Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Genomic and evolutionary study from SARS-CoV-2 virus isolates from Bangladesh during the early stage of pandemic strongly correlate with European origin and not with China

Mohammad Fazle Alam Rabbi [a,b], Mala Khan [c,*], Mamudul Hasan Razu [c], Md. Imran Khan [a], Saam Hasan [a], Mauricio Chalita [d], Kazi Nadim Hasan [e], Abu Sufian [a,f], Md. Bayejid Hosen [f], Mohammed Nafiz Imtiaz Polol [a], Jannatun Naima [a], Kihyun Lee [d], Yeong Ouk Kim [d], Md. Mizanur Rahman [a,g], Jongsik Chun [d], Md. Abdul Khaleque [e], Zabed Bin Ahmed [c], Nur A. Hasan [h,i], Rita R. Colwell [i], Sharif Akhteruzzaman [a,j,**]

[a] *NGS Lab, DNA Solution Limited, Dhaka, Bangladesh*
[b] *Department of Soil, Water and Environment, University of Dhaka, Dhaka, Bangladesh*
[c] *Bangladesh Reference Institute for Chemical Measurements, BRiCM, Dhaka, Bangladesh*
[d] *ChunLab Inc., Seoul, South Korea*
[e] *Department of Biochemistry and Microbiology, School of Health and Life Sciences, North South University, Dhaka, Bangladesh*
[f] *National Forensic DNA Profiling Laboratory, Dhaka Medical College, Dhaka, Bangladesh*
[g] *NIPRO JMI Pharma, Dhaka, Bangladesh*
[h] *EzBiome Inc, Gaithersburg, MD, USA*
[i] *Center for Bioinformatics and Computational Biology, University of Maryland, College Park, USA*
[j] *Department of Genetic Engineering and Biotechnology, University of Dhaka, Dhaka, Bangladesh*

ARTICLE INFO

Keywords:
COVID-19
SARS-CoV-2
Phylogeny
Pandemic
GISAID
Whole genome sequencing

ABSTRACT

The goal of this study was to identify the genomic variants and determine molecular epidemiology of SARS-CoV-2 virus during the early pandemic stage in Bangladesh. Viral RNA was extracted, converted to cDNA, and amplified using Ion AmpliSeq™ SARS-CoV-2 Research Panel. 413 unique mutants from 151 viral isolates were identified. 80% of cases belongs to 8 mutants: 241C toT, 1163A toT, 3037C toT, 14408C toT, 23403A toG, 28881G toA, 28,882 G toA, and 28883G toC. Observed dominance of GR clade variants that have strong presence in Europe, suggesting European channel a possible entry route. Among 37 genomic mutants significantly associated with clinical symptoms, 3916CtoT (associated with sore-throat), 14408C to T (associated with cough-protection), 28881G to A, 28882G to A, and 28883G to C (associated with chest pain) were notable. These findings may inform future research platforms for disease management and epidemiological study.

1. Introduction

In December 2019, several cases of unknown pneumonia were reported in the Hubei province of China which raised concerns among world health experts [1]. The etiology was later diagnosed as a novel coronavirus and was dubbed by Chinese authorities as “COVID-19” or “2019-nCoV” [2,3]. The virus was later designated as Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) based on taxonomic and genetic relationship with the previously identified SARS-CoV virus [4]. It's high rate of transmissibility [5,6] allowed the virus to achieve global transmission rapidly [7,8]. The World Health Organization (WHO) announced COVID-19 as a pandemic on March 11, 2020 [9]. As of December 14, 2020 nearly 70 million infections have been reported with over 1.6 million deaths [10].

SARS-CoV-2 is a positive-sense single-stranded RNA virus believed to be transmitted in aerosols and common surfaces [11]. It has been shown to be adept at transmission and possesses a strong pathogenic capacity, raising the need for an in-depth understanding of its genetic

* Corresponding author.

** Corresponding author at: NGS Lab, DNA Solution Limited, Dhaka, Bangladesh.

E-mail addresses: bricmdg@yahoo.com (M. Khan), sharif_akhteruzzaman@yahoo.com (S. Akhteruzzaman).

https://doi.org/10.1016/j.ygeno.2022.110497
Received 15 May 2022; Received in revised form 19 September 2022; Accepted 25 September 2022
Available online 28 September 2022

characteristics [12,13]. Extensive molecular studies revealed that single-stranded, positive sense RNA genome of this virus is enclosed within a capsid. The capsid is composed of the nucleocapsid protein N. A membrane surrounds the capsid, which is made up of three proteins: membrane protein (M), envelope protein (E), and spike glycoprotein (S). These proteins play a crucial role in binding host receptors, mediating membrane fusio`1n, and facilitating virus entry into host cells. The 5′ cap and 3′ poly-A tail on the positive-stranded RNA genome enable translation from the host translation machinery. S (spike glycoprotein), N (nucleocapsid protein), M (membrane protein), and E (envelope protein) are structural proteins that are encoded at the 3′ end. By adhering to the viral genome, the nucleocapsid protein helps the genome pack against the interior surface of the envelope. S, M, and E proteins make up the viral envelope instead. The 3′-end encodes four structural proteins as well as nine accessory proteins (Orf3a, Orf3b, Orf6, Orf7a, Orf7b, Orf8, Orf9b, Orf9c, Orf10). Accessory proteins were thought to be crucial to virulence and host interaction and these proteins show high variability. At the 5′-end of SARS-CoV-2, a frameshift between two Orfs, Orf1a and Orf1b, allows the production of two polypeptides, which are then processed by proteases to create 16 non-structural proteins (Nsp1–16). These proteins are crucial for several steps in the virus infection cycle including viral replication, transcription, and the production of envelope proteins [14].

A large amount of information has accumulated with regard to genomic and proteomic mutations found subtypes of virus mainly circulating in developed parts of the world [5,15,16]. However, information on mutations of the SARS-CoV-2 virus circulating in low- and middle-income countries (LMIC) is sparse.

Bangladesh, a developing country in South Asia, is one of the most densely populated (above 1000 inhabitants/km2) [17]. The first COVID-19 case in Bangladesh was reported on March 08, 2020 [18]. Since then, the country has suffered a steeply rising number of new COVID-19 cases. As of August 23, 2022, >2 million COVID-19 cases have been reported, and >29,000 people have died [19]. The country also has a large population who are settled and work abroad, especially in the Middle East and Europe [20]. These workers tend to visit families in Bangladesh during the summer season, March to June. Furthermore, China is a strategic partner in economic development of Bangladesh, with significant traffic between these two countries [21]. These factors indicate possible routes by which the virus entered Bangladesh.

Whole genome sequences of the SARS-CoV-2 virus have been publicly deposited, most of which are accessible through the Global Initiative on Sharing All Influenza Data (GISAID). By December 7, 2021, almost 6.0 million viral sequences had been uploaded [22]. These data are crucial for genomic epidemiological studies. A large body of SARS-CoV-2 genomic information is available on strains from the developed world, but comparatively less information is available from resource-poor countries like Bangladesh [23]. In this study, we sequenced and comprehensively analysed the entire genomes of 151 SARS-CoV-2 strains from patients with varying severity of the disease in the early stages of the pandemic. To compare isolates from Bangladesh with those from other countries, a comparative analysis was performed that included additional SARS-CoV-2 genomes from strains that were simultaneously circulating in other parts of the world. Based on the above discussion, our aim was to determine the path of the SARS-CoV-2 outbreak in Bangladesh. Finally, we performed a machine learning-based analysis focused on the association of SARS-CoV-2 single mutation and clinically relevant parameters such as symptomatic status, fever, etc.

2. Results

2.1. Data quality

A total of 151 individuals positive for SARS-CoV-2 participated in this study. Patient gender, age, and geographical region were requested for all positive individuals. Although all positive individuals provide written consent to participate in the study, only 104 individuals provided metadata information. Among those 104 individuals, sixty-five (65) were male and 39 were female. The mean age was 41.09 ±1.75 (SEM). Geographical data analysis showed that the samples considered in this study were scattered throughout Dhaka city (Fig. 1), which is the capital and most densely populated city of Bangladesh [24]·

For all isolates from patients ($n = 151$) >98% of the sequence reads generated aligned successfully to the reference genome (NC_045512.2). Similarly, coverage for each genome was determined showing that their range was between 800× to >6000×, with a mean of 3000×. These indicated that all data generated in this study were high quality and could be further analysed with confidence.

2.2. Phylogenetic distribution of Bangladeshi isolates into distinct GISAID clades

Phylogenetic analysis of 151 SARS-CoV-2 genome sequences decoded in this study along with concurrent reference isolates, classified Bangladeshi isolates into three paraphyletic clades according to GISAID clade classification system (Fig. 2). This system classifies all SARS-CoV-2 isolates into 7 different clades defined by the presence of unique SNP signatures. Among the genomes of 151 SARS-CoV-2 isolates from Bangladesh, 132 genomes (87.4%) were placed in the GR clade, followed by 13 genomes (8.7%) in G and 6 genomes (3.9%)) into the GH clade. This appearance coincided with the recent deposition of genome sequences from Asia, where the majority of newly deposited genomes also were in the GR clade, followed by GH, G, and O [25]. An additional randomly selected 20 viral genomes were deposited in the GISAID database by various laboratories in Bangladesh. The clade distribution pattern of these other Bangladesh isolates was similar to our sample sets (17 belong to the GR clade and all other 3 belong to the G, S and L clades). Among them, one of the isolates (gene accession number MT5664683.1) belonging to the L clade was an interesting observation, mainly due to the fact that it had a very close resemblance to the Wuhan reference genome (NC_045512.2). GH, GR and G are the clades with the most member isolates worldwide. The V clade isolates are rarer but have discovered in Asia both previously and more recently. However, the L clade has been very rare in Asia recently and the scenario coincides with the Bangladesh isolates except the one isolate considered in our study. Gene bank accession numbers of 151 SARS-CoV-2 virus genome sequences used in this study are summarized in Supplementary Table 2.

2.3. Mutation analysis

The 151 isolates contained a combined total of 1753 single nucleotide Mutation (SNVs); minus those which occurred within ambiguous codons and were not considered for further analysis (Supplementary Table 1). These mutations were spread across the 412 positions in the SARS-CoV-2 genome. Eight of these mutations appeared in >79% of the isolates (Fig. 3A), namely the C to T change at 241, the A to T change at 1163, the C to T change at 3037, and at 14408 the change from A to G change at 23403, G to A change at 28881 and 28,882, and finally G to C change at 28883. Among these, 241C to T is a 5-UTR SNP mutation while 23,403 A to G is synonymous and all others are not synonymous. As further validation, all these mutations were also present in the other Bangladeshi isolates we included in our analysis (Fig. 3B). 1600 (91.27%) of our mutations occurred within the coding regions, while the rest occurred in UTR regions. Among the mutations that occur within genes on 1179 (73.68%) of them were nonsynonymous, resulting in amino acid changes. This included 244 unique mutations and 242 unique positions where base substitutions led to said mutations. Among one hundred and fifty one (151) SARS-CoV-2 virus isolates, DNAS_isl_29_MT860690 had the highest number of variants (total eighteen; 18), whereas, DNAS_isl_136_MT581417, DNAS_isl_138_MT581416 and DNAS_isl_148_MT566437 had the least number of mutations (only six;

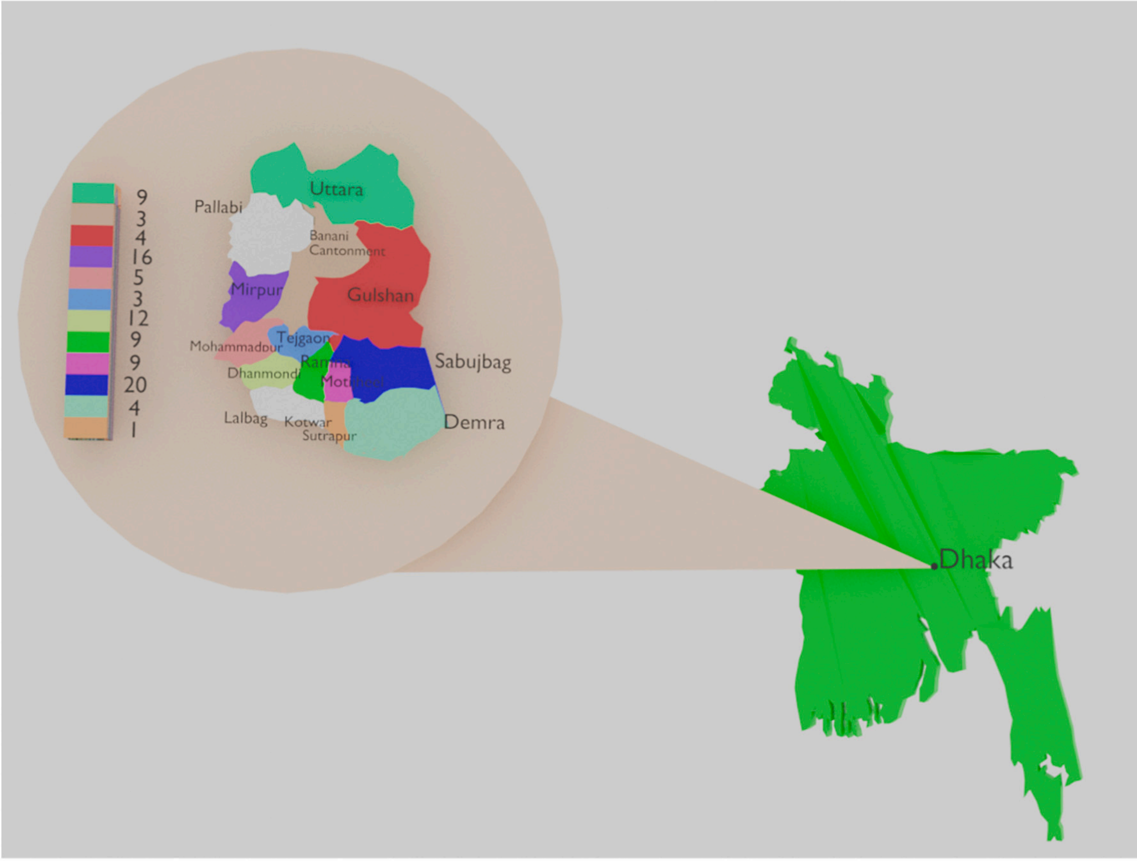

Fig. 1. Schematic diagram of geographical distribution of COVID-19 samples.
A total of 151 COVID-19 positive patients were considered in this study. Among them, 95 individuals gave consent to reveal their geographical location. Among them, it is clear the majority of samples came from the communities in or around the Sabujbag, Dhanmondi, Mirpur, Uttara and Ramna areas. A relatively fewer numbers of samples came from the central and southern most regions of Dhaka.

6). Among our isolates, DNAS_isl_29_MT860690 had the highest number of non-synonymous mutation (total fifteen; 15). The amino acid mutations with the highest occurrence were the D to G change in the S protein caused by the mutation at nucleotide position 23,403, the R to K and G to R change in the N protein caused by the multiple nucleotide mutations from nucleotide position 28,881 to 28,883, the P to L change in ORF1ab caused by the 14,408 mutation, and the I to F change also in ORF1ab resulting from the mutation at nucleotide position 1163. All of these mutations were in over 120 (79%) of our samples. There was one position with more than one kind of base substitution among our samples. This was position 28,727 where we found a G to A change as well as a G to T change, occurring in different samples. For amino acid variants, there were two positions in the genome where multiple mutations led to different amino acid changes. First was the aforementioned 28,727, where different variants led either to an A to S or an A to T change in the N protein. The other was the amino acid substitution caused by the 28,883 mutation. While this mutation occurs alongside those at 28881 and 28,882 as an MNV, the 28,883 base is part of a different codon than the other two. In most cases the G to C mutation at this position led to a G to R substitution in the N gene product. However, there was one isolate which also contained a mutation at position 28,884. This altered the resultant amino acid substitution to G to Q change instead (Supplementary Table 3).

2.3.1. Spike protein homology modeling

We have modeled structures of SARS-CoV-2 spike protein using DNAS_isl_29_MT860690, DNAS_isl_136_MT581417, DNAS_isl_138_MT581416, DNAS_isl_148_MT566437 sequences as well as with the reference sequence (NC_045512.2) (Supplementary Fig. 1). We have showed the Ramachandran Plot analysis of all our modeled structures and our modeled structures have favorable region >93% which indicates that our modeled structures are acceptable (Supplementary Fig. 2) (acceptable range for favorable region >85%).

2.4. Gene distribution of variants

ORF1ab contained the highest number of mutations, which was expected considering- it is the largest among the SARS-CoV-2 genes. The nucleocapsid phosphoprotein encoded by *N* gene had the second highest number of mutations, followed by the surface glycoprotein encoded by the *S* gene. The remaining genes had fewer mutations compared to these three. Lowest number of mutations was in *ORF6*, *ORF7b* and envelope protein encoded by *E* gene. Distribution of unique mutation among the genes followed a similar pattern as the distribution of total variants. *ORF1ab* contained highest number of unique mutants followed by *S* and *N* genes (Table 1).

The number of times each possible amino acid change occurred with a given gene was determined. Fig. 4, heat map, shows the frequency of occurrence for each type of change for all 11 genes, finding a total of 76 types of amino acid changes in our analysis. Overall, D to G change was the most common (161 of 1175 amino acid variants), R to K (132), I to F (121), P to L (133) and G to R (132) were most frequent. Other common amino acid substitutions included A to V (15 for ORF1ab, ORF3a and S proteins), T to I (34), L to F (25), Q to H (28), and S to L (22). It should be noted that most amino acid changes are the same variant occurring in a large number of samples.

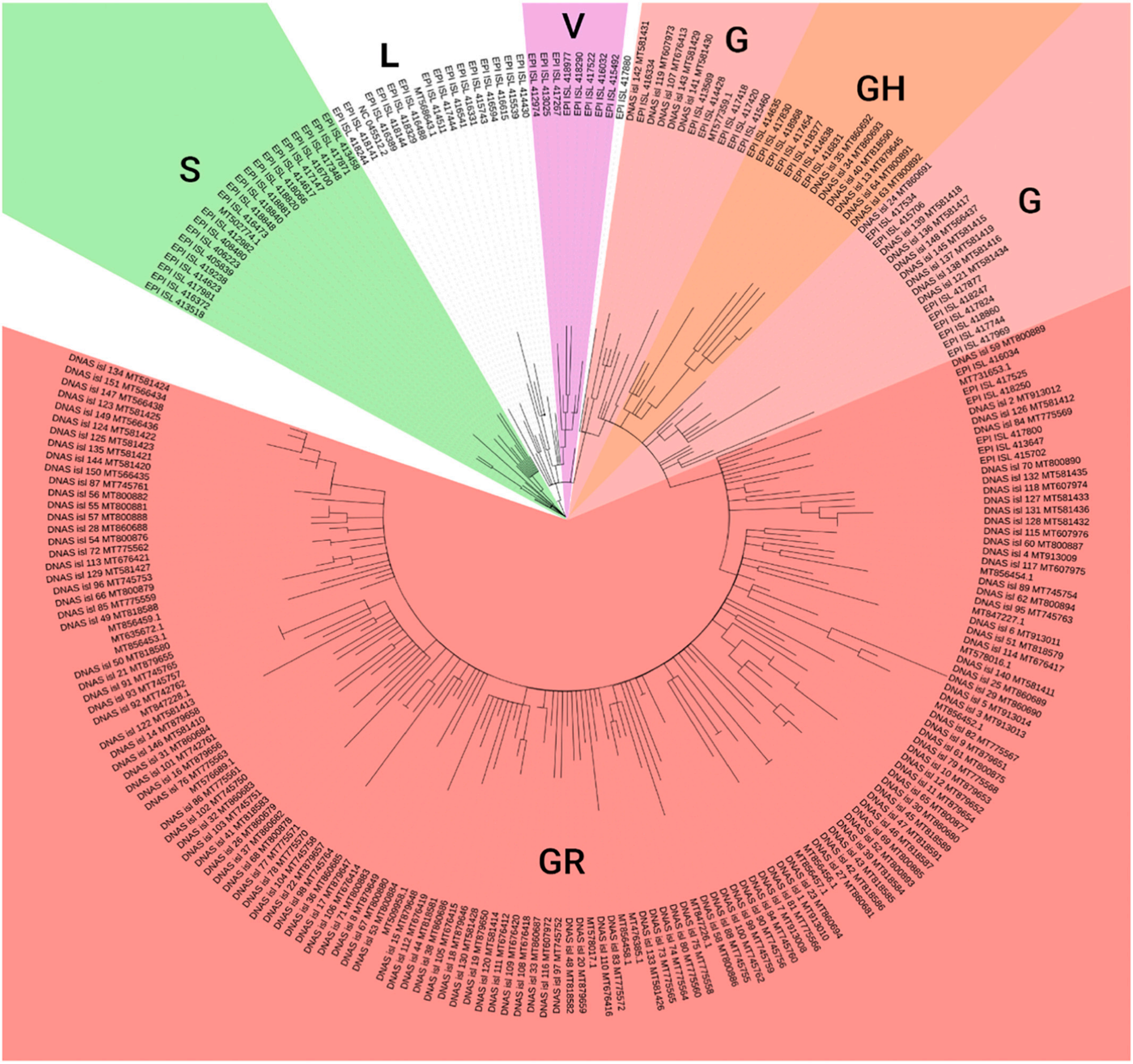

Fig. 2. Radial SNP based phylogenetic tree displaying the classification of 151 SARS-CoV-2 virus isolates according to GISAID clade classification system. 132 out of 151 isolates considered in this study belonged to GR clade, 13 belonged to G clade and rest belonged to GH clade. Among the 20 randomly picked isolates from other Bangladeshi laboratories, 17 belonged to GR clade, 1 belonged to G clade, 1 to S clade and another to L clade which is where the Wuhan reference (NC_045512.2) genome is placed. This is in line with the recent identification and deposition of genomes from Asia. Majority of recently deposited Asian isolates have belonged to the GR clade.

2.5. *Epidemiological sub-typing of Bangladeshi isolates*

Isolates were grouped based on the EzCOVID19 SNP profile subtyping system as described in the Methodology section (“EzCOVID19” n.d.). The 151 isolates comprised eight groups based on type assignment according to the EzCOVID19 algorithm (Fig. 5). Group 1 was most closely related to type 9. Both shared a common horizontal distance from the root and grouped together in the same branch. No mismatch was observed in the 41 SNP sites between this group and type 9. Nine virus isolates belonged to this group. Type 9 are most prevalent in Europe. Groups 2 and 3 were most closely related to type 2. Group 3 was identical to this type with regards to 41 SNP profile. However, group 2 contained one SNP difference at position 14,408. The base at this position was G, whereas for type 2 it was T. Three isolates belonged to group 2, while one belonged to group 3. This subtype is most prevalent in Europe, and found in parts of Asia, Africa and North America as shown in Fig. 5. For group 4, the closest subtype was type 61. Only 1 fall into group 4 and was not an exact match with type 61, which contains the 11,083 G to T mutation and 27,046C to T mutation. Both were absent in the group 4. It is worth noting that type 61 occurs exclusively in Europe. Study Group 5 closest related subtype was 15. The SNP profile was identical between these two. This type has a far more global distribution and occurs in North America, Europe and Asia. They group together with two member branches. Six belong to this study group. Group 6 shared highest similarity with type 26 with regards to the branch distance. Only one isolate joined group 6 and also has a SNP

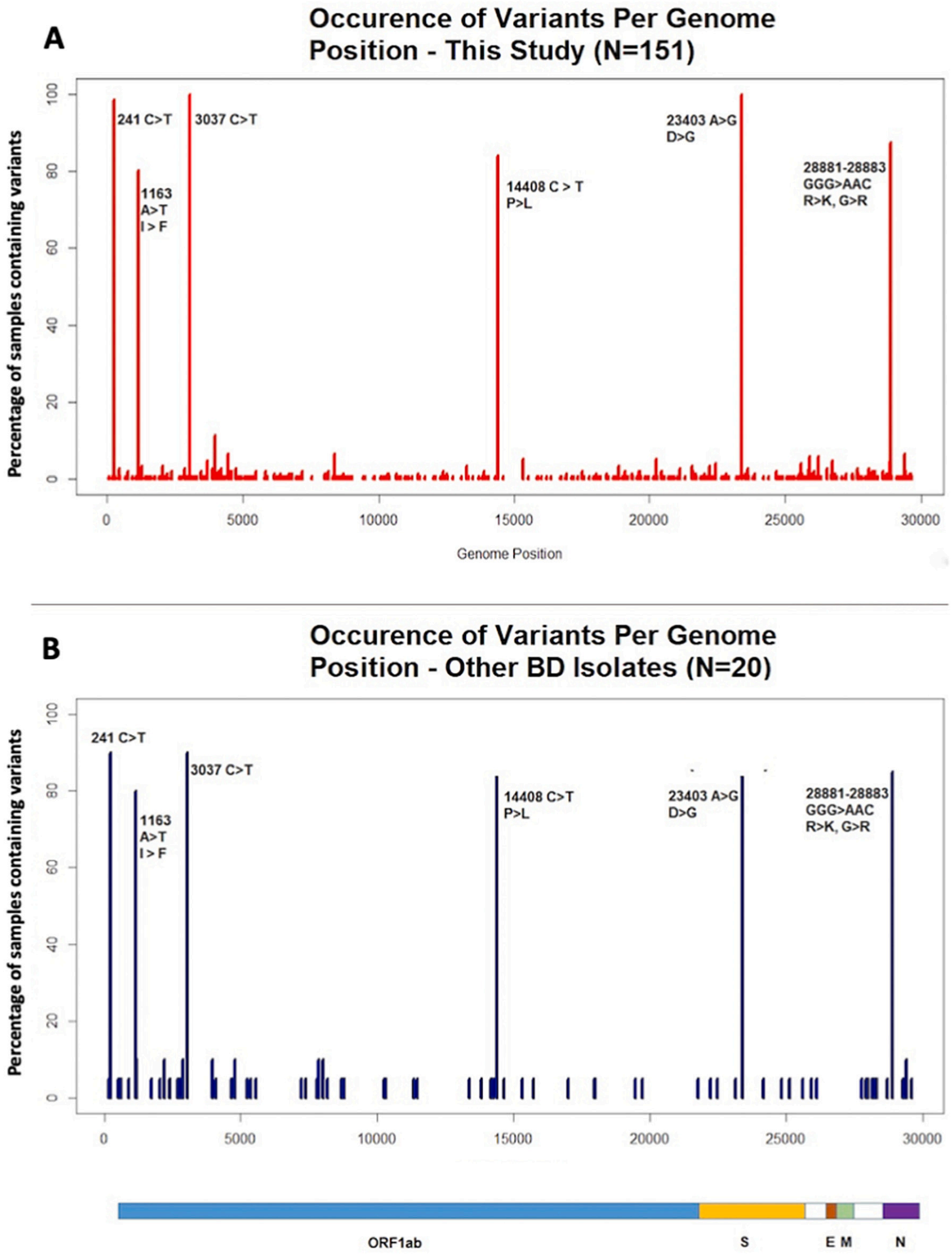

Fig. 3. Distribution of variants across the SARS-CoV-2 Genome, as compared to the Wuhan Reference Sequence (NC_045512.2).
A total of 1753 variants across 151 isolates spreading over 412 positions (3A in red). 8 of these variants occurred in over 120 isolates. These were the C to T change at 241, the A to T change at 1163, the C to T change at 3037 and at 14408, the A to G change at 23403, G to A change at 28881 and 28,882, and finally the G to C change at 28883. The 241C to T variant is the only one among 8 to occur in a non-coding region (it is a 5′ UTR SNP). The most common among the variants was the 23,403 A to G change, which results in the D to G mutation at position 614 of the spike glycoprotein. The 8 major variants were also found in a number of other Bangladeshi isolates submitted in GISAID (3B in blue). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

mismatch with type 26. Type 26 contained 25,563 G to T mutation, while our isolate did not. Isolates belonging to type 26 has been seen mostly in Europe and to a lesser degree, in South and North America. The closest matching subtype for Group 7 was type 64. Only one 1 isolate from our study belonged to study group 7 and it has contained a SNP mismatch with its most closely related type. Type 64 contained 11,083 G to T mutation, group 7 isolate did not. This is also a predominantly European subtype, while also occurring in Asia and South

Table 1
Total gene variants observed in SARS-CoV-2 virus isolates.

Gene	Total mutations	Unique mutations	Common mutations (with other BD samples)
envelope protein	2	2	0
membrane glycoprotein	23	12	0
nucleocapsid phosphoprotein	447	33	3
ORF10 protein	9	7	1
ORF1ab polyprotein	788	251	9
ORF3a protein	64	30	0
ORF6 protein	3	2	0
ORF7a protein	12	7	0
ORF7b	2	2	0
ORF8 protein	15	9	1
surface glycoprotein	235	53	3

The *ORF1ab* contained highest number of mutations and is the longest among eleven (11) genes in SARS-CoV-2 virus. The nucleocapsid phosphoprotein encoded by *N* gene had the second highest number of mutations, followed by the surface glyco protein encoded by the *S gene.*

America. Lastly, group 8 shared closest common ancestor with type 4, with key SNP profiles exact match and comprised the majority of the SARS-CoV-2 isolates (130 of 151). Virus isolates belonging to type 4 are also predominantly in Europe; with limited presence in Asia, Oceania, and North America.

2.6. Clinical importance of variants

A total of 37 mutations gave *p*–values <0.05 selected parameters (Supplementary Table 3), one of which was significantly associated with more than one disease symptoms, mainly 3961C to T mutant, shown an important determinant for patients developing sore throat and diarrhoea. The 14,408C to T, significantly associated with coughing, where individuals infected with a subtype with the variant appeared to suffer less from coughing. The particular mutant was also linked with a host of other parameter determined using the random forest model. However, neither of the statistical test returned significant *p*-value for them. Mutants 22,199 G to T, 19593C to T, 13902 T to C, 774C to T, and 21,597C to T were significantly associated with development of sore throat, 28,881 G to A, 28882 G to A, and 28,883 G to C with chest pain, 21,123 G to T with anorexia and 29,118C to T, 28178 G to T, 29262 G to T with pneumonia.

There were other mutations which coincided with individuals lacking specific disease characteristics for example, 4105 G to T, 3456 A to G, 28305 A to G, 26051 G to A, 24685 T to C not loss of taste and smell.

Some mutantss, despite significant association with certain clinical factors, were not consistent with respect to specific factors. 28,292C to A, 4300 G to T, 26526 G to T, 17193 G to T, 2731 G to A, 98264 A to G, 12025C to T, 8311C to T, 714 G to A, 8366 G to A, 18859 G to T, 9416 G to A, 20808 G to A were all significantly related to onset or protection from skin rash. 28,079 G to T and 3053 G to T were associated with itching and redness of eyes. Therefore, in all cases, the nature of the relationship could not be established based on the observations.

Finally, three mutants were significantly associated with overall symptomatic status of the patients, namely 28,580 G to A, 2363C to T, and the 3871 G to T, and were significantly more likely to be asymptomatic. Table 2 summarizes a few of these clinically relevant mutations and their associated symptoms based on significance.

3. Discussion

In this cross-sectional study, we have observed that SARS-CoV-2

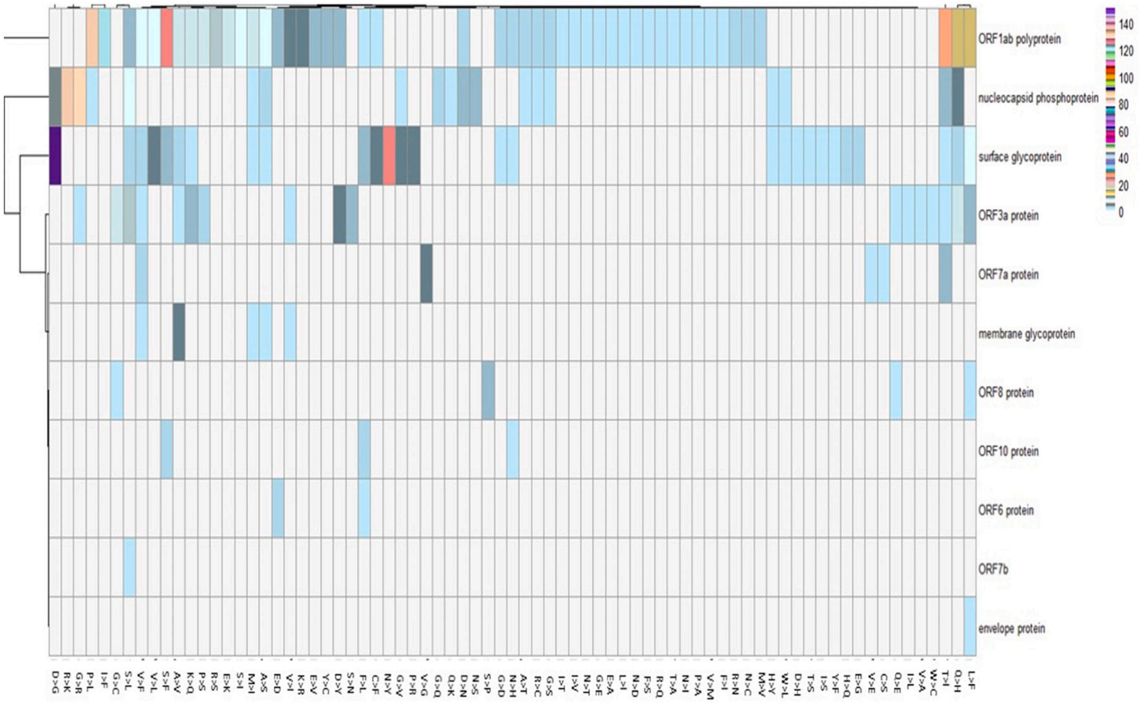

Fig. 4. Heatmap showing the amino acid change per gene in our SARS-CoV-2 virus isolates.
A total of 76 different kinds of amino acid changes could be observed across the 11 SARS-CoV-2 genes. Overall, D to G changes were the most abundant (161 out of 1175 amino acid variants). Aside from that, R to K (132 times), I to F (121 times), P to L (133 times) and G to R (132 times) occurred most frequently among the samples. Other common amino acid substitutions included A to V (occurring 15 times across the ORF1ab, ORF3a and S proteins), T to I (34 times), L to F (25 times), Q to H (28 times), and S to L (22 times). It should be noted that the majority of occurrences for each amino acid change are in fact the same variant occurring in a large number of samples. The majority of changes occurred only once, as indicated by the boxes with a light blue colour. As a result majority of these were only found in 1 of the genes. The white colour indicates that (i.e. the corresponding amino acid change for each of the white cells did not occur at all in the concerned gene). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

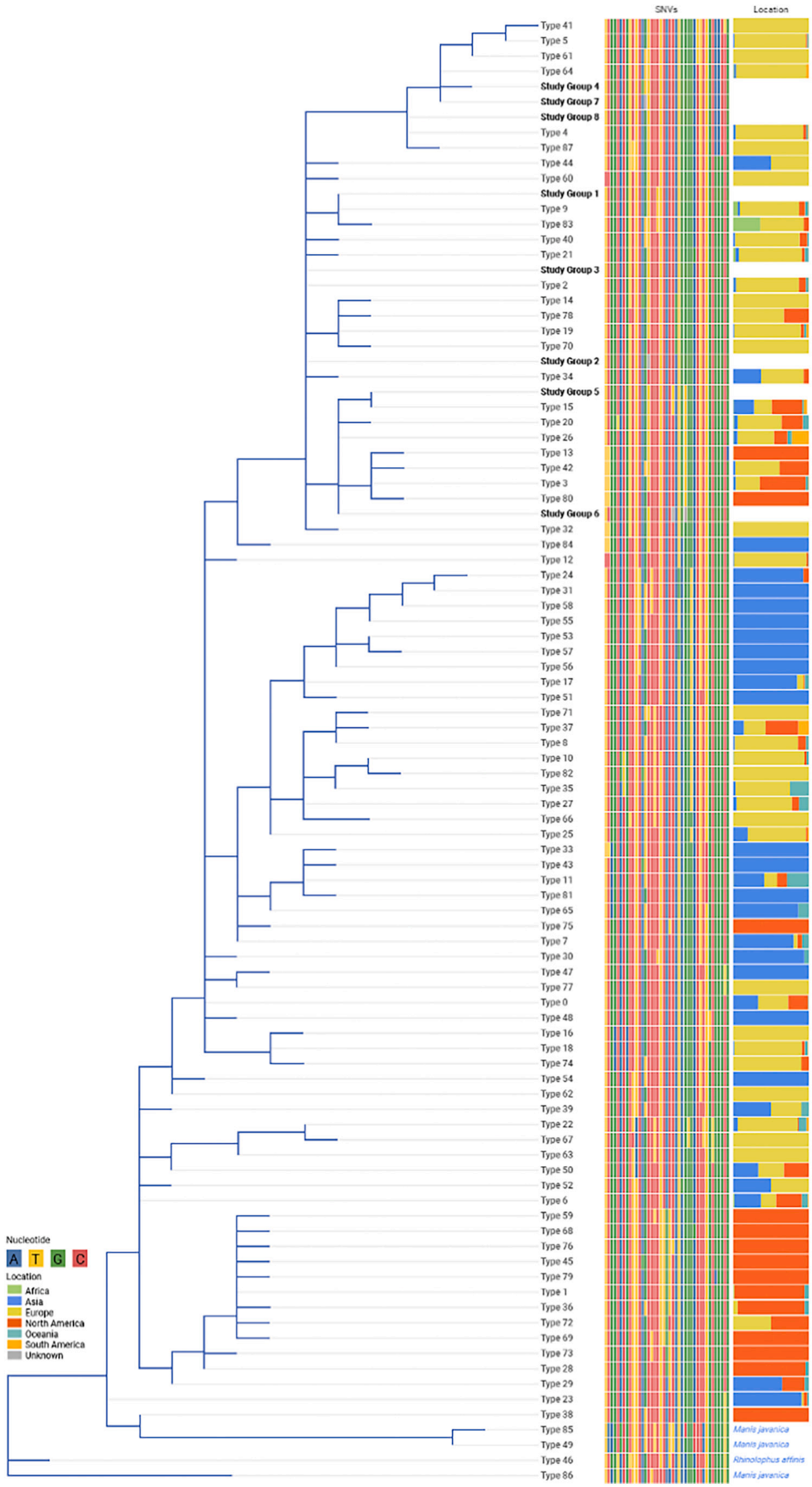

Fig. 5. SNP based phylogenetic tree displaying the classification of 151 SARS-CoV-2 virus isolates according to EZCovid19 classification system for SARS-CoV-2 virus.

Study group 1 (*N* = 9) closely relates to type 9 with an exact match to SNP profile. Study group 2 (*N* = 3) and Study group 3 (*N* = 1) closely related to type 2. Study group 2 had a mismatch with type 2 at 14408 position (14,408 T to G). Study Group 4 (N = 1) closely matched with type 61 with two mismatch at 11083 G to T and 27,046C to T variant. Study Group 5 (*N* = 6) has an exact match with type 15 in terms of branch distance and SNP profile. Study Group 6 (N = 6) closely matched with Type 26 with a mismatch at 25563 G to T variant. Study Group 7 (N = 1) has close match with type 64 with a SNP mismatch at 11083 G to T variant. Lastly, study group 8 (*N* = 130) shared its closest common ancestor with type 4 with an exact match. All the close related types of our virus isolates has a string presence in Europe. to EZCovid19 classification system for SARS-CoV-2 virus.

Table 2
Possible clinical importance of variants.

Variant(s)	Variant Type	Disease State Association	Effect on Disease Phenotype	Fisher's Test P-value of Significance	Extra Comment
14,408C to T	SNV	Cough	Protects from Cough	0.02691376	
3961C to T	SNV	Sore Throat	Causes Sore Throat	0.000570637	
	SNV	Diarrhoea	Protects from Diarrhoea	0.03658523	
28881G to A, 28882G to A, 28883G to C	SNV	Chest Pain	Causes Chest Pain	0.02478889	Co-variants
29118C to T, 28178G to T, 29262G to T	Co-Variants (SNVs)	Pneumonia	Causes Pneumonia	0.009615385	Co-variants. Only 1 patient developed pneumonia however they were infected with a subtype containing unique variants.

A total of 37 variants returned *p* values <0.05 for various parameters. The 8 most significant variants is summarized here. 3961C to T variant showed association with sore throat and diarrhoea development. The 14,408C to T inversely associated with cough development. The 28,881 G to A, 28882 G to A, and 28,883 G to C were associated with chest pain, and 29,118C to T, 28178 G to T, 29262 G to T were associated with pneumonia.

virus isolates are related European subtypes, concurrent with other studies of Bangladeshi SARS-CoV-2 [26]. SARS-CoV-2 virus isolated from neighbouring country, India, also were predominantly European sub-type [27]. This would suggest the virus entered Bangladesh via Europe, but multiple points of entry remain a very real possibility, reinforced by the observation that many of our isolates belonged to found in Africa, South America and North America. The GISAID clade classification offers a new perspective and most of our isolates belonged to clades that have recently seen high prevalence in Asia. However, possibilities that the virus may simply have entered these Asian countries through European sources earlier. Based on the data presented here, there is the possibility of SARS-CoV-2 entered Bangladesh through European countries, Asian countries, or both. A completely decisive conclusion is difficult to draw from SNP phylogeny. However, there is sufficient evidence for informed guess of European or Asian entry, with confidence inclining to the former.

The mutations provided several referral points of insight. The eight most common mutations were 241C to T, 1163 A to T, 3037C to T, 14408C to T, 23403 G to A, 28881 G to A, 28882 G to A, and 28,883 G to T, and were also present in the majority of other Bangladeshi isolates we included in our analysis. The 23,403 G to mutant, which results in D to G substitution at 614 position in the S protein [Supplementary Fig. 1:(A), (B),(C) &(D)], is one of the most prevalent amino acid mutations of the virus. Two amino acids mutation (Position 18 and 614) is found in the homology modeled structure of spike protein of DNAS_isl_29_MT860690 [Supplementary Fig. 1:(D)] sequence. In particular, it is a defining signature of the L, GH, G and GR clades. The majority of recently sequenced European isolates belongs to the latter three, lending further credibility to European origin claim since a large portion of recently sequenced Asian isolates have been classified as belonging to other clades (not the most common six).

Presence of an L clade member among the Bangladeshi isolates opens up an interesting perspective. The 28,882 mutant, with those at 241, 3037, and 23,403 are characteristic markers for the L clade which contains the Wuhan reference genome and other early Chinese isolates. An evolutionary timeline would suggest one possible sequence would be SARS-CoV-2 entering Bangladesh via a Chinese source for example, several students had to be evacuated from Wuhan, arriving in Bangladesh midway into the outbreak. Hence, the aforementioned four mutants in our samples. The virus responded to local selection pressures and incorporated the other common mutants, such as 14,408, 1163, and those mutants at 28881 and 28,883, which appear to have become co-mutants. One reason why this may have sound more credible is the complete absence of a number of common European mutants in our samples. Mutants such as 14,805C to T, 1440 G to A, 2558C to T and others were not present in any of our isolates, nor were they present in the other Bangladeshi samples analysed. Further 11,083 G to T were also rare, occurring in only one or a few more samples. An issue with this logic could be that recently sequenced isolates from Europe have generally belonged to the G clade, i.e. containing mutants at 241, 3037, and 23,403 and not 11,083, the other most common European mutant of the clade typing scheme. This raises question whether the major route of transmission arose when the more recent European isolates arrived in Bangladesh. However, the possibility of virus subtypes in different parts of the world accumulating similar mutations independently must be considered.

A few of the amino acid mutants showed clear prevalence compared to others. Six occurred in >80 of the isolates. The most common were 614 D to G for the S protein, which occurred in all samples, followed by, in decreasing order of frequency, 203 R to K and the 204 G to R also in N protein (132), P to L at 323 in RNA Polymerase of ORF1ab (127), and I to F in NSP2 of ORF1ab (121). The N protein R to K mutant in particular is curious. One of the nucleotide mutants, 28,881 G to A, was an earlier mutant been reported. It was found predominantly in European isolates, but nor two other co-variants, 28,882 G to A and 28,883 G to C. Subsequently, 28,882 was detected and used as a marker in the GISAID's clade classification. It's perplexing when these three mutants became associated (association is presumed since they occur together whenever they have been found in a genome). One possibility be that the 28,881 and 28,882 mutants arose separately. The 28,882 genotype perhaps accumulated the 28,881 and 28,883 mutants, as it spread across Asia (28,881 variant was not observed in Asia before). While the 28,881 genotype isolate did the same and incorporated the 28,882 and 28,883 mutant, completing the worldwide distribution pattern of these three mutants that observed today [28]. From the Bangladesh standpoint, both European and Asian origins remain a possibility, with regard to arrival of strains with this genotype. The European origin theory remains possible as other common European mutants are largely absent, meaning less likely Europe to Bangladesh transmission. Finally, in one isolate there is a 28,884 G to A mutant eliminating the change of 204 G to R mutants. Instead, the four nucleotide MNV leads to G to Q at 204 of the N protein. While seemingly similar, glutamine is not positively charged in its side chain, unlike arginine. Thus far it is indicated, that there is a lower mortality rate in Bangladesh compared many other parts of the world, the majority of the viral strains active in Bangladesh have to date harboured G to R change. Here an isolate incorporating a new mutation that, while not reversing the glycine to arginine substitution, does nonetheless alter it from a positively charged amino acid to one that is neutral (though polar).

There has not been significant focus this far on linking between individual SARS-CoV-2 mutants and the disease manifestation associated with the virus. By employing machine learning and statistics, it was possible to report a significant association between mutants and a specific clinical factors. Two mutants, by using random forest model, were 14,408C to T and 3916C to T. The second is a synonymous change. The first however results in a proline to leucine mutation in the RNA polymerase protein. Although it's fisher's test *p* value was only significant for the development of cough, the potential importance for overall symptomatic status deserves further attention. In general, individuals infected with the virus subtype containing this mutant were asymptomatic. The RNA polymerase is critical for viral replication and alteration in function may well affect a function and consequently prevent onset of symptoms. In addition the 28,881–28,883 multiple nucleotide mutant showed strong association with onset of chest pain. Heart failure (with other organ failures) has been suggested as a major causes of fatality in SARS-CoV-2 [29]. These three variants appeared to be correlated with the onset of chest pain, suggesting a key pathogenic determinant.

4. Conclusion

Our goal in this study was to identify the phylogenetic origin and genetic variation among the SARS-CoV-2 isolates in Bangladesh. The findings indicate the 151 virus isolates from our study has strongest phylogenetic link with European isolates; although this link cannot be completely verified with this data alone. In addition, we identified 8 mutants that are very common among the Bangladeshi mutants considered in this study. These mutants have been found in most parts of the world and have been validated by other studies carried out in Bangladesh. This study enriches our knowledge revealing the phylogenetic relationship of SARS-CoV-2 virus circulating in Bangladesh which opens up new area of research to combat with this pandemic efficiently.

5. Materials and methods

5.1. Ethical approval

The cross-sectional study included 151 Bangladeshi patients diagnosed positive for COVID-19 based on real-time reverse transcriptase PCR (rRT-PCR). Samples were collected from April 28, 2020 to July 21, 2020 at the Outdoor Patient Unit (OPD) of the Central Police Hospital (CPH) and other tertiary medical centres in Dhaka, Bangladesh. Bangladesh Reference Institute for Chemical Measurements (BRICM) in collaboration with, DNA Solution Ltd. (DNAS) provided the COVID-19 diagnosis and carried out subsequent whole-genome sequencing. All procedures in the study were according to ethical standards of the Helsinki Declaration of 1975, as revised in 2000 [30]. Informed consent was obtained from each individual providing sample. The study protocol was approved by the Bangladesh Council of Scientific and Industrial Research's (BCSIR) ethics review committee (Ref No# 5600.8400.02.037.20).

5.2. Sample collection and real-time PCR

Oro-pharyngeal swabs from suspected patients were collected in Government approved viral transport medium (VTM) and sent to DNAS through a cool box. Samples were tested for SARS-CoV-2 RNA using a Government approved commercial real-time one-step COVID-19 test PCR kit (Sansure Biotech Inc., Changsha, China) according to the manufacturer's instructions. The real-time PCR kit uses PCR fluorescent probe technology and targets two genes, ORF 1 and conserved coding regions of the nucleocapsid protein N gene. A positive internal control of human ribonuclease P (RNAase P) was used along with a positive and a negative control to neutralize the presence of PCR inhibitors. Real-time PCR was performed in an ABI7500 Fast DX instrument (Thermo Fisher Scientific, Massachusetts, USA). Samples with a ct value of <30 for both viral target genes were selected for subsequent viral RNA isolation and whole genome sequencing.

5.3. RNA extraction and cDNA preparation

RNA extraction was performed using the QIAamp®DSP Virus Spin Kit (Qiagen, Hilden, Germany) according to the instructions for use. Briefly, 200 μL of VTM containing the oropharyngeal swab was used as starting material for viral RNA extraction using silica membrane technology. The samples were lysed, bound to the silica membrane column, washed to remove impurities and eluted with RNase-free elution buffer. The cDNA was prepared on the same day using both the random hexamers and the oligo dT primers using the ProtoScript II First Strand cDNA Synthesis Kit (NEB, Ipswich, MA, USA). Prepared cDNA was stored at −20 °C until further use.

5.4. Library preparation and whole-genome sequencing

Ion AmpliSeq™ SARS-CoV-2 Research Panel (Thermo Fisher Scientific, Massachusetts, USA) was used to amplify the SARS-CoV-2 genome using prepared cDNA as a template. The panel contains 237 pairs of specific primers covering >99% of the SARS-CoV-2 genome. Amplified fragments were carried forward to prepare libraries for massive parallel sequencing using Ion AmpliSeq™ Library Kit Plus (Thermo Fisher Scientific, Massachusetts, USA), following manufacturer's instructions. Each of the prepared libraries was diluted to 100pM and pooled together for clonal amplification on the Ion One Touch 2 instrument. Clonally amplified libraries were enriched on using Ion One Touch ES followed by loading the enriched libraries on an Ion 530 chip. 15–20 samples were multiplexed simultaneously on the 530 chip during each run.

5.5. Bioinformatic analysis

Genome assembly of the raw data was performed by the EzCOVID19 [31] cloud service provided on the EzBioCloud website [32]. Assembly is performed by aligning reads to a predefined reference genome (NC_045512.2) while preserving the unique variations of the input raw data by creating a consensus genome. The consensus genome was then compared to the same reference genome to calculate single nucleotide variations (SNV) and positions. The SNVs were compared to GISAID clade variation markers [33]. Fig. 1 was generated by extracting all SNVs provided by EzCOVID19 and using RAxML [34] using all default parameters.

The phylogenetic and group typing analysis was accomplished using the EzCOVID19 cloud service, where a pre-build type grouping system based on occurrence of signature mutants was provided. In brief, EzCOVID19 considered 2761 SARS-CoV-2 genomes available at GISAID until April 01, 2020. Using a pairwise alignment approach (Myer-Miller's method) each/all 2761 genomes matched against Wuhan-Hu-1 reference genome (NC_045512.2) were aligned. From the resulting alignment, homopolymeric stretches of bases that cause frameshift errors were manually removed. The alignment matrix was then searched for mutation sites and, in this process, sites at which $\geq$ 99% genomes showed a valid nucleotide character (not gap or ambiguous) were used. Among the mutations positions, sites with $\geq$ 1% minor allele frequency (to avoid using sites that only have infrequent/spurious mutations) were selected. This resulted in 41 SNV sites (T514, C1059, G1397, G1440, C2416, A2480,C2558, G2891, C3037, C8782, T9477, C9962, A10323, G11083, C14408, C14724, C14805, C15324, T17247, C17747, A17858, C18060, C18877, A20268, T21584, A23403, G25563, G25979, G26144, A26530, C27046, T28144, C28657, T28688, G28851, C28863, G28881, G28882, G28883, C29095, G29553) which resulted in 88 unique allele combinations considered types. All isolates were typed according to typing system and isolates belonging to the same group were considered a similar study group.

The clinical importance of each of the mutant present in our samples was assessed. To do this clinical metadata for 104 of the 151 individuals providing samples were collected. Clinical parameters included fever, skin rash, diarrhoea, sore throat, chest pain, pneumonia, cough, anorexia, redness and itching of eyes, and overall symptomatic status of each individual (i.e. asymptomatic, mildly symptomatic, or severely symptomatic).

A random forest model was implemented to determine association between all mutants and each clinical factor (Andy Liaw, Wiener, and

Andy Liaw 2018). Mutants classified as important determinants of the category variable (clinical trait in question) were selected for further statistical analysis. Chi square and fisher's exact test were performed for each mutants to establish whether effect of the presence of each mutant was significant, with respect to the selected clinical factor. The *p*-value threshold for significance was set to 0.05.

5.5.1. Homology modeling of spike protein

We have downloaded the nucleotide fasta sequence of spike protein from the NCBI database. Amino acids fasta sequence is extracted from Expasy translate [35]. For Structure modeling we used SWISS-MODEL [36]. Structures assessments were done in MolProbity software.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2022.110497.

Funding information

The study was funded by the Ministry of science and Technology, Peoples republic of Bangladesh. The funding number is 5600.8400.02.037.20.

Author statement

MFR contributed in the conceptualization. AS, MBH, MIK, MHR contributed in methodology, investigation, formal analysis. MNP and JN contributed in data curation. SH, MC, KL, YOK and JC carried out formal analysis, software, validation and visualization. MIK and SH wrote the original draft. MFR, KNH, MMR, MK, MAK, NAH, RRC SA and ZBA review and editing the manuscript aMK contributed in supervision, project administration and funding acquisition.

Conflicts of interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Data availability

Data will be made available on request.

References

[1] W. Du, S. Han, Q. Li, Z. Zhang, Epidemic update of COVID-19 in Hubei Province compared with other regions in China, Int. J. Infect. Dis. 95 (2020) 321–325.

[2] F. Wu, et al., A new coronavirus associated with human respiratory disease in China, Nature 579 (2020) 265–269.

[3] P. Zhou, et al., A pneumonia outbreak associated with a new coronavirus of probable bat origin, Nature 579 (2020) 270–273.

[4] M. Pal, G. Berhanu, C. Desalegn, V. Kandi, Severe acute respiratory syndrome Coronavirus-2 (SARS-CoV-2): an update, Cureus 12 (2020).

[5] P. Skums, A. Kirpich, P. Icer Baykal, A. Zelikovsky, G. Chowell, Global transmission network of SARS-CoV-2: from outbreak to pandemic, medRxiv Prepr. Serv. Heal. Sci. (2020), https://doi.org/10.1101/2020.03.22.20041145.

[6] J. Chen, Pathogenicity and transmissibility of 2019-nCoV—A quick overview and comparison with other emerging viruses, Microbes Infect. 22 (2020) 69–71.

[7] Y. Roussel, A. Giraud-gatineau, M. Jimeno, J. Rolain, Since January 2020 Elsevier has Created a COVID-19 Resource Centre with Free Information in English and Mandarin on the Novel Coronavirus COVID- 19. The COVID-19 Resource Centre is Hosted on Elsevier Connect, The Company ' s Public News and Information, 2020.

[8] Y. Yang, et al., The deadly Coronaviruses: the 2003 SARS pandemic and the 2020 novel Coronavirus epidemic in China, the company' s public news and information, J. Autoimmun. 109 (2020), 102487.

[9] WHO, Director-General's Opening Remarks at the Media Briefing on COVID-19, Available at, https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19—11-march-2020, 11 March 2020.

[10] Coronavirus Disease (COVID-19) Situation Reports, 2022.

[11] H.J. Maier, E. Bickerton, P. Britton, Coronaviruses: methods and protocols, Coronaviruses Methods Protoc. 1282 (2015) 1–282.

[12] T.M. Belete, A review on promising vaccine development progress for COVID-19 disease, Vacunas 21 (2020) 121–128.

[13] F. Amanat, F. Krammer, SARS-CoV-2 vaccines: status report, Immunity 52 (2020) 583–589.

[14] G. Mariano, et al., Structural characterization of SARS-CoV-2: where we are, and where we need to be, Front. Mol. Bios. 7 (2020).

[15] N. Biswas, P. Majumder, Analysis of RNA sequences of 3636 SARS-CoV-2 collected from 55 countries reveals selective sweep of one virus type, Indian J. Med. Res. 151 (2020) 450–458.

[16] D. Mercatelli, F.M. Giorgi, Geographic and genomic distribution of SARS-CoV-2 mutations, Front. Microbiol. 11 (2020) 1–13.

[17] Population density (people per sq. km of land area) - Bangladesh | Data, 2022.

[18] S.K. Dey, M.M. Rahman, U.R. Siddiqi, A. Howlader, Exploring epidemiological behavior of novel coronavirus (COVID-19) outbreak in Bangladesh, SN Compr. Clin. Med. 1 (2020), https://doi.org/10.1007/s42399-020-00477-9.

[19] Coronavirus Disease (COVID-2019) Bangladesh Situation Reports, Available at, https://www.who.int/bangladesh/emergencies/coronavirus-disease-(covid-19) update/coronavirus-disease-(covid-2019)-bangladesh-situation-reports (Accessed: 23rd August 2022).

[20] M. Islam, N. Migration from Bangladesh and Overseas Employment Policy, 2022.

[21] M.J. Uddin, M. Bhuiyan, Sino-Bangladesh relations: an appraisal, biiss J. 32 (2011) 1–24.

[22] Y. Shu, J. McCauley, GISAID: global initiative on sharing all influenza data – from vision to reality, Eurosurveillance 22 (2017) 2–4.

[23] S. Anwar, M. Nasrullah, M.J. Hosen, COVID-19 and Bangladesh: challenges and how to address them, Front. Public Health 8 (2020) 154.

[24] Population of Cities in Bangladesh, Available at, https://worldpopulationreview.com/countries/cities/bangladesh, 2020 (Accessed: 27th October 2020).

[25] GISAID Initiative, Available at, https://www.epicov.org/epi3/frontend#lightbox104973013 (Accessed: 3rd November 2020).

[26] S. Saha, et al., Complete genome sequence of a novel coronavirus (SARS-CoV-2) isolate from Bangladesh, Microbiol. Resour. Announc. 9 (2020).

[27] R. Devendran, M. Kumar, S. Chakraborty, Genome analysis of SARS-CoV-2 isolates occurring in India: present scenario, Indian J. Public Health 64 (2020) 147.

[28] M. Pachetti, et al., Emerging SARS-CoV-2 mutation hot spots include a novel RNA dependent-RNA polymerase variant, J. Transl. Med. 18 (2020) 1–9.

[29] L. Wu, et al., SARS-CoV-2 and cardiovascular complications: from molecular mechanisms to pharmaceutical management, Biochem. Pharmacol. 178 (2020), 114114.

[30] World Medical Association declaration of Helsinki, Ethical principles for medical research involving human subjects, JAMA (2013), https://doi.org/10.1001/jama.2013.281053.

[31] EzCOVID19, Available at, https://www.ezbiocloud.net/tools/sc2/ (Accessed: 27th October 2020).

[32] S.H. Yoon, et al., Introducing EzBioCloud: a taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies, Int. J. Syst. Evol. Microbiol. 67 (2017) 1613–1617.

[33] GISAID, Clade and Lineage Nomenclature Aids in Genomic Epidemiology of active hCoV-19 Viruses, Available at, https://www.gisaid.org/references/statementsclarifications/clade-and-lineage-nomenclature-aids-in-genomic-epidemiology-of-active-hcov-19-viruses/ (Accessed: 27th October 2020).

[34] A. Stamatakis, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies, Bioinformatics 30 (2014) 1312–1313.

[35] E. Gasteiger, A. Gattiker, C. Hoogland, I. Ivanyi, R.D. Appel, A. Bairoch, ExPASy: theproteomics server for in-depth protein knowledge and analysis, Nucleic Acids Res. 31 (13) (2003) 3784–3788.

[36] A. Waterhouse, M. Bertoni, S. Bienert, G. Studer, G. Tauriello, R. Gumienny, et al., SWISS-MODEL: homology modelling of protein structures and complexes, Nucleic Acids Res. 46 (W1) (2018) W296–W303.