Novel mutations identified from whole-genome sequencing of SARS-CoV-2 isolated from Noakhali, Bangladesh

Maqsud Hossain (muhammad.maqsud@northsouth.edu)
North South University

Tahrima Saiha Huq
North South University

Aura Rahman
North South University

Md. Aminul Islam
Noakhali Science and Technology University

Syeda Naushin Tabassum
North South University

Kazi Nadim Hasan
North South University

Abdul Khaleque
North South University

Abdus Sadique
North South University

Mohammad Salim Hossain
Noakhali Science and Technology University

Newaz Mohammed Bahadur
Noakhali Science and Technology University

Firoz Ahmed
Noakhali Science and Technology University

Hasan Mahmud Reza
North South University

Research Article

Keywords: novel mutations, SARS-CoV-2, phylodynamic analysis, Bangladesh

Posted Date: April 30th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-437228/v1

Abstract

Whole-genome sequencing is increasingly being used to investigate the spatial and temporal distribution of viral pathogens including the Severe Acute Respiratory Syndrome Coronavirus Variant 2 (SARS-CoV-2) which is responsible for the ongoing COVID-19 pandemic. In this study, we determined 55 complete genome sequences of SARS-CoV-2 strains isolated from patients from Noakhali, a South-Eastern district in Bangladesh. Variant analysis of our sequenced genomes identified sixteen rare variations in S, six in N, two in M, one in E protein and the S protein variation, Y204F, identified in two of our sequenced strains, has not been reported from any other countries in the GISAID database. Comparison of the prevalence pattern across the country showed GH clade lineages B.1.36 and B.1.36.16 to be abundant in Noakhali and the South-Eastern region of Chittagong when compared to the rest of the country. Phylodynamic analysis of our sequenced genomes revealed that the virus was estimated to be evolving at the rate of $1.065 \text{ X } 10^{-4}$ subs/site/year. The study results demonstrated the necessity of initiating a concerted, country-wide genomics surveillance effort to determine any novel mutation of functional significance, understanding virus evolution, transmission, and spread in Bangladesh.

Short running title: Genome sequencing of Noakhali isolates SARS-Cov-2 in Bangladesh

Introduction

In December 2019, several pneumonia cases of unknown etiology were notified to World Health Organization (WHO). The outbreak was associated with a seafood Market in Wuhan, and the causative agent was identified as a Beta-coronavirus, later named Severe Acute Respiratory Syndrome Coronavirus Variant 2 (SARS-CoV-2). Due to extensive global transmission, the WHO declared SARS-Cov-2 mediated COVID-19 as a pandemic on March 11, 2020[1].

On February 01, 2020, several Bangladeshi people were brought back from Wuhan and tested SARS-CoV-2 negative in the Institute of Epidemiology, Disease Control and Research (IEDCR). The first SARS-CoV-2 positive detected in Bangladesh was on March 08, 2020. As of February 15, 2021, 541,038 confirmed cases and 8,285 deaths were reported in Bangladesh. According to Our World In Data (https://ourworldindata.org/), Bangladesh has administered at least 2.08 million first-dose vaccines (February 20, 2021). Since the initiation of vaccination, Bangladesh holds 11th position in the global vaccination race and has been ahead of several European Union countries (February 07, 2021). While preparing the manuscript the highest number of cases observed on July 09, 2020 (~4000 cases/day); a second short wave started in early November, which started declining in mid-December 2020 (~1000 cases/day). The number decreased to ~300 cases/day on February 06, 2021, when the country's vaccination started. However, recently we observed a sharp increase in the number of infected people and currently raised to >7000 cases/day (https://www.worldometers.info/coronavirus/country/bangladesh/).

Coronaviruses are a family of positive-sense, single-stranded, non-segmented RNA viruses with a genome size of approximately 30kbp. The SARS-CoV-2 genome encodes several smaller open reading frames

(ORFs) such as ORF1ab, ORF3a, ORF6, ORF7a, ORF7b, ORF8, and ORF10 which are predicted to encode for the replicase polyprotein, the spike (S) glycoprotein, envelope (E), membrane (M), nucleocapsid (N) proteins, accessory proteins, and other non-structural proteins (NSP)[2–4].

The mutation rate of RNA viruses has an exceptionally high correlation with increased infection capabilities and viral fitness. Pathogen genome sequencing has been playing a pivotal role in understanding the epidemiology of infectious diseases and also allowed to identify genes responsible for host-pathogen interactions. Moreover, constrained genomic regions identified by genomic approaches can be preferential target for drug and vaccine development to avoid the escape of rapidly evolving pathogen like SARS-CoV-2[1,5], in which new mutation hotspots are emerging[2,3].

In a developing country like Bangladesh community-level testing is not a feasible option and passive surveillance is unlikely to detect asymptomatic or presymptomatic infections. Genome sequencing can be considered as an alternative to overcome this problem in understanding viral transmission dynamics and were instrumental during the West African Ebolavirus outbreak in 2014–2016 and again during the emergence of Zika virus in the Americas in 2015–2016[6,7].

Bangladesh is a low and middle-income country (LMIC) with over 166 million people, where 63% of the population live in rural regions[8]. There is a scarcity of information on the mutations taking place in the SARS-CoV-2 genome circulating in Bangladesh due to its resource-poor settings. Although Bangladesh Government measures have expanded testing capacity and infrastructure development, it has been challenging to investigate the prevalence in resource-poor settings, especially outside Dhaka, Bangladesh's capital city. Several public and private research organizations in Bangladesh have completed SARS-CoV-2 genome sequencing on a reasonable number of samples collected, majorly from Dhaka and Chittagong City. Genomic databases of those are available in both Global Initiative on Sharing All Influenza Data (GISAID) and National Centre for Biotechnology Information (NCBI). However, the epidemiological data and the type of variants circulating outside Dhaka and Chittagong have been poorly investigated and made it difficult to understand the characteristics of SARS-CoV-2 circulating in different regions of the country, complicating decision making on taking effective interventions, including therapeutics and vaccines[8].

NSU Genome Research Institute (NGRI) at North South University, the first private University in Bangladesh, took the initiative of doing SARS-CoV-2 genome sequencing on real-time PCR positive samples collected from Noakhali, a district located in the South-eastern region of Bangladesh. Noakhali Science and Technology University was Bangladesh Govt. approved national COVID-19 testing center. This study reports the complete genome sequences of 56 SARS-CoV-2 from the Noakhali region of Bangladesh. The study also reports variant and phylodynamic analysis results obtained by comparing with the country and global specific SARS-CoV-2 sequencing datasets.

Materials And Methods

Sample collection and RT-qPCR

Nasopharyngeal and oropharyngeal swabs were collected from specific health complexes or Sadar hospital, stored in normal saline and transported in refrigerated condition within 24 hours after collection in Noakhali Science and Technology University (NSTU) Covid-19 Diagnostic Laboratory.

The study was reviewed and approved by the ethics committee of the Directorate General of Health Sciences (DGHS), Bangladesh and by National Research Ethics Committee (NREC) (Ref.: BMRCAIREC/2O 1 9 -2022 I 7 08) of Bangladesh Medical Research Council (BMRC).

The study was conducted according to the relevant guidelines and regulations in accordance with the Declaration of Helsinki. Relevant demographic, clinical and laboratory data were retrieved from the clinical records of the patient and signed written informed consent was obtained from participants and/or their legal guardians.

RNA was isolated from positive SARS-CoV-2 stored patient samples using QIAamp Viral RNA Mini Kit (QIAGEN) extraction Kit. We used two gene-based RT-PCR COVID-19 detection technique where the novel coronavirus (2019-nCoV) ORF 1ab and the specific conserved sequence of coding nucleocapsid protein N genes were used as target regions as per the recommendation of the Institute of Epidemiology, Disease Control and Research (IEDCR), Bangladesh. After extraction, RNAs were analyzed for the detection of SARS-CoV-2 by RT-PCR (CFX96, BioRad) using the Sansure RT-PCR kit (Sansure Biotech Inc., China).

cDNA synthesis, library preparation and sequencing

In total, we sequenced 56 COVID-19 positive samples between 5 June 2020 to 20 December 2020 available from NSTU COVID-19 testing laboratory, Noakhali, Bangladesh. Viral cDNA was synthesized with ProtoScript II First Strand cDNA Synthesis Kit (ThermoFisher, Waltham, MA, USA), followed by the NEBNext® Ultra™ II Non-Directional RNA Second Strand Synthesis (New England BioLabs, USA) (Appendix-2) . RNA with Ct values less than 31 were taken for cDNA preparation. The cDNAs were quantified using Quantus™ Fluorometer and Quantifluor® dsDNA system (Promega, US). Illumina's Respiratory Virus Oligo Panel was used in combination with Illumina DNA Prep with Enrichment and IDT for Illumina Nextera UD Indexes to construct target enriched libraries from ~150 ng of input DNA. Approximately 22.5 pmole of pooled and enriched libraries were loaded onto a MiSeq v2 kit (300 cycles) to produce paired-end reads from 152 cycles at NSU Genome Research Institute (NGRI), North South University. The sequenced reads were assembled and consensus fasta files were generated by mapping onto SARS-CoV-2 reference strain (GenBank Accession no. MN908947.3) using Illumina® DRAGEN RNA Pathogen Detection App. Sequence data have been submitted in the GISAID and metadata are available in Supplementary Table S1.

Identification of unique and rare mutations and prediction of functional effects

The mutations in each protein in each strain were obtained using the CoV server mutations analysis tool from GISAID (https://www.gisaid.org/epiflu-applications/covsurver-mutations-app/) which runs the input

FASTA sequence against the reference strain from Wuhan 2019 (hCoV-19/Wuhan/WIV04/2019). The frequency of occurrence of each mutation and its global distribution was also obtained using this software. The statistics of the global distribution of the mutations are accurate as of January 2021.

The effect of the unique mutation identified was predicted using PROVEAN (v1.1)[9] and PolyPhen-2 (http://genetics.bwh.harvard.edu/pph2/index.shtml)[10]. In both cases, the FASTA file of the reference Spike protein (https://www.ncbi.nlm.nih.gov/protein/QHD43416.1, accession MN908947.3, with 1273 amino acids) was used as the input file, and the substitution of Y for F at position 204 was specified. The PROVEAN threshold was set to the default -2.5. If the score is higher than the threshold value, it indicates a neutral effect, whereas if the score is equal to or lower than the threshold value, it indicates a deleterious effect[9]. For PolyPhen-2, the closer the score is to 1, the more deleterious the mutation, and the closer it is to 0, the more neutral it is[11], and the specificity and sensitivity scores refer to the prediction confidence[10]. PolyPhen-2 predicts the outcome for 2 models- The human divergence model (HumDiv) is the preferred model for analyzing rare alleles and natural selection while the human variance (HumVar) model is the preferred model for identifying Mendelian disease including deleterious alleles[11].

Lineage distribution and phylogenetic analysis

In addition to the genomes sequenced in this study, the lineages of 620 Bangladeshi and 68,870 global SARS-CoV-2 sequences that had been deposited in GISAID (Global Initiative on Sharing All Influenza Data) as of December 30 2020, were analyzed. From GISAID, 620 complete genomes of SARS-CoV-2 isolated from Bangladesh till 30th November, 2020 (as of 20th January), were downloaded. Only complete genomes with a size of greater than 29,000 bp were selected and those with low coverage i.e. possessing > 5% of N's, were filtered out. The Wuhan-1 strain (MN908947.3) was used as the reference[12]. The genomes of those 614 strains along with the reference and the strains of this study were aligned with MAFFT (Multiple Alignment using Fast Fourier Transform)[13], following which positions with mutations were extracted using SNP-sites[14]. A maximum-likelihood tree was generated using IQTree and the ultrafast bootstrap method of 1000 replicates[15]. The tree was visualized and edited using iTOL (Interactive Tree of Life)[16].

Phylodynamic analysis

The dataset for the phylodynamic inference of Bangladeshi genomes contains a total of 564 low coverage excluded, high coverage, and completed genomes from Bangladesh. The genomes were obtained from GISAID on January 23, 2021, and all of the genomes spanning from April 08, 2020, to December 19, 2020, were obtained.

A reduced dataset was designed to analyze the Noakhali strains, which included 207 Bangladeshi strains of the total Bangladeshi strains, and 56 SARS-CoV-2 genomes (from this study) collected from Noakhali, sequenced at the NGRI, both within the above-mentioned timeline. Two-thirds of the total Bangladeshi strains were discarded and only 20% of the strains were selected at random from the major clusters

observed in the initial phylodynamic tree. Furthermore, the sequences were selected to ensure a fair representation of the epidemiologically identified clusters and to mitigate sampling bias.

The phylodynamic analysis of all the genomes from Bangladesh was analyzed using BEAST (Bayesian Evolutionary Analysis Sampling Trees) version 1.10.4[17]. BEAST is focused explicitly on reconstructing time-scaled trees using strict or relaxed molecular clock models, and this cross-platform tool uses Markov Chain Monte Carlo (MCMC) algorithm to average over tree space so that each tree is weighted proportional to its posterior probability.

The datasets were imported in BEAUTi, where the dates were parsed, and a Hasegawa-Kishino-Yano (HKY) + I^- nucleotide substitution model with a gamma category count of 4 and a strict molecular clock was selected as the nucleotide substitution model. The tree prior parameter was set to a coalescent exponential growth model, which was more suitable for the analysis of early viral samples as the epidemic growth is assumed to be exponential early-on. The default option of a “Random starting tree” was selected to start the inference process. The Bayesian analysis for all data containing all the Bangladeshi strains was run for 10 million MCMC steps with sampling parameters and trees every 1000 generations, and the Noakhali strains analysis was run for 30 million MCMC steps with sampling parameters and trees every 3000 generations. The effective sample sizes (ESS), 95% highest posterior density intervals for evolutionary rate, were inspected using Tracer (v1.7.1)[18], and the tree file was summarized in TreeAnnotator by setting the burnin percentage to ten and the target tree type to maximum clade credibility tree. The MCC tree based on MCMC analysis of the 564 Bangladeshi genomes was visualized in FigTree v1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/).

Results

Unique and rare mutations in Noakhali strains

The surface spike glycoprotein recognizes the host cell receptors and mediates the virion’s binding and fusion with the host cell membrane, thus facilitating the virus’s entry into the cells. Furthermore, this protein increases the adhesion of infected cells with neighbouring non-infected cells to allow the spread of the virus across cells[19]. This protein has been the target of much investigation. Of the strains sequenced and analyzed in Bangladesh, the majority of the strains belong to the GR clade, a derivative of the G clade. 274 strains of the 324 sequenced belonged to the GR clade (GISAID). The G clade is the parent clade of GR and GH and is classified by the mutation D614G[20]. This mutation was first detected in Europe in February 2020, with an especially high incidence in Italy[21]. The D614G mutation is dominant in Bangladesh, reflecting the global trend and has been linked with a higher infectivity[22]. Each strain we sequenced exhibited the D614G mutation in the spike protein. However, 21 out of the 56 sequenced strains also showed other mutations in the critical spike protein, which may have consequences for the protein’s structure and function (Fig. 1).

Table 1: Unique and rare mutations found in the spike protein of SARS-CoV-2 strains isolated from Noakhali, Bangladesh.

Protein	Strain Name	Mutation	No. of occurrences (global)	No of countries with this mutation	No. of BD strains with this mutation	Notes
Structural Proteins						
S	NGRI-NSTU-2	G769V	617	31	4	
NGRI-NSTU-21	Q675H	464	34	4		
NGRI-R2-4	H245Y	233	12	0		
NGRI-R2-O10	G261R	59	8	1		
NGRI-R2-8	G261R	59	8	1		
NGRI-R2-18	S46L	8	5	0		
NGRI-R3-2	K1191N	155	22	1	Non-structural (C-term)	
NGRI-R3-3	K1191N	155	22	1	Non-structural (C-term)	
NGRI-R3-12	A684S	6	5	1		
E654V	15	3	0			
NGRI-R3-14	C1247S	3	2	0	Non-structural (C-term)	
NGRI-R3-15	C1247S	3	2	0	Non-structural (C-term)	
NGRI-R3-16	Y204F	No data	No data	No data		
NGRI-R3-O1	D80Y	1447	25	0		
NGRI-R3-O5	C1247F	124	14	0	Non-structural (C-term)	
NGRI-R3-O15	S1252F	203	13	0	Non-structural (C-term)	
NGRI-R4-OR18	Y204F	No data	No data	No data		
NGRI-R4-OR19	L54F	965	34	1		
NGRI-R4-O17	R21K	106	7	0		
Q1201H	24	6	0	Non-structural		

N	NGRI-NSTU-2	G18C	66	10	1
NGRI-NSTU-3	P67L	57	10	1	
NGRI-R2-22	G238C	261	23	0	
NGRI-R2-O9	R203S	71	10	0	
NGRI-R3-10	A211T	13	4	0	
NGRI-R3-11	A211T	13	4	0	
NGRI-R3-17	M210I	743	25	1	
E	NGRI-R2-O9	L19F	5	5	0
M	NGRI-R2-17	H125Y	537	26	7
NGRI-R4-O17	D209Y	476	20	1	
Non-structural proteins					
NGRI-R2-4:	NS3	N119Y	No data	No data	No data
NS7a	S81del, V82del	No data	No data	No data	
NGRI-R2-27	NS7a	E92stop	No data	No data	No data
NGRI-R3-17	NSP3	L1259R	No data	No data	No data

A total of 17 different mutations besides the D614G are found in our strains, with four of them appearing in more than one strain. Of these, the A262S mutation found in strain NGRI-R2-O9 has not been included as it is not an uncommon mutation, having appeared more than 3000 times across 29 countries. All other additional mutations were found to be globally rare. Two of these recurring four mutations, K1191N appearing in strains NGRI-R3-2 and NGRI-R3-3, and C1247S in NGRI-R3-14 and NGRI-R3-15 are non-structural mutations found in the C-terminal region of the spike protein. Furthermore, of the four recurring mutations, one particular mutation, the Y204F in strains NGRI-R4-OR18 and NGRI-R3-16, has not been recorded in any strains from any other countries in the GISAID database previously. Another, C1247S, in strains NGRI-R3-14 and NGRI-R3-15 have only been recorded 3 times from across 2 countries in the GISAID database, and never before in Bangladesh.

Some of the mutations that appear only once are also quite rare, including the A684S in strain NGRI-R3-12 which was previously seen 6 times in 5 countries, and the S46L mutation in strain NGRI-R2-18, which has only been recorded 8 times from across 5 countries. The mutation A262S in NGRI-R2-O9 was the

most common worldwide, with 3003 previously recorded cases from 29 countries. However, it has only been reported once before in Bangladesh.

Of the 17 mutations included in Table 1, 7 were reported in Bangladesh previously. The mutations found in the strains NGRI-NSTU-2 and NGRI-NSTU-21 were both reported 4 times, while the other 5 appeared previously only once each. For the remaining 10 mutations, our sequenced strains mark their first recorded appearance in Bangladesh.

The circular nucleocapsid proteins bind the 2981 nucleotide long viral genome to form a ribonucleotide core which helps in the virus's entry into the host cell. It is also important in viral RNA synthesis. Following entry, this protein also affects cellular processes that are responsible for inflammation in the lungs, progression of the cell cycle, immune responses and viral degradation in the host body[19]. The most common mutations found across our sequenced strains were the GR clade-defining RG203KR mutations and the S194L mutations, which are among the globally dominant mutations for this protein[20]. These mutations are also known to cause major changes in the protein structure and intraviral interactions[23]. However, there were also 6 different unusual mutations that appeared across our strains, which will be the focus of this report. Of these, 3 are being reported for the first time in Bangladesh.

The mutation A211T appeared twice, in the strains NGRI-R3-10 and NGRI-R3-11. While this mutation has been reported 13 times globally across 4 countries, it has so far not been reported in Bangladesh (GISAID). The other mutations that are being reported for the first time in Bangladesh include the G238C mutation (which has appeared 261 times in 23 countries) and the R203S mutation (which has appeared 71 times in 10 countries). Uncommon mutations which have appeared in Bangladesh previously and also appeared in our strains include the G18C (66 times in 10 countries previously), P67L mutation (57 times in 10 countries previously), and the M210I (807 times in 30 countries previously) mutations. These have all been recorded once in Bangladesh till now.

The envelop protein is a small membrane protein that plays a role in viral assembly and releases[19]. The envelope protein is the most conserved structural protein both globally[20], and across our strains (GISAID). The only mutation that appeared in this protein in our strains was the L19F mutation which was only seen 5 times in 5 countries. This appeared in the strain NGRI-R2-O9.

The membrane protein exists abundantly in the virus and makes up the membrane on which the spike proteins are attached. This protein also acts as a scaffold for viral assembly. This protein has important consequences in host cells, including enhancing viral pathogenesis in the cells and promoting apoptosis of the cells[19]. The most common mutation in the membrane protein globally was reported to be the T175M[20]. This did not appear in our strains. Only 2 mutations were found in the membrane protein across our sequenced strains, the H125Y in strain NGRI-R2-17, and the D209Y in NGRI-R4-O17. Both of these have appeared in Bangladesh before, the former being recorded 7 times across the country and the latter being recorded once in Dhaka. Globally, the H125Y mutation has been reported 580 times in 27 countries previously, and the D209Y has been reported 494 times in 22 countries previously (GISAID).

Besides the mutations in the structural proteins, 5 unique mutations were also found in the non-structural proteins, 3 of which were on the same strain (NGRI-R2-4). 3 of our strains carried these mutations. The mutations are listed in Table 1.

Functional implication of unique mutation Y204F:

Using PROVEAN, we obtained a score of -1.297 with 693 supporting sequences for prediction (Supplementary Data S1 and Supplementary Table S2). As this is above the threshold value, it indicates that the mutation is neutral – it will not have any significant effects on the protein function.

However, using PolyPhen-2, quite different results were obtained. Using the human divergence model HumDiv, we obtained a score of 0.995 (sensitivity 0.68, specificity 0.96) and using the human variation model HumVar, we obtained a score of 0.993 (sensitivity 0.47, specificity 0.96). As both these scores are close to 1 (Fig. 1), they indicate that the mutation is probably damaging [23].

Noakhali shows high prevalence of B.1.36 and B.1.36.16 lineages

Phylogenetic tree and lineage analysis enabled the pattern of lineage distribution to be observed for the district of Noakhali for the period of July to November 2020 (Figure 2a). The majority of the strains belonged to the GR clade parent lineage B.1.1 (n=21; 37.5%) and its descendant lineages which together contributed towards ~60.7% (n = 34) of Noakhali's SARS-CoV-2 lineages (Supplementary Table S3). The rest were of B.1.36.16 (n=12; 21.4%) and its parent lineage B.1.36 (n=9, 16.0%), both of which belong to the GH clade. All three dominant lineages, i.e., B.1.1, B.1.36, and B.1.36.16 were found across the sampling period and therefore, it may be assumed that no shift in prevalence of various dominant lineages was observed during the studied timeframe.

Notably, two of the strains from the study, NGRI-R4-2 and NGRI-R4-3 were found to be of lineages B.1.141 and B.1.186 both of which haven't been reported from Bangladesh thus far. While globally B.1.1.141 is predominantly found in the UK, it is also dispersed in other European countries, Russia and Brazil. B.1.186 on the other hand originated in mid-March, 2020, in Saudi Arabia where it is the most prevalent (57.3% of total cases). The lineage also shows some representation in Italy (11.0%) and India (8.5%). Both of these lineages (B.1.186 and B.1.1.141) were isolated on the first week of November 2020 and therefore were possibly introduced at the beginning of the second wave of COVID-19 pandemic in the country.

This study has made >50 strains of SARS-CoV-2 from the Noakhali district available, enabling lineage analysis of Bangladeshi SARS-CoV-2 strains at a smaller geographic scale. Two lineages of the GH clade, B.1.36 and B.1.36.16 were particularly abundant in the Chittagong division when compared to the rest of the country as shown in Figure 2b. Chittagong Division alone contributed to 57.1% of Bangladesh's B.1.36.16 lineage distribution, a percentage that rose to 72.7% after inclusion of Noakhali's strains of our study into the dataset. Similarly, while prior to the inclusion of our strains, 46.2% of the B.1.36 lineage strains were shown to have originated from Chittagong Division, the updated dataset propelled this statistic to 68.2%. While B.1.1 was relatively rare in Chittagong Division for the observed time period,

contributing to only 10.8% of the national aggregate, it was noticeably prolific among Noakhali's strains as it accounted for 40.7% (n = 22) of lineages from the district.

Correlation of disease severity with lineage and mutations

Metadata of patients infected by Noakhali's strains revealed that most of the prolific lineages from the district, such as B.1.1, B.1.36 and B.1.36.16 exerted varying degrees of severity in their hosts, ranging from asymptomatic cases to severe infections (Supplementary Table S4 and Supplementary Table S5). However, no death cases were found to bear strains of B.1.36 or its descendant lineages. Only one death case was found among the 21 B1.1 lineage infected patients, while one strain belonging to B.1.1.10 lineage also resulted in death. Notably, however, two strains belonging to the B.1.1.25 lineage caused patient deaths while the third produced severe symptoms in its host.

Of the two patients carrying the unique spike mutation Y204F, one of them (NGRI-R3-16) was asymptomatic, while the other (NGRI-R4-OR18) showed moderate symptoms. Therefore, it is unlikely that this mutation affects disease severity. Of all the rare mutations identified, three patients carrying them were deceased. They were the hosts of the strains NGRI-NSTU-3 containing the nucleocapsid protein mutation P67L, NGRI-R2-17 containing the membrane protein mutation H125Y, and NGRI-R2-8 containing the spike protein mutation G261R.

Furthermore, some of the patients carrying the listed rare mutations in the structural proteins also had severe symptoms. Patients infected by strains exhibiting rare mutations in the spike protein who displayed severe symptoms included those carrying strains NGRI-R2-O10 containing the mutation G261R, NGRI-R3-O15 containing the mutation S1252F, NGRI-R2-18 containing the mutation S46L, NGRI-R3-2 containing the mutation K1191N, and NGRI-R2-4 containing the mutation H245Y. It is interesting to note that the strain NGRI-R2-4 also had 3 unique mutations in non-structural proteins. Of the patients carrying strains with a rare mutation in the membrane protein, only one displayed severe symptom, and this was caused by the strain NGRI-R4-O17 carrying the mutation D209Y.

Finally, the strains with rare mutations in the nucleocapsid protein that caused severe symptoms include strains NGRI-R2-22 containing the mutation G238C, NGRI-R3-17 containing the mutation M210I (this strain also had a unique mutation in a non-structural protein), and NGRI-R3-11 containing the mutation A211T. It is interesting to note that the other patient carrying this A211T mutation (in strain NGRI-R3-10) showed only moderate symptoms. The patient carrying the only strain containing a mutation in the envelope protein, strain NGRI-R2-O9, was asymptomatic.

Phylodynamic inferences from Bangladeshi strains

Analysis of 546 genomes from Bangladesh estimates the evolutionary rate and the date of the most recent common ancestor (TMRCA) of the genomes used in the analysis. The evolutionary rate, measured in substitutions per site per year (subs/site/year), tells us the rate at which the virus evolves to bring a

change in their existing lineage. The MRCA is the point where all the sampled viruses were in the same host, whether human or non-human and so its timing can represent when the epidemic began to diverge.

The analysis from the Coalescent Exponential Growth model shows a clock rate of 1.66 X 10^{-4} subs/site/year (95% BCI: 7.059 X 10^{-5} – 3.7889 X 10^{-4}) estimated to be the average evolutionary rate of Bangladeshi genomes (Table 2). Simultaneously, the global estimated rate of SARS-CoV is between 0.80 – 2.38[24]. With appropriate time-scale settings, the node age of the root is 2020.1538 in decimal years, which when shows the most recent common ancestor (TMRCA) to first appear around February 25, 2020 (95% BCI: October 11, 2019 – October 10, 2020), in Bangladesh, which is consistent with the other global estimates. Moreover, the first positive SARS-CoV-2 was detected around March 8, 2020, within two weeks of the predicted introduction. Figure 3 and Figure 4 shows the maximum clade credibility (MCC) tree representing the distribution of 564 genomes from Bangladesh.

Phylodynamic inferences from Noakhali strains

Analysis of the Noakhali genomes along with the strains from all over Bangladesh shows the average evolutionary rate to 1.065 X 10^{-5} subs/site/year (95% BCI: 8.26 X10^{-9} - 8.42 X 10^{-5}) and the most recent common ancestor (TMRCA) were February 11, 2020 (Table 2). The time-scaled phylodynamic tree of this analysis is shown in Figure 5 and Figure 6.

Table 2: Summarization of the evolutionary rates (in subs/site/year) and estimated MRCA

Analysis	Clock Model	Coalescent Model	Estimated Rate (Mean Rate)	Substitution Rate (95% HPD)	Estimate MRCA	95% HPD Interval
Bangladesh	Strict Clock	Exponential Growth	1.66E-04	7.059 X 10-5 – 3.7889 X 10-4	February 25, 2020	October 11, 2019 – October 10, 2020
Noakhali	Strict Clock	Exponential Growth	1.07E-05	8.2648 X 10-9 - 8.4286 X 10-5	February 11, 2020	-

Discussion

The first positive case of COVID-19 was recorded on March 8, 2020, as a result of the SARS-CoV-2 pandemic with its first outbreak in Hubei, Wuhan, China around late December of 2019. The data collected span from early April 2020 till late December 2020 which is within two months of the first outbreak thus, holds information about the early stages of the SARS-CoV-2 epidemic, and the evolutionary changes that accumulated in the viral genome[25,26].

The SARS-CoV-2 virus shares the important structural proteins with the rest of the coronavirus family. These proteins are the spike (S), nucleocapsid (N), envelope (E), and membrane (M) proteins. They are critical in the virus's functions, including its attachment to host cell receptors[19]. Previous studies have

shown Bangladesh to have a slightly higher mutation rate than the global average[20]. This report summarizes the rare mutations seen in these structural proteins across the Bangladeshi strains sequenced by our lab. In this report, we define rare as less than 1000 globally.

The unique mutation we found involves the substation of one aromatic amino acid, tyrosine (Y), to another aromatic amino acid, phenylalanine (F), at position 204 of the spike sequence. A previous study found PROVEAN to be one of the best predictors of pathogenicity while ranking PolyPhen-2 as average[7]. Furthermore, Polyphen-2 is designed to predict the effects of mutations in human proteins[10], and therefore its prediction with regards to the SARS-CoV-2 spike protein may not be reliable. While these are only *in silico* analyses and further structure-function studies are needed to reach a conclusion about the effects of the Y204F mutation in the spike protein, it is likely that the mutation may not have any severe effects on the function of the protein. This possibility is further backed by the fact that both tyrosine and phenylalanine are aromatic amino acids with similar structures.

While we were able to observe certain mutations present in patients exhibiting varying disease severity, further analysis using a larger sample size is required to establish a causative correlation between the mutations and disease severity. In the future, it would be interesting to further study the correlation between the mutations reported in these strains with phenotypes such as disease severity and mortality rates of infected patients using statistical analyses. Structure-function studies based on these mutations could help explain any correlations found. Drug design targeting specific mutations would also be an interesting avenue of investigation.

The study's focus on Bangladesh's south-eastern district of Noakhali aided in further confirming the higher prevalence of B.1.36 and descendant lineages in the southernmost regions of the country. While no clear correlation was observed between lineage and disease severity, higher frequencies of certain lineages may be explained by possible local outbreaks of these lineages owing to their greater adaptability in the region. The lineage B.1.36 can be considered as one of the most virulent lineages and the observation is based on the genome sequences available in GISAID for deceased patients from India and Saudi Arabia, respectively (data not shown). Among the 52 Indian genomes available (deceased) at GISAID, 10 were infected by this lineage, whereas, in Saudi Arabia among the 121 available genomes, almost half of them were affected by this lineage (January 18, 2021).

The phylodynamic analysis using Bayesian inferences enabled us to estimate the average rate of evolution of the Bangladeshi genomes and estimate the most recent common ancestor. The evolutionary rate was found to be 1.66×10^{-4} substitution per site per year with the MRCA around late February 2020. Taking the epidemiological information into account, along with the results from the analysis, the estimated date of TMRCA have appeared to be approximately within two weeks of the first case. An analysis of the Noakhali strains was combined and analyzed to obtain an estimate of the evolutionary rate and TMRCA. The virus was estimated to be evolving at the rate of 1.065×10^{-5} subs/site/year with TMRCA around mid-February.

The dataset for the Noakhali analysis spanned from April 2020 to December 2020 during which Bangladesh has experienced its first wave which began early-April and peaked around mid-June. During that time frame, there was a spike in mid-November which declined towards the end of December. As a result of the presence of outliers, the average rate of evolution was a bit lower than expected. If the dataset was partitioned and analyzed for shorter time frames, it would have provided a much better estimate. Adding to it, the use of other coalescent models like the Birth-death model could have provided an estimate for the basic reproductive number which would allow us to have a better scenario of the phylodynamics and transmission pattern. Given that the Chittagong division came in second to the Dhaka division consistently, in terms of the increase in the number of active cases, and deaths, application of the SEIR model could have opened up a lot more information that could enable to assess the situation in more depth[27].

Administration of recombinant spike proteins has shown reduced immunity in many subjects and responses varied between individuals although there has been no report on the loss of neutralization activity. Moreover, there are some reports on the more severe cases of infection after vaccination due to SARS-CoV-2 although such outcome depends on number of factors including since how many days preceding the person was vaccinated, administration of booster dose, adaptive evolution through new mutation acquisition etc. and such outcome should be taken into serious consideration. Whole genome sequencing should therefore be in place for continuous surveillance for identification of the changing nature of this deadly pathogen. Sequencing information will be useful to determine the genomics changes and their effect on immunogenicity of the viruses and develop effective therapeutics and help policy makers to take public health measures accordingly.

Conclusion

Most of the studies represented a small fraction of the documented number of positive cases at the time in Bangladesh as they were selected to be a country-wide representative given limited resources for genome sequencing and consequently, the introduced viral diversity may also have been underestimated[8]. The number of genomes sequenced in the study are representatives form particular geographic locations of Bangladesh and data from considerable positive cases (15% of cases; 56 genomes from 365 positive cases) were available. Recent reports from ICDDR, B on the evidence of variants of interest emerging in the UK and South Africa have heightened calls for systematic genomic surveillance in Bangladesh. Our results suggest patterns of SARS-CoV-2 transmission may vary substantially even in nearby communities. Understanding these local patterns will enable better targeting of public health interventions and continued surveillance to sequence SARS-CoV-2 viruses across multiple spatiotemporal scales remain critical for tracking viral transmission dynamics within and between communities.

Declarations

Contributions

M.H.: fund acquisition, conceptualization, supervision, computational analysis, writing, reviewing, editing; T.S.H.: laboratory experiments , computational and data analysis, writing, editing; A.R.: laboratory experiments, data analysis, writing, editing: A. I.: sample collection, laboratory experiments, data analysis; S.N.T: bioinformatics data analysis, writing, editing; A.S.: laboratory experiments; K.N.H., A.K., M.S.SH., and N.M.BAr.H.: supervision and reviewing; F.A.: sample collection, supervision and reviewing; H.M.R.: conceptualization, supervision, and reviewing. All authors read and reviewed the final manuscript.

Corresponding authors

Correspondence to Hasan Mahmud Reza and Maqsud Hossain

Competing interests

The authors declare no competing interests.

Ethical statements

The study was reviewed and approved by the ethics committee of the Directorate General of Health Sciences (DGHS), Bangladesh and by National Research Ethics Committee (NREC) (Ref.: BMRCAIREC/2O 1 9 -2022 I 7 08) of Bangladesh Medical Research Council (BMRC).

Noakhali Science and Technology University (NSTU) is a Bangladesh Government approved national Covid-19 testing center. Relevant demographic, clinical and laboratory data were retrieved from the clinical records of the patient according to the relevant guidelines and regulations. Signed written informed consent was obtained from the patient.

Competing interests

The authors declare no competing interests

Acknowledgement

We gratefully acknowledge the Board of Trustees of North South University for funding this project.

References

1. van Dorp, L. et al. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infect. Genet. Evol. 83, 104351 (2020). 2. Ahmed, S. F., Quadeer, A. A. & McKay, M. R. Preliminary identification of potential vaccine targets for the COVID-19 Coronavirus (SARS-CoV-2) Based on SARS-CoV Immunological Studies. Viruses 12, (2020). 3. Phan, T. Genetic diversity and evolution of SARS-CoV-2. Infect. Genet. Evol. 81, 104260 (2020). 4. Walls, A. C. et al. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. Cell 181, 281-292.e6 (2020). 5. Cotten, M. et al. Transmission and evolution of the Middle East respiratory syndrome coronavirus in Saudi Arabia: A descriptive genomic study. Lancet 382, 1993–2002 (2013). 6. Quick, J. et al. Real-time, portable genome sequencing for Ebola surveillance. Nature 530,

228–232 (2016). 7. Pshennikova, V. G. et al. Comparison of Predictive In Silico Tools on Missense Variants in GJB2, GJB6, and GJB3 Genes Associated with Autosomal Recessive Deafness 1A (DFNB1A). Sci. World J. 2019, 5198931 (2019). 8. Cowley, L. A. et al. Genomic and mobility data reveal mass population movement as a driver of SARS-CoV-2 dissemination and diversity in Bangladesh. medRxiv 2021.01.05.21249196 (2021). 9. Choi, Y., Sims, G. E., Murphy, S., Miller, J. R. & Chan, A. P. Predicting the Functional Effect of Amino Acid Substitutions and Indels. PLoS One 7, (2012). 10. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. Current Protocols in Human Genetics (2013). doi:10.1002/0471142905.hg0720s76. 11. Canhui Cao et al. Amino acid variation analysis of surface spike glycoprotein at 614 in SARS-CoV-2 strains. Genes Dis. 7, 567–577 (2020). 12. Wu, F. et al. A new coronavirus associated with human respiratory disease in China. Nature 579, 265–269 (2020). 13. Katoh, K., Kuma, K. I., Toh, H. & Miyata, T. MAFFT version 5: Improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. 33, 511–518 (2005). 14. Page, A. J. et al. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. Microb. genomics 2, (2016). 15. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the ultrafast bootstrap approximation. Mol. Biol. Evol. 35, 518–522 (2017). 16. Letunic, I. & Bork, P. Interactive Tree of Life (iTOL) v4: Recent updates and new developments. Nucleic Acids Res. 47, W256–W259 (2019). 17. Suchard, M. A. et al. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol. 4, (2018). 18. Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. Syst. Biol. 67, 901–904 (2018). 19. Satarker, S. & Nampoothiri, M. Structural Proteins in Severe Acute Respiratory Syndrome Coronavirus-2. Arch. Med. Res. 51, 482–491 (2020). 20. Mercatelli, D. & Giorgi, F. M. Geographic and Genomic Distribution of SARS-CoV-2 Mutations. Front. Microbiol. 11, 1–13 (2020). 21. Isabel, S. et al. Evolutionary and structural analyses of SARS-CoV-2 D614G spike protein mutation now documented worldwide. Sci. Rep. 10, 1–9 (2020). 22. Yurkovetskiy, L. et al. SARS-CoV-2 Spike protein variant D614G increases infectivity and retains sensitivity to antibodies that target the receptor binding domain. bioRxiv Prepr. Serv. Biol. 13, 14 (2020). 23. Wu, S. et al. Effects of SARS-CoV-2 mutations on protein structures and intraviral protein–protein interactions. J. Med. Virol. 1–9 (2020) doi:10.1002/jmv.26597. 24. Zhao, Z. et al. Moderate mutation rate in the SARS coronavirus genome and its implications. BMC Evol. Biol. 4, 1–9 (2004). 25. Paul, R. Bangladesh confirms its first three cases of coronavirus. Reuters https://www.reuters.com/article/us-health-coronavirus-bangladesh-idUSKBN20V0FS (2020). 26. COVID-19 Situation Updates. IEDCR https://iedcr.gov.bd/covid-19/covid-19-situation-updates. 27. WHO. Bangladesh COVID-19 Morbidity and Mortality Weekly Update (MMWU). vol. 57 (2021).

Figures

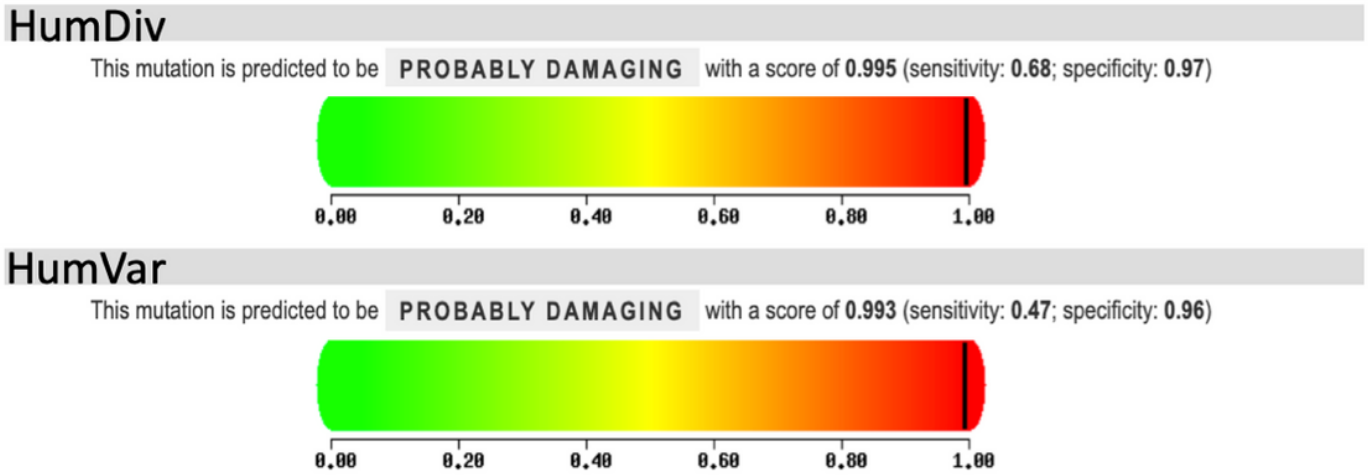

Figure 1

The HumDiv and HumVar scores obtained from PolyPhen-2. The scores are 0.995 and 0.993 respectively. Both values are close to 1 and predict that the mutation has a damaging effect on the spike protein

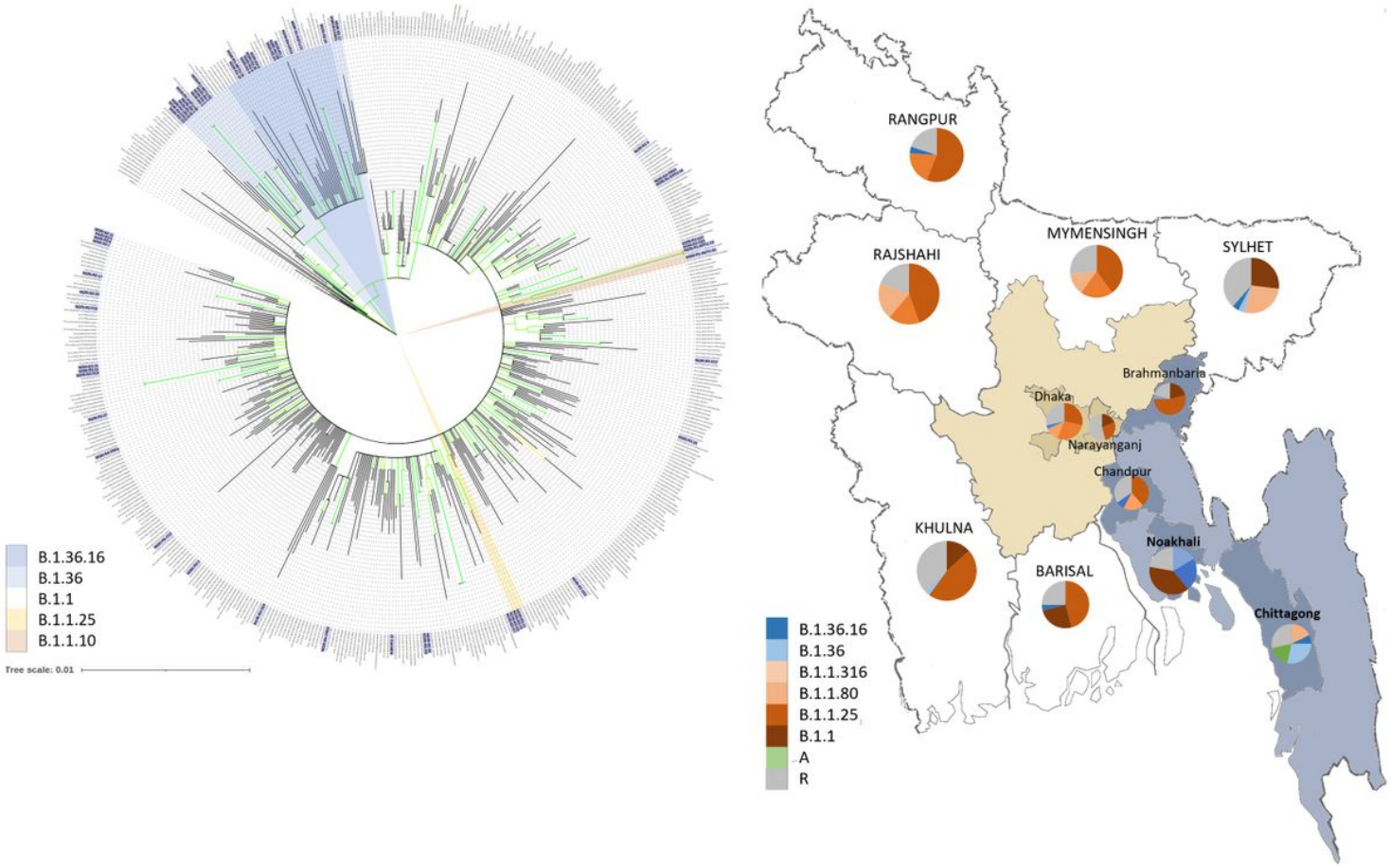

Figure 2

a) Maximum-likelihood tree of all Bangladeshi SARS-CoV-2 strains isolated till 30th November, 2020. The Wuhan reference strain (MN908947.3) has been used as root. Strain labels with purple background show

positions of the strains of this study, while high bootstrap branches are represented in green. The tree was visualized and edited using iTOL v516. b) Lineage distribution pattern in different divisions and districts of Bangladesh. The prevalence of GH clade lineages B.1.36 and B.1.36.16 appears higher in the districts of Noakhali and Chittagong in comparison to other divisions. The lineages under clades G and GR are represented in warm colors and those under clade GH are represented in cool colors. Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

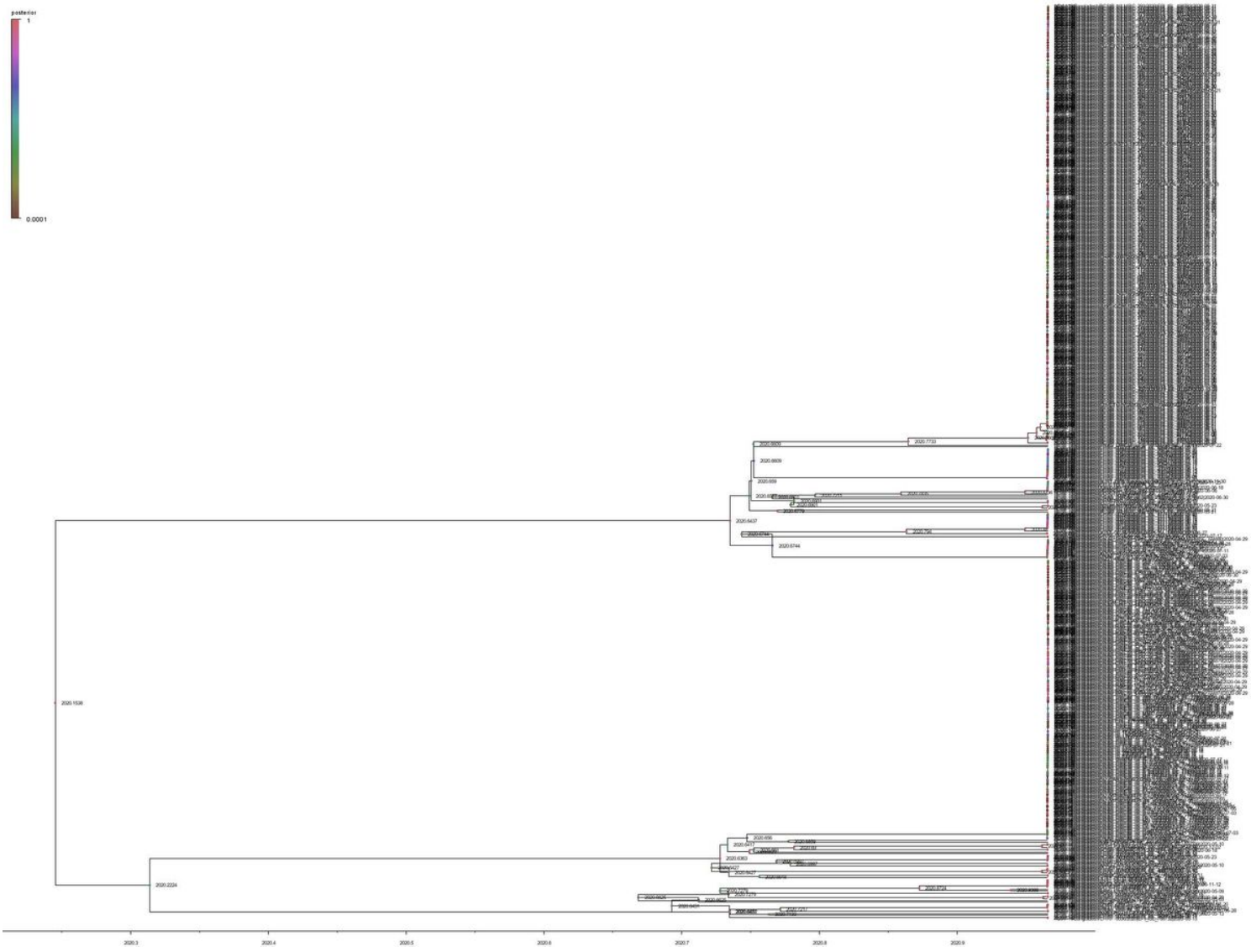

Figure 3

Estimated Maximum-Clade-Credibility Tree of 564 complete sequences of SARS-CoV-2 from Bangladesh. The scale below the tree dates backward from the present to the root where the most recently sampled strain at time zero and the root is somewhere in the past (in years). Each clade has been labeled with node ages in decimal years, indicating an estimate of divergence.

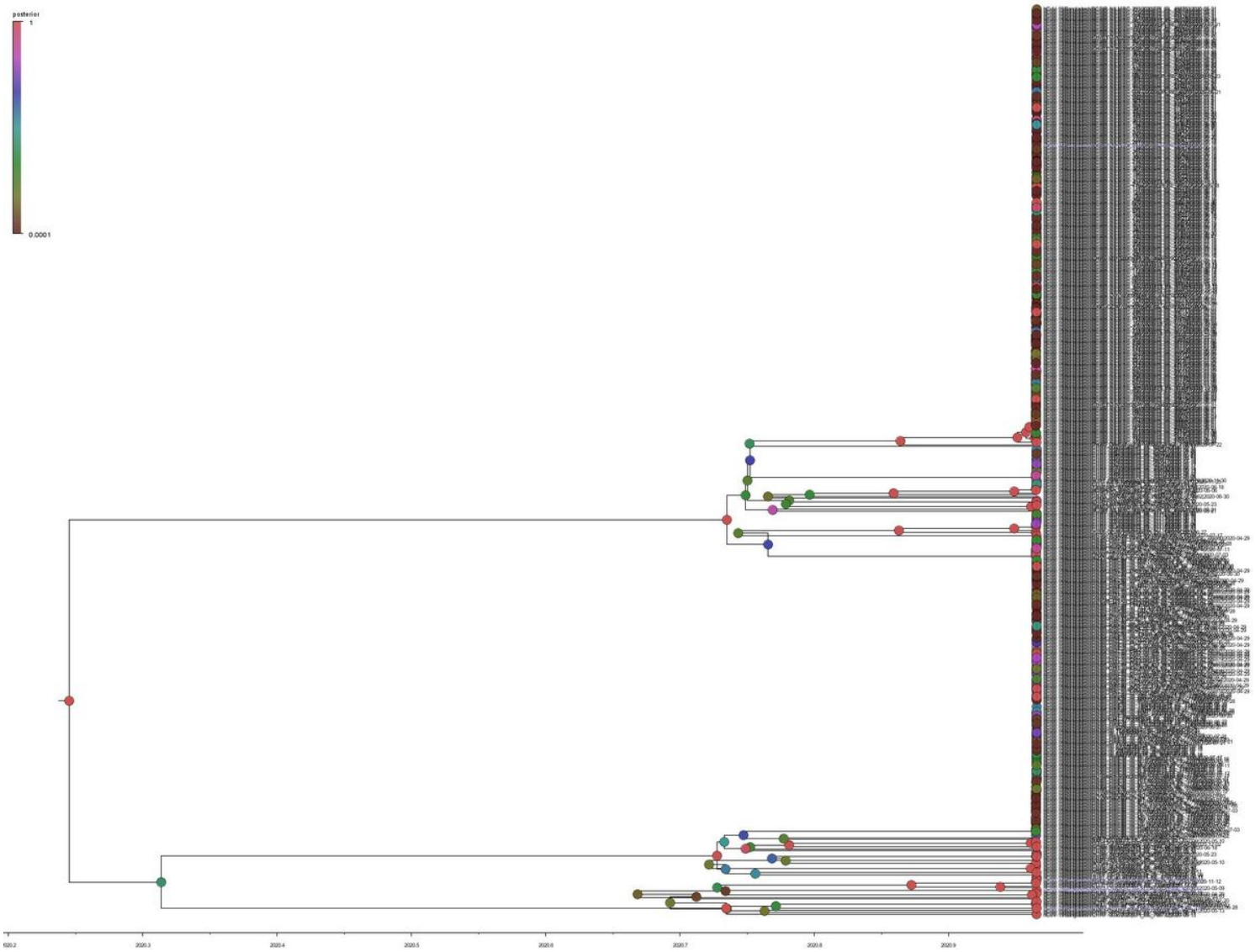

Figure 4

Modified time-scaled tree representing Bangladeshi genomes and their posterior (divergence) of the strains from the root. The legend on the top left corner ranges from 1 to 0.001, where the root is considered to be 1. The colors on the legend allows us to find a pattern in divergence of the strains from the root, circulating in Bangladesh. As the tree is in increasing order, by looking at the colors at nodes, strains at the bottom of the tree to have diverged less in comparison to the strains at the top.

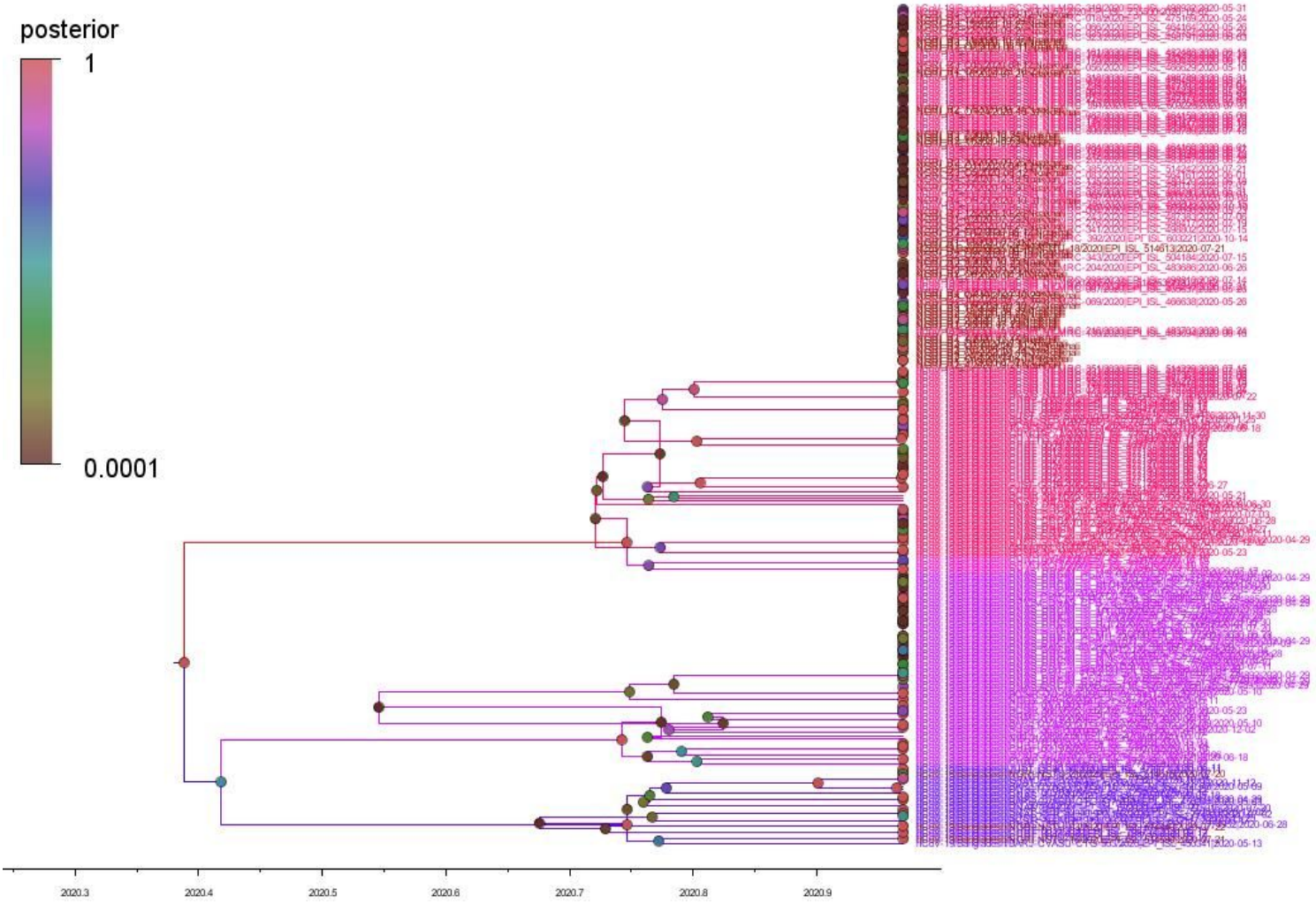

Figure 5

The time-scaled tree generated for the phylodynamic analysis of Noakhali strains. The rooted tree contains 207 genomes from all over Bangladesh and 58 genomes from the Noakhali district, sequenced at NSU, Genome Research Institute. The colored nodes of the tree represent posterior (divergence) values and allows us to see the divergence between the strains. The analysis reveals the strains have grouped into four clusters and tips of the tree have been colored to distinguish the genomes in each of the four clusters.

Figure 6

The time-scaled tree of the 265 genomes analyzed using BEASTv1.10.4. The tree estimates the evolutionary rate to be 1.065 X 10-4 subs/site/year with TMRCA to be around mid-February 2020 (2020.1158 in decimal years). The taxa highlighted in pink above are the genomes from Noakhali, sequenced at NSU, Genome Research Institute, to express the distribution across the time-scaled rooted tree generated.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryDataS1.fasta
- SupplementaryTableS1.xlsx
- SupplementaryTableS2.xlsx
- SupplementaryTableS3.xlsx
- SupplementaryTableS4.xlsx
- SupplementaryTableS5.xlsx
- SupplementaryDataS1.fasta
- SupplementaryTableS1.xlsx
- SupplementaryTableS2.xlsx
- SupplementaryTableS3.xlsx
- SupplementaryTableS4.xlsx
- SupplementaryTableS5.xlsx