Comparative Genomic Study for Revealing the Complete Scenario of COVID-19 Pandemic in Bangladesh

Ishtiaque Ahammad[1], Mohammad Uzzal Hossain[1], Arittra Bhattacharjee[1,2], Zeshan Mahmud Chowdhury[2], Md. Tabassum Hossain Emon[3], Md. Golam Mosaib[4], Keshob Chandra Das[5], Chaman Ara Keya[2], Md. Salimullah[5*]

[1]*Bioinformatics Division,* [5]*Molecular Biotechnology Division, National Institute of Biotechnology, Ganakbari, Ashulia, Savar, Dhaka-1349, Bangladesh*

[2]*Department of Biochemistry and Microbiology, North South University, Bashundhara, Dhaka-1229, Bangladesh*

[3]*Department of Biotechnology and Genetic Engineering, Life Science Faculty, Mawlana Bhashani Science and Technology University, Santosh, Tangail-1902, Bangladesh.*

[4]*Department of Biochemistry and Molecular Biology, Faculty of Health & Medical Sciences, Gono Bishwabidyaloy, Ashulia, Savar, Dhaka-1344, Bangladesh*

***Correspondence:**

Dr. Md. Salimullah

Chief Scientific Officer

Molecular Biotechnology Division

National Institute of Biotechnology

Ganakbari, Ashulia, Savar, Dhaka-1349, Bangladesh

Tel: 880-2-7788443

E-mail: salim2969@gmail.com

Abstract

As the COVID-19 pandemic continues to ravage across the globe and take millions of lives, worldwide efforts to understand its causative agent, SARS-CoV-2 at the genomic level are also running in full swing. Such studies are providing precious insights about the pathogenesis, evolution, strengths and weaknesses of the virus. As of October 1st, 2020, 323 SARS-CoV-2 genomes have been sequenced across Bangladesh. The current study is aimed at answering some vital questions about these sequences. From our analyses, it was discovered that the majority of the SARS-CoV-2 found in Bangladesh belonged to the lineage B 1.1.25 of GR clade. Dhaka and Chittagong division were the most diverse in terms of SARS-CoV-2 clades while Mymensing was the least. There are more variety of clades in southern parts of Bangladesh than the northern parts. The most commonly found SARS-CoV-2 mutations found in the country were Spike_D614G, NSP12_P323L, N_G204R and N_R203K. Even though no significant pattern of distribution could be drawn between mutations found in Bangladesh and the countries with similar mortality rates and the countries with large Bangladeshi diaspora, to a certain degree they match with those in the UK, Oman, Italy, Greece, South Africa and Russia. Therefore, careful eye should be kept on the performance of vaccines in those countries in the near future as they are likely to work well in Bangladesh if they work well there. Mutational events in Bangladesh were found to increase between April and July, 2020 and decrease since August, 2020. The number of mutations per SARS-CoV-2 virus sample in Bangladesh was calculated to be 6.88 which is lower than the global average of 7.23. The decrease and the lower rate of mutation raise the possibility of a vaccine or drug working sustainably to protect the people. Based on these insights, a clear picture about the ongoing pandemic can be drawn in the context of Bangladesh which will help the country take appropriate measures to combat the virus.

Keywords

SARS-CoV-2; COVID-19; Genome Sequencing; Comparative Genomics; Bangladesh

1. Introduction

Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2), the causative agent of Coronavirus Disease-2019 (COVID-19), has already infected > 44,000,000 people with >1,168,000 deaths till October, 2020 (https://www.worldometers.info/coronavirus/). COVID-19 pandemic has passed through the first wave in many countries. [1]–[5] Insights regarding the transmission and evolution of the virus during this first wave are essential to break the subsequent chain of infections. [6], [7] Genomic data can provide some of these crucial insights which can help make upgraded public health policies.[8], [9] Besides, genomic surveillance can deliver deep understandings about the virus and reduce fatality during any new wave of infection.[10]–[13]

The onset of SARS-CoV-2 occurred in Wuhan, Hubei Province, China in December, 2019 [14]–[16]. Initially, clinicians diagnosed this disease as virus-induced pneumonia based on blood tests and chest radiographs. Later, the analysis of genomic data and phylogenetic tree led to the recognition of the pathogen as a member of the *Coronaviridae* family. [17] *Coronaviridae* family encompasses the largest known enveloped, single stranded RNA viruses with genome size ranging from 25-32 kilo base pairs (Kb) [18], [19]. The family is divided into two subfamilies, the *Coronavirinae* and the *Toronavirinae*. The subfamily *Coronavirinae* is further organized genotypically and serologically into 4 genera: α, β, γ, and δ-CoVs. [20] The *betacoronavirus* genus comprises of the Severe Acute Respiratory Syndrome (SARS)-CoV which was first identified in 2002-2003 and Middle East Respiratory Syndrome (MERS)-CoV in 2012. The genome sequences of SARS-Cov-2 has a 79.6% identity with SARS-CoV/ SARS-CoV-1 and 67.06% identity with MERS-CoV, indicating that they belong to the *betacoronavirus genus*. [21] All human coronaviruses are considered to be of zoonotic origin, with Chinese bats being the most likely host for SARS-CoV-2 [22]–[24]. Genetically, about 96% identity was observed between SARS-CoV-2 and bat coronavirus (BatCoV RaTG13). [17] Conversely, since bat habitats remain distanced from human life, the virus may have involved an intermediate animal such as pangolins before transmitting to humans. [25]–[29]

China Center for Disease Control and Prevention (CDC) primarily suggested the Huanan local seafood market as the origin of the COVID-19 outbreak. [30] Despite this, none of the animals in the area were tested positive for the virus. This suggests that the virus has already moved around Wuhan, long before its outbreak. Since then, the control of viral transmission through non-

therapeutic interventions suggested by the World Health Organization (WHO) has been attempted. [31] However, the violation of these preventive measures and absence of proper antiviral therapeutics and vaccinations has led to an uncontrollable transmission of the disease. The virus spread rapidly both inside and outside of China. In March 2020, the disease was declared as a global pandemic by the World Health Organization (WHO). [32] Although, at the beginning of the pandemic, the intensity of the disease was higher in American and European countries but later it also spread to Asian and south-east Asian countries.

Bangladesh, as one of the most densely populated countries of the world, has been susceptible to the coronavirus disease 2019 (COVID-19). Today, it has been labelled as the second-most affected country in South Asia.[33] The country with limited finance, and scarce facilities experiences major challenges at combating this transmission. The first case of this virus was confirmed in two men coming from Italy and a female relative by the country's epidemiology institute, the Institute of Epidemiology, Disease Control and Research (IEDCR) on March 7th, 2020. Although many Bangladeshi citizens came from Wuhan beforehand, they were reported negative for SARS-CoV-2. As a response, Bangladesh government took a number of preventive measures including imposing restrictions on international flights, strengthening of screening procedures and shutting down of educational institutions and so on. [34]

The first complete genome sequencing of SARS-CoV-2 in Bangladesh was announced by Child Health Research Foundation on 12th May, 2020. [35] Soon after, National Institute of Biotechnology announced the sequencing of SARS-CoV-2 genome by Sanger sequencing method [36]. The SARS-CoV-2 genome sequencing effort in Bangladesh flourished afterwards and as a result, 323 genomes have been sequenced by October 1st, 2020. The goal of this study is to probe all these sequences and find some crucial answers which will make it easier to comprehend the trajectory of this pandemic in Bangladesh.

2. Materials and Methods

2.1 Retrieval of the SARS-CoV-2 Genome Sequences

Genomes of SARS-CoV-2 isolates were retrieved from the Global Initiative on Sharing All Influenza Data (GISAID) database (www.gisaid.org). [37] Genomes were collected based on two criteria. First of all, they should have had complete length (>29,000 base pair length). Secondly, they had to have high coverage according to GISAID. Three groups of SARS-CoV-2 genome sequences were collected- sequences from Bangladesh, countries with large Bangladeshi diaspora and countries with similar mortality rate due to COVID-19. The mortality rate due to COVID-19 data was collected from (https://www.worldometers.info/) website.

2.2 Classification of the Genomes

The genomes were separately collected from GISAID clade GR, GH, G, S, O and L. Linages of these genomes were identified using Phylogenetic Assignment of Named Global Outbreak LINeages (Pangolin) COVID-19 Lineage Assigner (https://pangolin.cog-uk.io/). [38] After that their distribution, timeline and root/ origin were visualized through Microreact. [39]

2.3 Phylogenetic Analysis of the Classified SARS-CoV-2 Genomes

The classified genomes were subjected to phylogenetic analysis. The phylogenetic tree was constructed using previously reported methods. [36] In short, the genomes were aligned by MAFFT and the tree was constructed with FastTree. [40], [41] All of these steps were executed via the Galaxy Platform. [42] The constructed tree was visualized and edited by iTOL. [43] However, here we deducted the genomes that contain more than one 'X' amino acid substitutions according to CoVsurver application (https://www.gisaid.org/epiflu-applications/covsurver-mutations-app/). These genomes might give confusing results.

2.4 Detection of Genomic Variances with Their Consequences in Molecular Diagnosis

The genomic variances were detected using CoV-GLUE (http://cov-glue.cvr.gla.ac.uk). [44] CoV-GLUE identified Single Nucleotide Polymorphisms (SNPs), issues regarding primers, probes and whole genome sequencing. The web application also provided information about mutations in protein coding regions.

2.5 Revealing the Distribution of SARS-CoV-2 Clades and Mutations

Using the clade filter in GISAID, the distribution of clades within the three groups of samples (samples from Bangladesh, countries with large Bangladeshi diaspora and countries with similar mortality rate) were revealed. Only countries that had at least 100 genomes sequenced were considered. Distribution of major SARS-CoV-2 mutations found worldwide were also checked for these three groups. Mutation filter in GISAID was used for this analysis. Distribution of clades and mutation among the administrative divisions in Bangladesh was also evaluated using the same method.

2.6 Dynamics of SARS-CoV-2 Mutations in Bangladesh

Major mutations detected in each month starting from the beginning of April, 2020 till the end of September, 2020 were scrutinized for revealing the increase/decrease of the mutations over time. Rate of mutations per isolate in Bangladesh for the same period was also determined.

3. Results

3.1 Classification of the SARS-CoV-2 genomes from Bangladesh

From the first instance of SARS-CoV-2 genome submission from Bangladesh (12th May, 2020) to the time of the present study (1st October, 2020), the GISAID database recorded the genomes of 337 SARS-CoV-2 isolates. Among them, 302 (around 89%) isolates had complete length and high coverage according to the definition of GISAID. Among 302 viral isolates, 223 genomes were the member of lineage B 1.1.25 from GR clades (**Table 1**). The second largest group was B 1.1 of GR clade with 48 isolates. The root of these major groups were identified in UK (**Supplementary File 1**). Both B 1.1.25 and B 1.1 represented 1 and 2 isolates from G clade respectively. B.1 of G and O clades have 10 isolates totally. Only two members were found in B.1.1.59 and B.1.1.60 of GR clade. These members share common genomic features with Wales and Northern Ireland. Besides, 4 viral isolates from S and L clades were originated in China. These 4 genomes, especially from the O clade, share close ancestral relationships with some Chinese SARS-CoV-2 isolates which initiated COVID-19 pandemic.

Most of the viral isolates analyzed showed some issues regarding diagnostic amplification and whole genome sequencing. Many of them are known issues. However, some of them are yet to be

discussed. Most of these issues have raised due to the SNPs of N gene. In contrast, issues for ORF1ab gene are relatively low. Protocols provided by US and Chinese CDC demonstrated 21 and 13 unknown issues respectively. On the other hand, procedures provided by HKU Med and Ministry of Public Health, Thailand only showed 2 issues (**Table 4**). Wuhan coronavirus (2019-nCoV) real-time RT-PCR N gene methodology also faced some unknown issues but this approach is currently not recommended.

3.2 SARS-CoV-2 Clades within Bangladesh

It has been observed that the GR clade of SARS-CoV-2 is present in the highest proportion in all divisions within Bangladesh (**Fig 2a**). Dhaka and Chittagong had two of the most diverse array of SARS-CoV-2 viral clades. Mymensing solely possessed GR clades while Rajshahi and Sylhet contained only GH and GR clades. Hence these three divisions can be considered as regions within Bangladesh with lesser diversity of SARS-CoV-2. The largest number of viruses found within Bangladesh belonged to GR clade (**Fig 2b**). Among them, Dhaka division alone possessed 75 members of the GR clade. The smallest number of viruses pertained to O clade with only 3 members. Two of them were found in Dhaka and 1 in Chittagong division.

3.3 SARS-CoV-2 Mutations within Bangladesh

Our analyses have revealed that currently there are 9 major SARS-CoV-2 mutations present within the Bangladeshi population (**Table 2**). Among them the most commonly found were Spike_D614G and NSP12_P323L which were present in 100% of the SARS-CoV-2 isolates in almost all divisions (**Fig 3a**). The next most commonly detected mutations were N_G204R and N_R203K which were found among over 90% of the samples in almost all divisions. The most rarely found mutation was N_P13L which was found among 6.85% of the samples from Chittagong. In terms of number of mutations, highest number of mutations were detected in Dhaka followed by Chittagong (**Fig 3b**). The least number of mutations were detected in Mymensing division. NSP12_P323L and Spike_D614G were detected from 84 and 82 isolates from Dhaka division respectively. These are two of the largest instances of mutation within any division in terms of number. **Table 3** depicts the number of mutations per isolate at amino acid level in Bangladesh. So far the overall the number of mutations per isolate stands at 6.88. In terms of number of occurrences of each mutation there seemed to be only a few mutations that tower over the rest. The top 5 mutations namely NSP12_P323L, Spike_D614G, N_G204R, N_R203K, NSP2_I120F

occurred 314, 316, 278, 278, and 254 times respectively. NS3_Q57H, N_S194L, and NSP15_V35F occurred 30, 25, and 16 times respectively. Three mutations have occurred 11 times. Apart from these, all the rest of the mutations occurred less than 10 times (**Supplementary file 2**). Two mutations in the Spike Glycoprotein Receptor Binding Domain (RBD) were observed in our study. All the mutations of S were filtered considering diagnostics issues and their presence in the RBD was visualized by CoVsurver application. L518I mutations were observed in 3 viral isolates (Accession: EPI_ISL_483703, EPI_ISL_480447, EPI_ISL_483694) of GH clade (Lineage: B.1.36) (**Fig 1**). Another mutation, E516Q, was found in one isolate (Accession: EPI_ISL_475571) of O clade (Lineage: B.1). These 4 viral isolates were found from 4 different divisions of Bangladesh on June 2020. E516Q already occurred 4 times and first identified in hCoV-19/Belgium/ULG-9641/2020 on March 2020. This substitution might increase the flexibility of the domain (http://biosig.unimelb.edu.au/covid3d/mutation/QHD43416/AB/E516Q/E). L518I already occurred 3 times in Bangladesh and once in hCoV-19/USA/UT-UPHL-01685/2020 on April 2020.

3.4 Dynamics of SARS-CoV-2 Mutations in Bangladesh

Type of mutation in SARS-CoV-2 in Bangladesh was found to increase between April-July, 2020 (**Fig 4**). The diversity and overall accumulation of mutation seemed to follow a downward trajectory since August. The variety of mutations was highest in May while the overall amount of mutations reached its peak in July. In July, the mutations NSP12_P323L (95.83%), Spike_D614G (96.67%), and N_G204R (85.83%) was present in the highest percentage. In the following month, only 2 types of mutations were detected in Bangladesh namely N_G204R and N_R203K. Both were present in 100% of the SARS-CoV-2 isolates collected during that month. September saw a drastic fall in the percentage of mutations with each of the mutations, Spike_D614G, N_R203K, and N_G204R being present among 20% of the samples.

3.5 Distribution of Major SARS-CoV-2 Mutations in Relevant Countries

When compared with countries that have significant amount of Bangladeshi diaspora, it can be seen that in almost all them the mutations Spike_D614G and NSP12_P323L were most prominent (**Fig 5a**). N_R203K and N_206R were found in almost the same percentage in each country although they differ significantly between countries. The rest of the mutations also vary widely in their distribution. Data from countries that had similar mortality rates due to COVID-19 revealed

that the mutations Spike_D614G and NSP12_P323L were most prevalent among those countries (**Fig 5b**). However, their level varied widely between the countries. The rest of the mutations varied in even greater degree in between those countries.

4. Discussion

Some sequencing issues have been detected within the SARS-CoV-2 isolates that have been analyzed (**Table 4**). Therefore continuous modifications to COVID-19 Reverse Transcriptase (RT)-Polymeric Chain Reaction (PCR) protocols are being suggested especially for the Nucleoprotein (N) coding gene for accurate detection of the virus. Identification of clades and mutations of a certain virus in an area is very important because certain drugs and vaccines might not be fully effective against viruses belonging to certain clades or possessing certain mutations. From our analysis, GR clade has been identified as the predominant clade found within Bangladesh. It has been found in the highest percentage is every division within the country and even was the only clade found in one of the divisions (Mymensing) (**Fig 2a**). Overall it can be observed that the diversity of SARS-CoV-2 clades in Bangladesh is higher in southern parts of the country that the northern parts. A metadata analysis of 60,703 SARS-CoV-2 genomes revealed GR clade to be the most predominant clade worldwide. [45] GR clade is defined by mutations at certain positions in the SARS-CoV-2 genome (A23403G, C14408T, C3037T, G28881A, G28882A, G28883C and C241T in the 5' UTR). [46] According to another study, G and GR clade of SARS-CoV-2 is mainly observed in Europe while S and GH is found in the Americas. The L clade is found mostly in Asia. [47]

Spike_D614G, NSP12_P323L, N_G204R and N_R203K were the most widely found mutations found in Bangladesh (**Fig 3a**). In total, mutations in 448 locations have been found within the 323 SARS-CoV-2 genomes sequenced in Bangladesh till October 1st, 2020. Out of these 448 locations, 437 locations had mutational events occurring less than 10 times. Moreover, 304 locations had mutations only once and 75 locations had only twice. The top 5 locations in the SARS-CoV-2 genomes in Bangladesh where mutations have taken place were NSP12_P323L, Spike_D614G, N_G204R, N_R203K, NSP2_I120F which occurred 316, 314, 278, 278 and 254 times respectively (**Supplementary File 2**). Some GH Clade B.1.36 lineage SARS-CoV-2 isolates have been found with mutations in Spike Glycoprotein Receptor Binding Domain (RBD). These mutations might made the virus less sensitive to neutralizing antibodies that were generated against wild type Spike

Glycoprotein. However, the mutation D614G is found in the interface between the protomers that stabilize the mature spike trimer on the surface, instead of the RBD. [48] It has been observed that D614G mutant did not bind ACE2 more efficiently than its native counterpart. However, it has been found to decrease S1 shedding and make the Spike protein more stable than the native one. It also elevates the incorporation of S-protein into the virion. Epidemiological studies have also revealed that the viruses with D614G transmit more efficiently. [49] NSP12_P323L has been observed to co-evolve with Spike_D614G across the world. SARS-CoV-2 NSP12 is an RNA-dependent RNA polymerase. P323L has been found to enhance the interaction between NSP12 and NSP8 which is likely to increase the processivity of NSP12 and thus boost viral replication. [50] The third and fourth most commonly found mutations in Bangladesh namely, R203K and G204R in N were suggested to be favourable to SARS-CoV-2 adaptability and proliferation. Modelling of N_R203K and N_G204R mutants suggested that these mutations could cause drastic alterations in the N protein structure. [51]

We calculated the number of mutations per sample in Bangladesh to be 6.88 (**Table 3**). It happens to be lower than the global average which was reported to be 7.23 mutations per sample. [47] It was observed from our analysis that the type and the number of mutations in Bangladesh began to drop since August, 2020 (**Fig 4**).

When compared with countries that have significant amount of Bangladeshi diaspora, it can be seen that in almost all them the mutations Spike_D614G and NSP12_P323L were most prominent (**Fig 5a**). N_R203K and N_206R were found in almost the same percentage in each country although they differ significantly between countries. The rest of the mutations also vary widely in their distribution. Data from countries that had similar mortality rates due to COVID-19 revealed that the mutations Spike_D614G and NSP12_P323L were most prevalent among those countries (Fig 5b). However, their level varied widely between the countries. The rest of the mutations varied in even greater degree in between those countries. There was no significant correlation found between the percentage of each mutation in Bangladesh and the ones in countries with large Bangladeshi diaspora as well as countries with COVID-19 related mortality rates. The only thing that was found in common is the prominence of Spike_D614G and NSP12_P323L and equal distribution of N_R203K and N_206R which can be regarded as almost universal. However, their

level varied greatly between countries evaluated in our study and no solid pattern could be drawn which is common between Bangladesh and these countries.

Lower rate of mutation and the fall in overall amount of mutations since August should provide hope for Bangladesh as it increases the likelihood for the success of a vaccine or drug in the near future. The response to vaccines in the countries that share some degree of similarity with Bangladesh in terms of distribution of mutations should be monitored closely as the vaccine might have a similar impact in this country.

5. Conclusion

In this study, we have analyzed 323 SARS-CoV-2 complete genomes sequenced in Bangladesh till October 1st, 2020. Phylogenetic analysis revealed that among the SARS-CoV-2 viruses detected in Bangladesh, the ones belonging to GH and GR clades originated in the UK, the ones in G and O clades in Italy and the ones in S and O clades originated in China. The most commonly found mutant SARS-CoV-2 in Bangladesh were Spike_D614G, NSP12_P323L, N_G204R and N_R203K. Therefore vaccines and drugs targeting these mutants are likely to work best in the country. The distribution of predominant mutations have been found to be similar in between Bangladesh and the countries with similar mortality rates due to COVID-19 and the countries with large Bangladeshi diaspora although their level varied greatly and so no concrete pattern could be drawn between Bangladesh and these countries. However, the distribution of mutations in the SARS-CoV-2 genomes in Bangladesh somewhat matches that in the UK, Oman, Italy, Greece, South Africa and Russia. Therefore, in the near future, close attention should be paid to vaccines working in those countries. If they work well there, there is a good chance they might work well also in Bangladesh. The good news for Bangladesh is that the number of mutations per sample in the country has turned out to be lower than the global average. The rate of mutation has been found to be decreasing in Bangladesh since August, 2020. These features are likely to make it easier for a vaccine or drug to be effective and sustainable in providing protection for the Bangladeshi population.

6. References

[1] D. J. Muscatello and P. B. McIntyre, "Comparing mortalities of the first wave of coronavirus disease 2019 (COVID-19) and of the 1918–19 winter pandemic influenza wave in the USA," *Int. J. Epidemiol.*, Sep. 2020, doi: 10.1093/ije/dyaa186.

[2] M. Hasanul Banna Siam, M. Mahbub Hasan, E. Raheem, M. Hasinur Rahaman Khan, M. H. Siddiqee, and M. Sorowar Hossain, "Insights into the first wave of the COVID-19 pandemic in Bangladesh: Lessons learned from a high-risk country," *medRxiv*, p. 2020.08.05.20168674, 1101, doi: 10.1101/2020.08.05.20168674.

[3] B. Salzberger *et al.*, "Epidemiology of SARS-CoV-2," *Infection*, vol. 1. Springer Science and Business Media Deutschland GmbH, p. 3, Oct. 08, 2020, doi: 10.1007/s15010-020-01531-3.

[4] G. Moore *et al.*, "Detection of SARS-CoV-2 within the healthcare environment: a multicentre study conducted during the first wave of the COVID-19 outbreak in England," *medRxiv*, p. 2020.09.24.20191411, Sep. 2020, doi: 10.1101/2020.09.24.20191411.

[5] S. Ryu, E. Noh, S. T. Ali, D. Kim, E. H. Y. Lau, and B. J. Cowling, "Epidemiology and Control of Two Epidemic Waves of SARS-CoV-2 in South Korea," *SSRN Electron. J.*, Sep. 2020, doi: 10.2139/ssrn.3687061.

[6] N. D. Grubaugh *et al.*, "Genomic epidemiology reveals multiple introductions of Zika virus into the United States," *Nature*, vol. 546, no. 7658, pp. 401–405, Jun. 2017, doi: 10.1038/nature22400.

[7] J. T. Ladner, N. D. Grubaugh, O. G. Pybus, and K. G. Andersen, "Precision epidemiology for infectious disease control," *Nature Medicine*, vol. 25, no. 2. Nature Publishing Group, pp. 206–211, Feb. 2019, doi: 10.1038/s41591-019-0345-2.

[8] C. C. Kalinich *et al.*, "Real-time public health communication of local SARS-CoV-2 genomic epidemiology," *PLoS Biol.*, vol. 18, no. 8, p. e3000869, Aug. 2020, doi: 10.1371/JOURNAL.PBIO.3000869.

[9] X. Deng *et al.*, "Genomic surveillance reveals multiple introductions of SARS-CoV-2 into Northern California."

[10] L. W. Meredith *et al.*, “Rapid implementation of SARS-CoV-2 sequencing to investigate cases of health-care associated COVID-19: a prospective genomic surveillance study,” *Lancet Infect. Dis.*, vol. 20, no. 11, pp. 1263–1272, Nov. 2020, doi: 10.1016/S1473-3099(20)30562-4.

[11] T. Covid- and G. U. consortium, “An integrated national scale SARS-CoV-2 genomic surveillance network,” *The Lancet Microbe*, vol. 1, no. 3, pp. e99–e100, Jul. 2020, doi: 10.1016/s2666-5247(20)30054-9.

[12] X. Deng *et al.*, “Genomic surveillance reveals multiple introductions of SARS-CoV-2 into Northern California,” *Science (80-.).*, vol. 369, no. 6503, pp. 582–587, Jul. 2020, doi: 10.1126/science.abb9263.

[13] P. Cristina Resende *et al.*, “Genomic surveillance of SARS-CoV-2 reveals community transmission of a major lineage 1 during the early pandemic phase in Brazil 2 3,” doi: 10.1101/2020.06.17.158006.

[14] D. M. Morens *et al.*, “The Origin of COVID-19 and Why It Matters,” *American Journal of Tropical Medicine and Hygiene*, vol. 103, no. 3. American Society of Tropical Medicine and Hygiene, pp. 955–959, Sep. 01, 2020, doi: 10.4269/ajtmh.20-0849.

[15] T. Bolsen, R. Palm, and J. T. Kingsland, “Framing the Origins of COVID-19,” *Sci. Commun.*, vol. 42, no. 5, pp. 562–585, Oct. 2020, doi: 10.1177/1075547020953603.

[16] T. Burki, “The origin of SARS-CoV-2,” *Lancet. Infect. Dis.*, vol. 20, no. 9, pp. 1018–1019, Sep. 2020, doi: 10.1016/S1473-3099(20)30641-1.

[17] P. Zhou *et al.*, “A pneumonia outbreak associated with a new coronavirus of probable bat origin,” *Nature*, vol. 579, no. 7798, pp. 270–273, Mar. 2020, doi: 10.1038/s41586-020-2012-7.

[18] R. A. Khailany, M. Safdar, and M. Ozaslan, “Genomic characterization of a novel SARS-CoV-2,” *Gene Reports*, vol. 19, p. 100682, Jun. 2020, doi: 10.1016/j.genrep.2020.100682.

[19] A. A. T. Naqvi *et al.*, “Insights into SARS-CoV-2 genome, structure, evolution, pathogenesis and therapies: Structural genomics approach,” *Biochimica et Biophysica Acta - Molecular Basis of Disease*, vol. 1866, no. 10. Elsevier B.V., p. 165878, Oct. 01, 2020,

doi: 10.1016/j.bbadis.2020.165878.

[20] S. Payne, “Family Coronaviridae,” in *Viruses*, Elsevier, 2017, pp. 149–158.

[21] M. U. Hossain *et al.*, “Recognition of Plausible Therapeutic Agents to Combat COVID-19: An Omics Data Based Combined Approach,” Apr. 2020, doi: 10.21203/rs.3.rs-25807/v1.

[22] C. K. Chang *et al.*, “Modular organization of SARS coronavirus nucleocapsid protein,” *J. Biomed. Sci.*, vol. 13, no. 1, pp. 59–72, Jan. 2006, doi: 10.1007/s11373-005-9035-9.

[23] N. Chen *et al.*, “Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study,” *Lancet*, vol. 395, no. 10223, pp. 507–513, Feb. 2020, doi: 10.1016/S0140-6736(20)30211-7.

[24] P. C. Y. Woo *et al.*, “Discovery of Seven Novel Mammalian and Avian Coronaviruses in the Genus Deltacoronavirus Supports Bat Coronaviruses as the Gene Source of Alphacoronavirus and Betacoronavirus and Avian Coronaviruses as the Gene Source of Gammacoronavirus and Deltacoronavirus,” *J. Virol.*, vol. 86, no. 7, pp. 3995–4008, Apr. 2012, doi: 10.1128/jvi.06540-11.

[25] T. T. Y. Lam *et al.*, “Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins,” *Nature*, vol. 583, no. 7815, 2020, doi: 10.1038/s41586-020-2169-0.

[26] B. Hong *et al.*, “SARS-CoV-2 and Malayan pangolin coronavirus infect human endoderm, ectoderm and induced lung progenitor cells,” *bioRxiv*, p. 2020.09.25.313270, Sep. 2020, doi: 10.1101/2020.09.25.313270.

[27] T. Zhang, Q. Wu, and Z. Zhang, “Probable Pangolin Origin of SARS-CoV-2 Associated with the COVID-19 Outbreak,” *Curr. Biol.*, vol. 30, no. 7, pp. 1346-1351.e2, Apr. 2020, doi: 10.1016/j.cub.2020.03.022.

[28] G. Z. Han, “Pangolins Harbor SARS-CoV-2-Related Coronaviruses,” *Trends in Microbiology*, vol. 28, no. 7. Elsevier Ltd, pp. 515–517, Jul. 01, 2020, doi: 10.1016/j.tim.2020.04.001.

[29] K. Xiao *et al.*, “Isolation of SARS-CoV-2-related coronavirus from Malayan pangolins,” *Nature*, vol. 583, no. 7815, pp. 286–289, Jul. 2020, doi: 10.1038/s41586-020-2313-x.

[30] S. P. Adhikari *et al.*, “Epidemiology, causes, clinical manifestation and diagnosis, prevention and control of coronavirus disease (COVID-19) during the early outbreak period: A scoping review,” *Infectious Diseases of Poverty*, vol. 9, no. 1. BioMed Central Ltd., pp. 1–12, Mar. 2020, doi: 10.1186/s40249-020-00646-x.

[31] M. M. Böhmer *et al.*, “Investigation of a COVID-19 outbreak in Germany resulting from a single travel-associated primary case: a case series,” *Lancet Infect. Dis.*, vol. 20, no. 8, pp. 920–928, Aug. 2020, doi: 10.1016/S1473-3099(20)30314-5.

[32] “Coronavirus Disease (COVID-19).” .

[33] “Coronavirus Update (Live): 44,431,995 Cases and 1,174,527 Deaths from COVID-19 Virus Pandemic - Worldometer.” .

[34] H. Kabir, M. Maple, and K. Usher, “The impact of COVID-19 on Bangladeshi readymade garment (RMG) workers,” *J. Public Health (Bangkok).*, pp. 1–6, Jul. 2020, doi: 10.1093/pubmed/fdaa126.

[35] S. Saha *et al.*, “Complete Genome Sequence of a Novel Coronavirus (SARS-CoV-2) Isolate from Bangladesh,” *Microbiol. Resour. Announc.*, vol. 9, no. 24, Jun. 2020, doi: 10.1128/mra.00568-20.

[36] M. Moniruzzaman *et al.*, “Coding-Complete Genome Sequence of SARS-CoV-2 Isolate from Bangladesh by Sanger Sequencing,” *Microbiol. Resour. Announc.*, vol. 9, no. 28, Jul. 2020, doi: 10.1128/mra.00626-20.

[37] S. Elbe and G. Buckland-Merrett, “Data, disease and diplomacy: GISAID’s innovative contribution to global health,” *Glob. Challenges*, vol. 1, no. 1, pp. 33–46, Jan. 2017, doi: 10.1002/gch2.1018.

[38] A. Rambaut *et al.*, “A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology,” *Nat. Microbiol.*, vol. 5, no. 11, pp. 1403–1407, Nov. 2020, doi: 10.1038/s41564-020-0770-5.

[39] S. Argimón *et al.*, “Microreact: visualizing and sharing data for genomic epidemiology and phylogeography,” *Microb. genomics*, vol. 2, no. 11, p. e000093, Nov. 2016, doi: 10.1099/mgen.0.000093.

[40] K. Katoh, J. Rozewicki, and K. D. Yamada, “MAFFT online service: Multiple sequence alignment, interactive sequence choice and visualization,” *Brief. Bioinform.*, vol. 20, no. 4, pp. 1160–1166, Mar. 2018, doi: 10.1093/bib/bbx108.

[41] M. N. Price, P. S. Dehal, and A. P. Arkin, “FastTree 2 - Approximately maximum-likelihood trees for large alignments,” *PLoS One*, vol. 5, no. 3, p. e9490, Mar. 2010, doi: 10.1371/journal.pone.0009490.

[42] E. Afgan *et al.*, “The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update,” *Nucleic Acids Res.*, vol. 46, pp. 537–544, 2018, doi: 10.1093/nar/gky379.

[43] I. Letunic and P. Bork, “Interactive Tree of Life (iTOL) v4: Recent updates and new developments,” *Nucleic Acids Res.*, vol. 47, no. W1, Jul. 2019, doi: 10.1093/nar/gkz239.

[44] J. B. Singer, R. J. Gifford, M. Cotten, and D. L. Robertson, “CoV-GLUE: A Web Application for Tracking SARS-CoV-2 Genomic Variation,” Jun. 2020, doi: 10.20944/preprints202006.0225.v1.

[45] S. M. Hamed, W. F. Elkhatib, A. S. Khairallah, and A. M. Noreddin, “Global dynamics of SARS-CoV-2 clades and their relation to COVID-19 epidemiology,” doi: 10.21203/rs.3.rs-89876/v1.

[46] B. Korber *et al.*, “Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus,” *Cell*, vol. 182, no. 4, pp. 812-827.e19, Aug. 2020, doi: 10.1016/j.cell.2020.06.043.

[47] D. Mercatelli and F. M. Giorgi, “Geographic and Genomic Distribution of SARS-CoV-2 Mutations,” *Front. Microbiol.*, vol. 11, p. 1800, Jul. 2020, doi: 10.3389/fmicb.2020.01800.

[48] N. D. Grubaugh, W. P. Hanage, and A. L. Rasmussen, “Making Sense of Mutation: What D614G Means for the COVID-19 Pandemic Remains Unclear,” *Cell*, vol. 182, pp. 794–795, 2020, doi: 10.1016/j.cell.2020.06.040.

[49] L. Zhang *et al.*, “The D614G mutation in the SARS-CoV-2 spike protein reduces S1 shedding and increases infectivity,” doi: 10.1101/2020.06.12.148726.

[50] S. R. Kannan *et al.*, "Infectivity of SARS-CoV-2: there Is Something More than D614G?," doi: 10.1007/s11481-020-09954-3.

[51] S. Wu *et al.*, "Effects of SARS-CoV-2 mutations on protein structures and intraviral protein–protein interactions," *J. Med. Virol.*, p. jmv.26597, Nov. 2020, doi: 10.1002/jmv.26597.

Table 1: Classification of SARS-CoV-2 genomes from Bangladesh.

Serial No.	Pangolin Lineage	GISAID Clade	Number of isolates	Lineage description	Most common countries
1	B 1.1.25	GR	223	Bangladesh lineage	Bangladesh (94%), UK (3%), Australia (2%)
2	B.1.1	GR	48	European lineage with 3 clear SNPs `28881GA`,`28882GA`,`28883GC`	UK (84%), USA (3%), Portugal (2%)
3	B.1.1.59	GR	1	Wales lineage	UK (99%), Bangladesh (1%)
4	B.1.1.60	GR	1	Northern Ireland lineage	UK (99%), Austria (0%), Bangladesh (0%)
5	B.1.36	GH	20	Global lineage with lots of representation of sequences from India, Saudi Arabia, Europe and UK.	India (45%), Saudi_Arabia (34%), Bangladesh (5%)
6	B.1	G	7	A large European lineage that corresponds to the Italian outbreak	UK (41%), USA (30%), Australia (3%)
7	B.1.1	G	2	European lineage with 3 clear SNPs `28881GA`,`28882GA`,`28883GC`	UK (84%), USA (3%), Portugal (2%)
8	B.1.1.25	G	1	Bangladesh lineage	Bangladesh (94%), UK (3%), Australia (2%)

9	B.1	O	3	A large European lineage that corresponds to the Italian outbreak	UK (41%), USA (30%), Australia (3%)
10	A	S	3	Root of the pandemic lies within lineage A	China (31%), India (10%), Japan (8%)
11	B	L	1	Base of this lineage lies in China. Contains two distinct SNPs `8782TC` and `28144CT	UK (42%), China (23%), USA (15%)

Table 2: Percentage of major mutations among the administrative divisions of Bangladesh

Divisions	**Mutations (%)**								
	Spike_ D614G	**N_P 13L**	**N_S1 94L**	**N_R2 03K**	**N_G2 04R**	**NS3_ Q57H**	**NS8_ L84S**	**NSP6_ L37F**	**NSP12_ P323L**
Barisal	100.00	0.00	4.35	91.30	91.30	13.04	0.00	0.00	100.00
Chittagong	91.78	1.37	19.18	73.97	73.97	19.18	6.85	6.85	93.15
Dhaka	96.47	0.00	1.18	91.76	91.76	2.35	0.00	3.53	98.82
Khulna	100.00	0.00	7.69	80.77	80.77	7.69	0.00	0.00	100.00
Mymensing	100.00	0.00	0.00	100.00	100.00	0.00	0.00	0.00	100.00
Rajshahi	100.00	0.00	6.06	93.94	93.94	6.06	0.00	3.03	100.00
Rangpur	100.00	0.00	4.00	96.00	96.00	8.00	0.00	4.00	100.00
Sylhet	100.00	0.00	9.52	90.48	90.48	9.52	0.00	4.76	100.00

Table 3: Rate of mutations at the amino acid level per isolate in Bangladesh

Overall number of mutations per isolate in Bangladesh	6.88
Number of mutations per isolate in April	5.75
Number of mutations per isolate in May	6.72
Number of mutations per isolate in June	6.97
Number of mutations per isolate in July	7.02

Table 4: Issues regarding molecular diagnosis of SARS-CoV-2 in Bangladesh for Various Protocols

Clade	**Gene name [Source of the protocols]**	**No. of unknown primer probe issues for sequence mismatches***
G	a) N [China CDC Primers and probes for detection 2019-nCoV (Link: https://tinyurl.com/y5jsjemt)] b) N [US CDC panel primer and probes â€“ U.S. CDC, USA (Link: https://tinyurl.com/y5blsevq)]	a) 3 b) 1
GH	a) N [China CDC Primers and probes for detection 2019-nCoV (Link: https://tinyurl.com/y5jsjemt)] b) N [PCR and sequencing protocol for 2019-nCoV - Ministry of Public Health, Thailand (Link: https://tinyurl.com/y355hecg)] c) N [US CDC panel primer and probes â€“ U.S. CDC, USA (Link: https://tinyurl.com/y5blsevq)] d) N [Wuhan coronavirus (2019-nCoV) real-time RT-PCR N gene 2020 (Link: https://tinyurl.com/y38jldqe)]	a) 1 b) 1 c) 3 d) 1 e) 1

	e) ORF1ab [China CDC Primers and probes for detection 2019-nCoV (Link: https://tinyurl.com/y5jsjemt)]	
GR	a) N [China CDC Primers and probes for detection 2019-nCoV (Link: https://tinyurl.com/y5jsjemt)] b) N [Detection of 2019 novel coronavirus (2019-nCoV) in suspected human cases by RT-PCR (HKU) (Link: https://tinyurl.com/y3zeduz8)] c) N [PCR and sequencing protocol for 2019-nCoV - Ministry of Public Health, Thailand (Link: https://tinyurl.com/y355hecg)] d) N [US CDC panel primer and probes â€“ U.S. CDC, USA (Link: https://tinyurl.com/y5blsevq)] e) N [Wuhan coronavirus (2019-nCoV) real-time RT-PCR N gene 2020 (Link: https://tinyurl.com/y38jldqe)] f) ORF1ab [China CDC Primers and probes for detection 2019-nCoV (Link: https://tinyurl.com/y5jsjemt)]	a) 8 b) 2 c) 1 d) 17 e) 2 f) 1

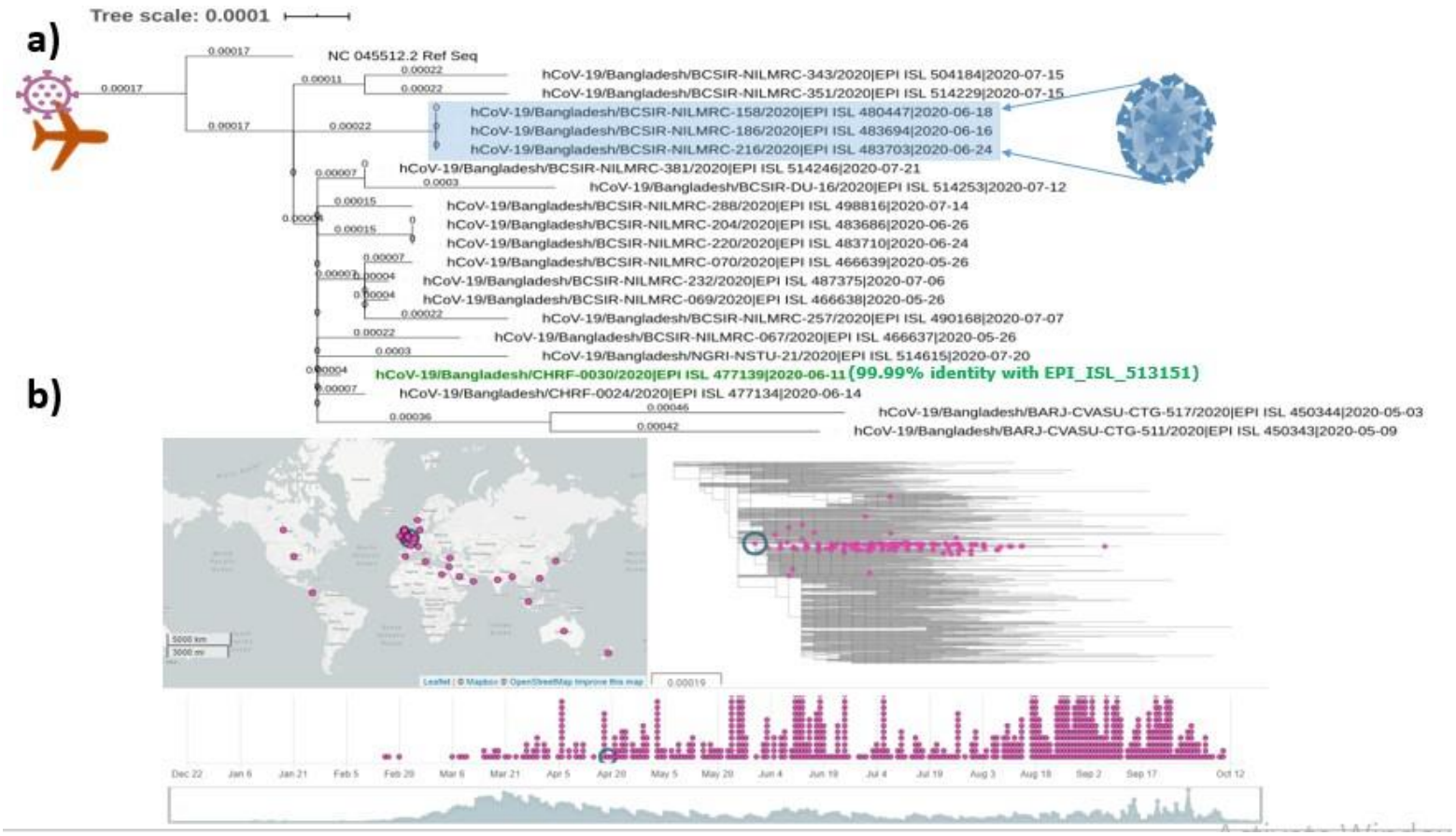


Fig 1: Phylogenetic characterization and transmission origin of the SARS-CoV-2 GH clade (Lineage: B.1.36) isolates from Bangladesh. **(a)** The phylogenetic relationships of the local GH clade isolates demonstrates 3 distinctive clusters. The blue labeled cluster contain mutation (L518I) in receptor binding domain of spike glycoprotein and the green colored isolate (near the root) has 99.99% genetic identity with the isolates of Saudi Arabia. **(b)** The root of these isolates were originated in UK and collected around April 29, 2020 in UK. Results of the phylogenetic analyses of the rest of the clades have been provided in **Supplementary File 1**.

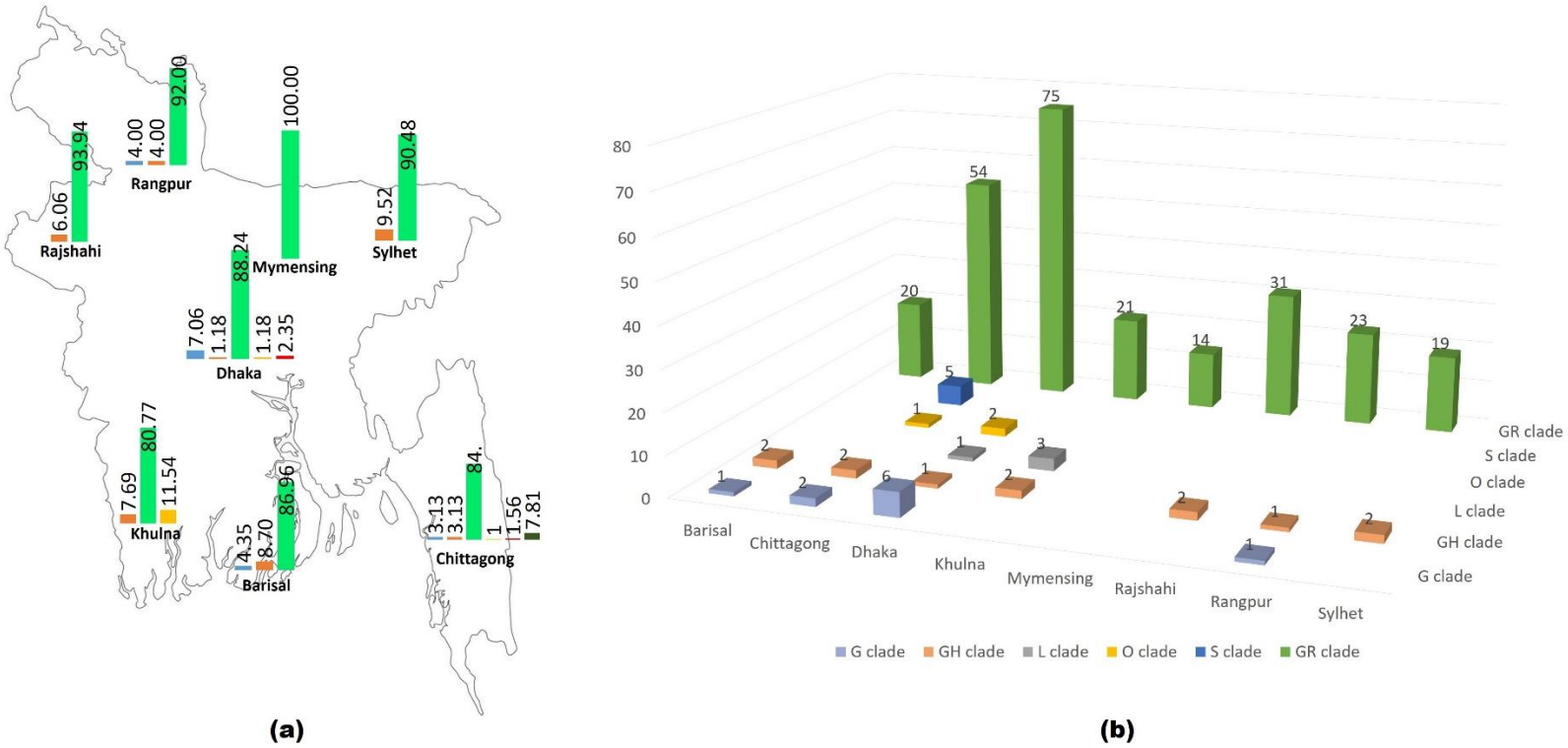
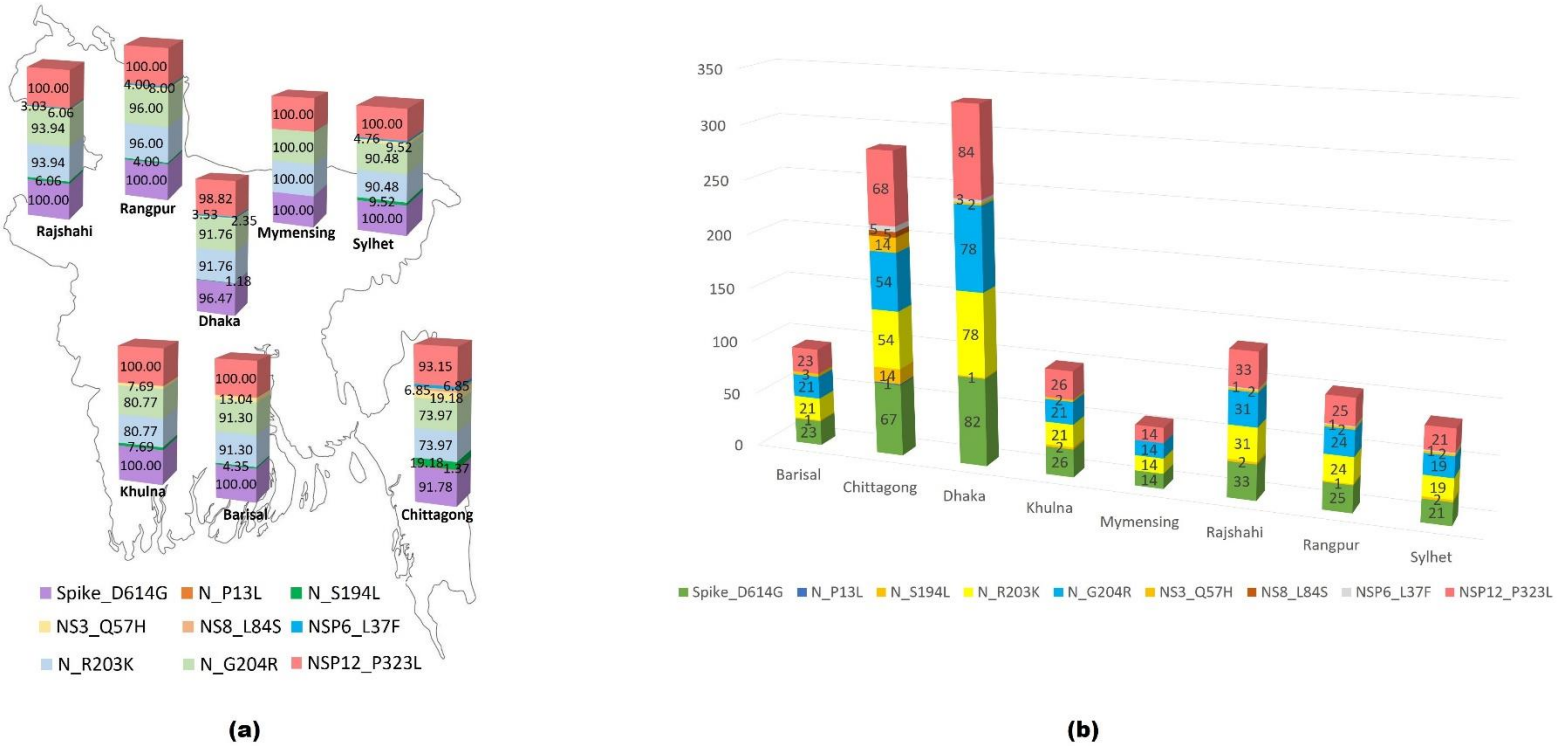


Fig 2: Distribution of SARS-CoV-2 clades in various administrative divisions of Bangladesh (**a**) In terms of percentage, Dhaka and Chittagong divisions were two of the most diverse areas in terms of clade variation while Mymensing was the least diverse. Sylhet and Rajshahi were also belonged to low diversity regions. GR clade is the most prominent clade in all divisions but their level differs from division to division. It was found among 100% of the samples from Mymensing division and 80.77% of the samples from Khulna division. (**b**) In terms of number, the number of GR clade towers over the rest of the clades in all divisions. Many of the clades were detected only once in a division such as G clade in Barisal and Rangpur, GH clade in Dhaka and Rangpur, L clade in Dhaka, and O clade in Chittagong.



(a) **(b)**

Fig 3: Distribution of major SARS-CoV-2 mutations in various administrative divisions of Bangladesh. (**a**) In terms of percentage, Spike_D614G and NSP12_P323L were detected in 100% of the samples in almost all divisions. Presence of N_G204R and N_R203K were detected in over 90% of the samples. The rarest mutation was N_P13L (found in only 6.85% of the samples from Chittagong). (**b**) In terms of number, Dhaka had the largest number of mutations while Mymensing had the least. Chittagong also exhibited high number of mutations. The mutations Spike_D614G and NSP12_P323L has been found in almost equal number in all divisions.

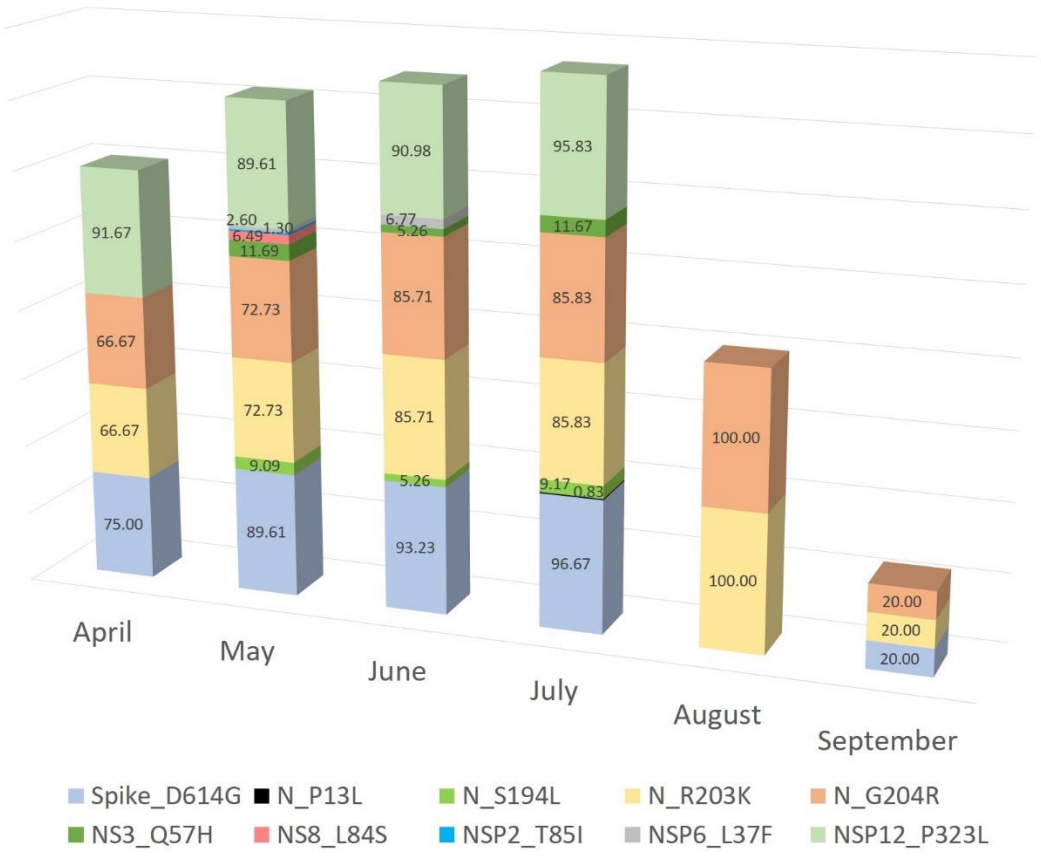


Fig 4: Dynamics of mutation over time in Bangladesh. Various types of mutations detected within the April-September, 2020 period has been presented in terms of percentage. The diversity of mutations was highest in May but the overall level of mutation was in its peak in July. Since August, the overall amount of mutations began to drop sharply.

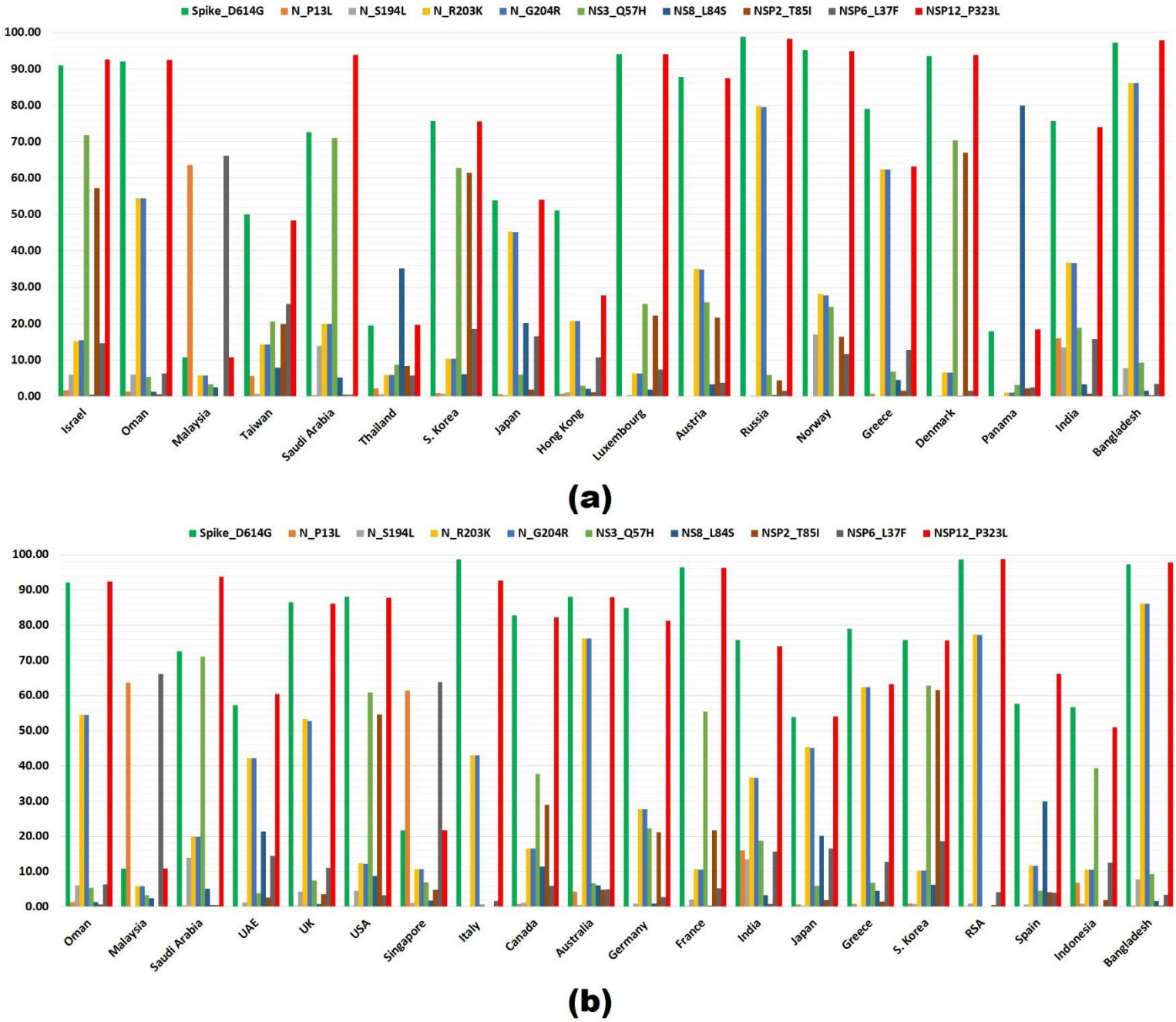


Fig 5: Distribution of mutations in Bangladesh and relevant countries. (**a**) Distribution of mutations in countries that had similar mortality rates as Bangladesh due to COVID-19. In all these countries Spike_D614G and NSP12_P323L were most prominent although their level varied greatly from country to country. The percentage of N_R203K and N_G204R was almost identical in all countries. Apart from these,

there is hardly any pattern that can be found in common between Bangladesh and the countries with similar mortality rates due to COVID-19. (**b**) Distribution of mutations in countries that host large Bangladeshi diaspora. Spike_D614G and NSP12_P323L were present in high percentage in all these countries. The percentage of N_R203K and N_G204R was almost identical everywhere but their level differed vastly between countries. Apart from these, there is hardly any pattern that can be found in common between Bangladesh and the countries with large Bangladeshi diaspora.