological week. Viruses 2021;13:243.

[5] Bosch BJ, van der Zee R, de Haan CAM, Rottier PJM. The coronavirus spike protein is a class I virus fusion protein: structural and functional characterization of the fusion core complex. J Virol 2003;77:8801–11.

[6] Ou X, Liu Y, Lei X, Li P, Mi D, Ren L, et al. Characterization of spike glycoprotein of SARS-CoV-2 on virus entry and its immune cross-reactivity with SARS-CoV. Nat Commun 2020;11:1620.

[7] Hoffmann M, Kleine-Weber H, Schroeder S, Krüger N, Herrler T, Erichsen S, et al. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. Cell 2020;181:271–280.e8.

[8] Lau SY, Wang P, Mok BWY, Zhang AJ, Chu H, Lee ACY, et al. Attenuated SARS-CoV-2 variants with deletions at the S1/S2 junction. Emerg Microbe Infect 2020;9:837–42.

[9] Li Q, Wu J, Nie J, Zhang L, Hao H, Liu S, et al. The impact of mutations in SARS-CoV-2 spike on viral infectivity and antigenicity. Cell 2020;182:1284–94.

[10] Conceicao C, Thakur N, Human S, Kelly JT, Logan L, Bialy D, et al. The SARS-CoV-2 Spike protein has a broad tropism for mammalian ACE2 proteins. PLoS Biol 2020;18:e3001016.

[11] Parvez MSA, Rahman MM, Morshed MN, Rahman D, Anwar S, Hosen MJ. Genetic analysis of SARS-CoV-2 isolates collected from Bangladesh: insights into the origin, mutational spectrum and possible pathomechanism. Comput Biol Chem 2021;90:107413.

[12] Islam MT, Alam ARU, Sakib N, Hasan MS, Chakrovarty T, Tawyabur M, et al. A rapid and cost-effective multiplex ARMS-PCR method for the simultaneous genotyping of the circulating SARS-CoV-2 phylogenetic clades. J Med Virol 2021. https://doi.org/10.1002/jmv.26818.

[13] Saha O, Hossain MS, Rahaman MM. Genomic exploration light on multiple origin with potential parsimony-informative sites of the severe acute respiratory syndrome coronavirus 2 in Bangladesh. Gene Rep 2020;21:100951.

[14] Shishir TA, Naser IB, Faruque SM. In silico comparative genomics of SARS-CoV-2 to determine the source and diversity of the pathogen in Bangladesh. PLoS One 2021;16:e0245584.

[15] Akter S, Banu TA, Goswami B, Osman E, Uzzaman MS, Habib MA, et al. Coding-complete genome sequences of three SARS-CoV-2 strains from Bangladesh. Microbiol Resour Announc 2020;9. e00764-20.

[16] Adnan N, Khondoker MU, Rahman MS, Ahmed MF, Sharmin S, Sharif N. Coding-complete genome sequences and mutation profiles of nine SARS-CoV-2 strains detected from COVID-19 patients in Bangladesh. Microbiol Resour Announc 2021;10. e00124-21.

[17] Hasan MM, Das R, Rasheduzzaman M, Hussain MH, Muzahid NH, Salauddin A, et al. Global and local mutations in Bangladeshi SARS-CoV-2 genomes. Virus Res 2021:198390.

[18] Kumar S, Stecher G, Li M, Knyaz C, Tamura K, Mega X. Molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol 2018;35:1547–9.

[19] Tamura K, Nei M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 1993;10:512–26.

[20] Rambaut A, Holmes EC, O'Toole Á, Hill V, McCrone JT, Ruis C, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol 2020;5:1403–7.

[21] van Dorp L, Acman M, Richard D, Shaw LP, Ford CE, Ormond L, et al. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infect Genet Evol 2020;83:104351.

[22] Dearlove B, Lewitus E, Bai H, Li Y, Reeves DB, Joyce MG, et al. A SARS-CoV-2 vaccine candidate would likely match all currently circulating variants. Proc Natl Acad Sci USA 2020;117:23652–62.

[23] Callaway E. The coronavirus is mutating – does it matter? Nature 2020;585:174–7.

[24] Groves DC, Rowland-Jones SL, Angyal A. The D614G mutations in the SARS-CoV-2 spike protein: implications for viral infectivity, disease severity and vaccine design. Biochem Biophys Res Commun 2021;538:104–7.

[25] Yurkovetskiy L, Wang X, Pascal KE, Tomkins-Tinch C, Nyalile TP, Wang Y, et al. Structural and functional analysis of the D614G SARS-CoV-2 spike protein variant. Cell 2020;183:739–751.e8.

[26] Papa G, Mallery DL, Albecka A, Welch LG, Cattin-Ortolá J, Luptak J, et al. Furin cleavage of SARS-CoV-2 Spike promotes but is not essential for infection and cell-cell fusion. PLoS Pathog 2021;17:e1009246.

[27] Johnson BA, Xie X, Bailey AL, Kalveram B, Lokugamage KG, Muruato A. Loss of furin cleavage site attenuates SARS-CoV-2 pathogenesis. Nature 2021;591:293–9.

[28] Hossain ME, Rahman MM, Alam MS, Karim Y, Hoque AF, Rahman S, et al. Genome sequence of a SARS-CoV-2 strain from Bangladesh that is nearly identical to United Kingdom SARS-CoV-2 variant B.1.1.7. Microbiol Resour Announc 2021;10:e00100–21.