

RESEARCH ARTICLE

In silico comparative genomics of SARS-CoV-2 to determine the source and diversity of the pathogen in Bangladesh

Tushar Ahmed Shishir^{1,2}, Iftekhar Bin Naser^{1*}, Shah M. Faruque²

1 Department of Mathematics and Natural Sciences, BRAC University, Mohakhali, Dhaka, Bangladesh, **2** School of Environment and Life Sciences, Independent University, Bangladesh (IUB), Bashundhara, Dhaka, Bangladesh

* iftekhar.naser@bracu.ac.bd**OPEN ACCESS**

Citation: Shishir TA, Naser IB, Faruque SM (2021) In silico comparative genomics of SARS-CoV-2 to determine the source and diversity of the pathogen in Bangladesh. PLoS ONE 16(1): e0245584. <https://doi.org/10.1371/journal.pone.0245584>

Editor: Houssam Attoui, Institut National de la Recherche Agronomique, FRANCE

Received: September 2, 2020

Accepted: January 5, 2021

Published: January 20, 2021

Copyright: © 2021 Shishir et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: Variant calling files of genome sequences of SARS-CoV-2 isolated in Bangladesh can be found at <https://doi.org/10.6084/m9.figshare.13124936> and the initial 64 SARS-CoV-2 genome sequences reported from Bangladesh used in this study are available at <https://doi.org/10.6084/m9.figshare.13125026>.

Funding: The author(s) received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Abstract

The COVID19 pandemic caused by SARS-CoV-2 virus has severely affected most countries of the world including Bangladesh. We conducted comparative analysis of publicly available whole-genome sequences of 64 SARS-CoV-2 isolates in Bangladesh and 371 isolates from another 27 countries to predict possible transmission routes of COVID19 to Bangladesh and genomic variations among the viruses. Phylogenetic analysis indicated that the pathogen was imported in Bangladesh from multiple countries. The viruses found in the southern district of Chattogram were closely related to strains from Saudi Arabia whereas those in Dhaka were similar to that of United Kingdom and France. The 64 SARS-CoV-2 sequences from Bangladesh belonged to three clusters. Compared to the ancestral SARS-CoV-2 sequence reported from China, the isolates in Bangladesh had a total of 180 mutations in the coding region of the genome, and 110 of these were missense. Among these, 99 missense mutations (90%) were predicted to destabilize protein structures. Remarkably, a mutation that leads to an I300F change in the nsp2 protein and a mutation leading to D614G change in the spike protein were prevalent in SARS-CoV-2 genomic sequences, and might have influenced the epidemiological properties of the virus in Bangladesh.

Introduction

The pandemic of coronavirus disease referred to as COVID-19 pandemic, which originated in Wuhan, China in December 2019 is ongoing and has now spread to 213 countries and territories. As of October 2020, the pandemic has caused about 45 million cases and over half a million death. This novel virus of the Coronaviridae family and *Betacoronavirus* genus [1, 2] designated as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), is the causative agent of COVID-19. Previously two other coronaviruses, namely SARS-CoV and MERS-CoV demonstrated high pathogenicity and caused epidemic with mortality rate ~10% and ~34% respectively affecting more than 25 countries each time [3–5]. However, SARS-CoV-2 has proven to be highly infectious and caused pandemic spread to over 213 countries and territories. Besides its devastating impact in North America and Europe, the disease has now rapidly spread in South America including Brazil, and in South Asian countries particularly India, Pakistan and Bangladesh [6].

The virus was first detected in Bangladesh in March 2020 [7]. Although infections remained low until the end of March it began to rise steeply in April. By the end of June, new cases in Bangladesh grew to nearly 150,000 and the rate of detection of cases compared to the total number of samples tested increased to about 22% which was highest in Asia [6].

SARS-CoV-2 is a positive-sense single-stranded RNA virus with a genome size of nearly 30kb. The 5' end of the genome codes for a polyprotein which is further cleaved to viral non-structural proteins whereas the 3' end encodes for structural proteins of the virus including surface glycoproteins spike (S), membrane protein (M), envelop protein (E) and nucleocapsid protein (N) [8]. Like other RNA viruses, SARS-CoV-2 is also inherently prone to mutations due to high recombination frequency resulting in genomic diversity [9–11]. Due to the rapid evolution of the virus, development of vaccines and therapies may be challenging. To monitor the emergence of diversity, it is important to conduct comparative genomics of viruses isolated over time and in various geographical locations. Comparative analysis of genome sequences of various isolates of SARS-CoV-2 would allow to identify and characterize the variable and conserved regions of the genome and this knowledge may be useful for developing effective vaccines, as well as in molecular epidemiological tracking. Thousands of SARS-CoV-2 virus genomes have been sequenced and submitted in public databases for further study. This include 66 SARS-CoV-2 genomic sequences submitted from Bangladesh in the Global initiative on sharing all influenza data (GISAID) database, till 11th June 2020 [12]. We conducted comparative analysis of publicly available genome sequences of SARS-CoV-2 from 27 countries to predict the origin of viruses in Bangladesh by studying a time-resolved phylogenetic relationship. Later, we analyzed the variants present in different isolates of Bangladesh to understand the pattern of mutations in relation to the ancestral Wuhan strain, find unique mutations, and possible effect of these mutations on the stability of encoded proteins, and selection pressure on genes.

Materials and methods

Genome sequence acquisition

A total of 435 whole genome sequences of SARS-CoV-2 isolated in 27 countries (S1 Table) with high frequency of infection including 64 sequences isolated in Bangladesh (S2 Table), and that of 5 isolates of each month between January to May, 2020 in the selected countries obtained from GISAID [12] were included in this analysis (S3 Table). However, since only a small number of sequences were reported from different African countries, we included all available sequences from the African countries and categorized collectively as African sequence. Reference sequence included in various analysis was the sequence of the ancestral strain from Wuhan, China (NC_045512.2) [13]. All the sequences were quality checked prior to further analysis and any sequence found to contain ambiguous characters, were removed.

Sequence annotation, phylogenetic analysis and clustering

The selected sequences were annotated by Viral Annotation Pipeline and iDentification (VAPiD) tool [14] to identify their coding regions and corresponding amino acid sequences for further analysis. Subsequently, multiple sequence alignment was carried out using Mafft algorithm [15] and a maximum likelihood phylogenetic tree was constructed with IQ-TREE [16] with bootstrap value of 1000 to identify closeness among the sequences. Then the generated tree was reconstructed based on time-calibration by TreeTime [17]. The tree was then annotated and visualized on iTOL server [18] putting the sequences into different clades based on specific mutations proposed in GISAID [19], and further classified as D614G type [20, 21]. In addition, another phylogenetic tree containing only SARS-CoV-2 sequences from

Bangladesh was constructed and categorized using the same tools, and further clustered with TreeCluster [22] to group them based on their identity with each other.

Analysis of mutations and their predicted effects

For analysis of mutations, sequence were mapped with minimap2 [23], and variants were detected using SAMtools [24]. Furthermore, snp-sites [25] was also used to detect variations among genomes with the reference, and finally common mutations from these two analysis were considered. In addition, SNPeff [26] was used to predict the effect of mutations in terms of changes in amino acid sequences. Subsequently, a haplotype network was generated based on mutations in genome using PopArt by Integer Neighbor-Joining method [27]. Then the direction of selection in sequences from Bangladesh was calculated by the SLAC algorithm [28] in the Datamonkey server [29] to understand their evolutionary pattern as to whether they are stabilizing or changing with time. Finally, since it is important to understand the functional consequences of mutations on corresponding proteins, we used a manually curated neural network DeepDDG [30] to predict the effects of the mutations on protein stability.

Results and discussion

Phylogenetic analysis of all SARS-CoV-2 sequences

A total of 435 Genomic Sequences of SARS-CoV-2 reported from various countries which included 64 sequences from Bangladesh and the sequence of the ancestral SARS-CoV-2 isolated in Wuhan, China were analyzed in the time-resolved phylogenetic tree. Sequences from Bangladesh belonged to three different clusters of which one cluster carried 43 of the 64 sequences, and shared the same node with sequence from Germany while they had a common ancestry with isolates from the United Kingdom (Fig 1). The remaining two clusters of SARS-CoV-2 sequences contained 4 and 5 sequences respectively from Bangladesh, and they shared the same node with sequence of SARS-CoV-2 reported from India, and also shared a common ancestry with isolates from Saudi Arabia. Besides, 12 lone sequences that did not belong to any of these clusters were found to have similarity with sequences from Europe including United Kingdom, Germany, France, Italy, and Russia. One of these sequences was closely related to sequence reported from the USA. Subsequently, all SARS-CoV-2 sequences from representative countries were clustered based on some specific mutations sustained, into 7 different clades as mentioned by GISAID. In this analysis, the sequences from Bangladesh were found to be distributed in all clades except V (Figs 1 and 2).

Phylogenetic analysis of SARS-CoV-2 sequences from Bangladesh

In order to understand the evolutionary relationship and possible transmission dynamics of SARS-CoV-2 in Bangladesh at a higher resolution, another time-resolved phylogenetic tree carrying only sequences of the pathogen isolated in various regions of Bangladesh was generated using the sequence of the first SARS-CoV-2 reported from Wuhan, China as a reference. Of the three clusters produced in this analysis, cluster-1 included mostly isolates from Chattogram and one isolate from Dhaka, cluster-2 included isolates from Dhaka, Narayanganj and Chattogram districts, whereas cluster-3 included isolates from Chattogram only. As mentioned above, the isolates from Bangladesh were found to be distributed in all 7 GISAID clades based on specific mutations, except in clade V (Figs 1 and 2). Most isolates of Dhaka and Narayanganj (47 of 52) belonged to the GR clade, whereas those of Chattogram belonged to five different GISAID clades (G, GH, GR, O, and S).

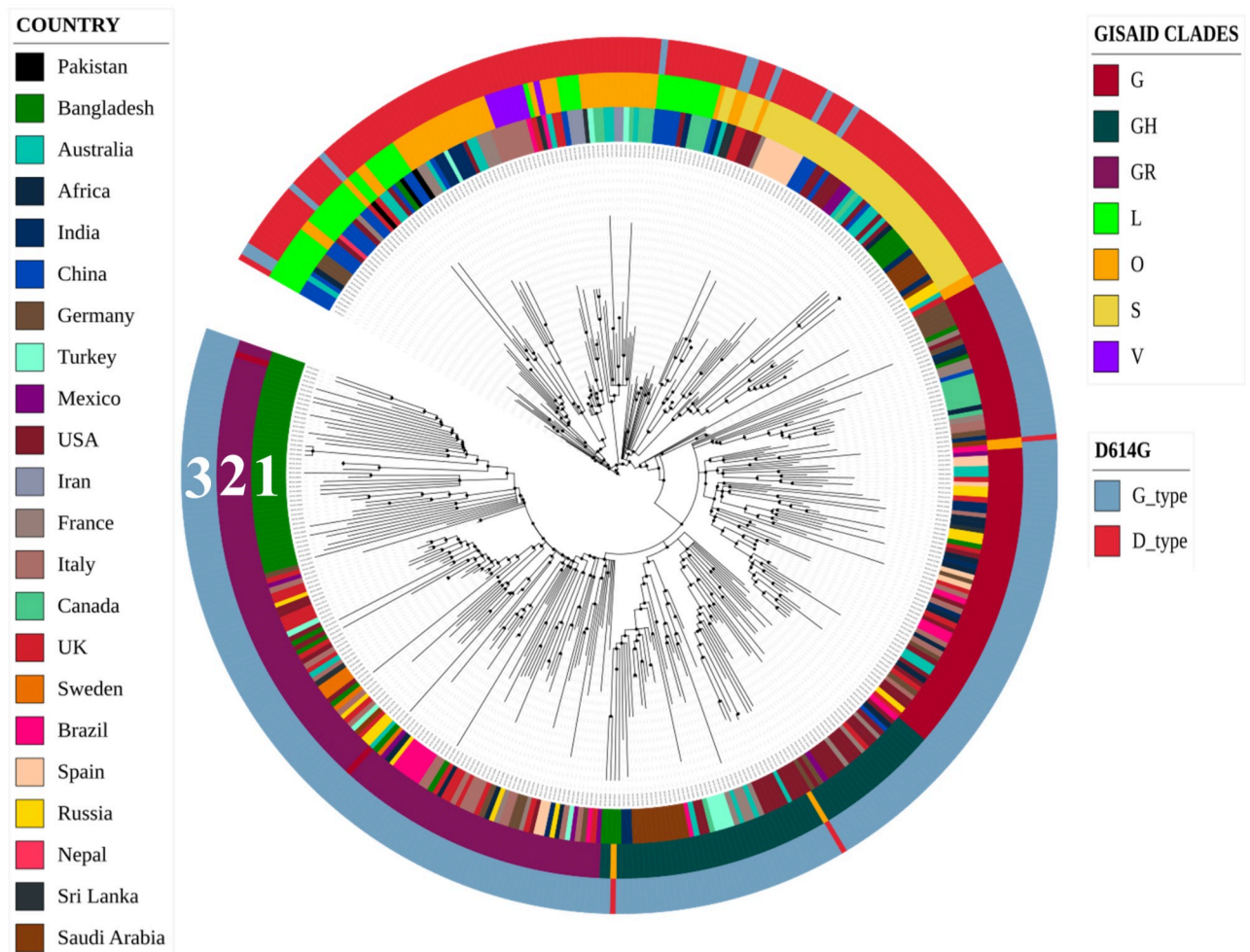


Fig 1. Time-resolved phylogenetic tree of SARS-CoV-2 from different countries. Circle 1 exhibits the location and distribution of genome sequences; circle 2 demonstrates GISAID clades based on some specific mutations, and circle 3 represents the D614G classification.

<https://doi.org/10.1371/journal.pone.0245584.g001>

D614G mutations in spike protein

The SARS-CoV-2 sequences were also categorized according to D614G type mutation (Fig 1). The D614G mutation generates an additional serine protease (Elastase) cleavage site near the S1-S2 junction of the Spike protein [21]. All but one sequence from Dhaka and Narayanganj were found to be of G type which carries Glycine at position 614. On the other hand, 5 out of 11 of the sequences of isolates in Chattogram were G type while the rest carried Aspartate at 614th position of the spike protein (Fig 2). In addition, the first sequence from Bangladesh carried G614 type of surface glycoprotein, which indicate that this dominant variant was present since the first isolation of SARS-CoV-2 in Bangladesh and the mutant virus might have been imported to the country from Europe. Since spike protein having glycine at 614 position is more stable and presumably increase the transmissibility [20, 21], we assume that, introduction of the strain carrying this mutation might have facilitated rapid viral transmission here in Bangladesh. This particular subtype with a non-silent (Aspartate to Glycine) mutation at 614th position of the Spike protein might have also rapidly outcompeted any preexisting subtype here, including the ancestral strain.

Tree scale: 0.1 ⇐

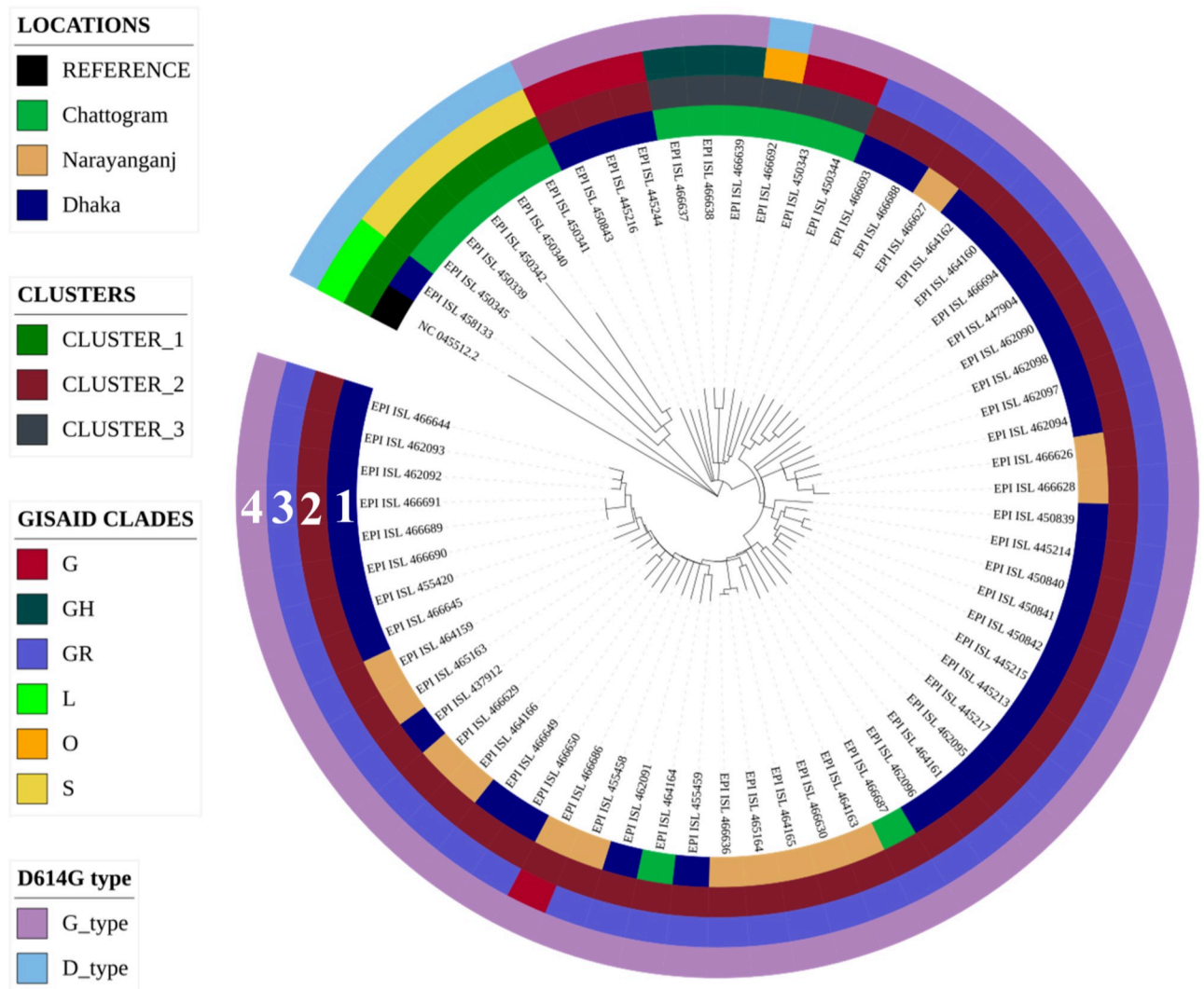


Fig 2. Time-resolved phylogenetic tree of SARS-CoV-2 sequences from Bangladesh. Circle 1 shows the location of isolation of the samples, circle 2 shows the three clusters in which various sequences belong based on their similarity, circle 3 shows GISAID clades of different sequences, and circle 4 represents whether the sequences are D type or G type.

<https://doi.org/10.1371/journal.pone.0245584.g002>

Haplotype analysis of SARS-CoV-2 sequences

Relationships among DNA sequences within a population are often studied by constructing and visualizing a haplotype network. We constructed a haplotype network by the Median Joining algorithm and found that 338 of 434 SARS-CoV-2 sequences from representative countries were alike, therefore formed a large haplo group (Fig 3A). However, there were presences of a significant number of unique lineages too consisting of a single or multiple SARS-CoV-2 sequences (Fig 3A). This network demonstrated the closeness of the sequences and their pattern of mutation beyond the geographical boundary. Several SARS-CoV-2 isolates appeared to have sustained certain common mutations along with certain unique mutations. Although a large proportion of sequences from Bangladesh belonged to the common cluster (Fig 3A),

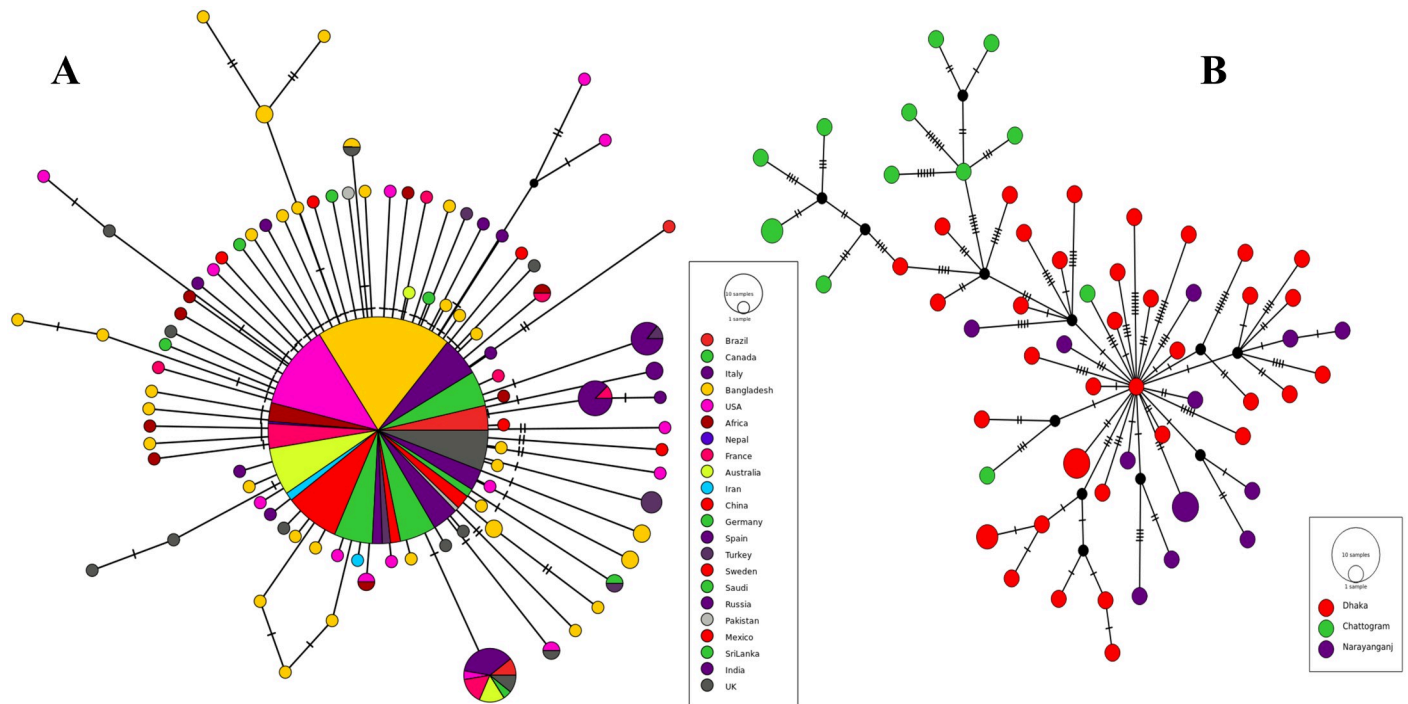


Fig 3. Haplotype network of selected SARS-CoV-2 genome sequences. (A) The largest circle represents a group of sequences from different countries that are similar and few sequences are connected with other sequences through undetermined intermediate due to harboring unique mutations denoted by small black circle. (B) Haplotype network of sequences from Bangladesh; the number of dashes in the connecting line denotes the number of mutations against each other.

<https://doi.org/10.1371/journal.pone.0245584.g003>

there was a significant number of unique nodes as well due to mutations overtime subsequent to being carried into Bangladesh (Fig 3A). Therefore analysis of the sequences from Bangladesh provided further insight of their mutation patterns. The haplotype network revealed that viruses isolated in Bangladesh had certain unique mutations in them, and as a result they belonged to different haplo groups and no significant cig group (Fig 3A). Most of the isolates sustained a significant number of mutations compared with each other. In addition it further confirms that most isolates from Chattogram (Fig 3B) were not directly related to those isolated in Dhaka or Narayanganj.

In consideration of phylogenetic relationship, spike protein mutation at position 614, and haplotype network we predict that SARS-CoV-2 were introduced into the country from different sources. Firstly, the viruses in Dhaka were imported from either of the European countries, including United Kingdom and France, while that in Chattogram was possibly imported from the Middle East, in particular the Kingdom of Saudi Arabia. Later on, Dhaka served as the infection hub inside the country to infect other cities like Chattogram and Narayanganj. Since Dhaka is the capital city, enormous number of people live here and uncontrolled movement of large number of people occurs between other districts and the capital city.

Location and predicted effects of the mutations

We detected the presence of 209 point mutations in 64 SARS-CoV-2 isolates from Bangladesh when compared to the reference sequence from Wuhan, China. In addition, 19 isolates were found to have lost significant portions of their genome. In 4 cases these deletions were associated with loss of non-structural proteins such as ORF7 and ORF8 while other deletions were upstream

or downstream gene variants (S4 Table). Among the point mutations, 29 mutations were in the non-coding region of the genome and 180 were in coding regions. Ten of the 29 non-coding mutations were in upstream non-coding region and rest was in downstream non-coding region of the genome. 70 mutations in the coding region were synonymous and 110 mutations predicted substituted amino acids. Among twelve predicted ORFs, ORF1ab which comprises approximately 67% of the genome encoding 16 nonstructural proteins had more than 60 percent of the total mutations while gene E encoding envelope protein and ORF7b were conserved and did not carry any mutation. Though ORF1 harbored the highest number of mutations, mutation density was highest in ORF10 considering ORF lengths. Details and distribution of the mutations are presented in (Table 1 and Fig 4) and full analysis report is placed in S3 Table.

Among SARS-CoV-2 sequences reported from Bangladesh, 241C>T and 3037C>T changes were the two most abundant mutations found in 58 out of 64 isolates, and were always found simultaneously (Table 2). Position 241 is located in the non-coding region whereas the mutation in position 3037 was synonymous. On the other hand, 57 sequences were found to harbor 14408C>T and 23403A>G mutations which altered amino acid Pro>Leu and Asp>Gly respectively, and these two mutations were found to be present simultaneously as well. In addition, other co-evolving mutations found were (241C>T, 3037C>T, 14408C>T, 23403A>G), (28881G>A, 28882G>A, 28883G>C), (8782C>T, 28144T>C), (4444G>T, 8371G>T, 29403A>G). Besides the highly abundant mutations, several less common and unique mutations were also present in the sequences analyzed.

As a result of continuous mutations genes are seemingly going through natural selection. ORF6 was predicted to have dN/dS value of 17.8 due to the presence of higher number of missense than synonymous mutations. This finding indicates that ORF6 is rapidly evolving and is highly divergent. The ORF6 protein is an accessory protein whose function is yet to be fully elucidated [31].

The ORF7a and Nucleocapsid phosphoprotein (N) had dN/dS values 1.08 and 1.55 respectively which confer their strong evolution to cope up with challenges under positive selection pressure. ORF10 is predicted to be conserved with dN/dS value 0 while envelope protein and ORF7b did not harbor any mutation and was conserved. On the other hand, ORF3a and surface glycoprotein (S) might approach toward positive selection pressure and evolve but the remaining protein coding genes were under negative selection pressure.

Table 1. Distribution of mutations sustained by various isolates of SARS-CoV-2 showing number of mutation in each gene, types of mutation, mutation density and directional selection value.

ORF	ORF length	Total No. of mutation	No. of Synonymous Mutations	No. of Missense Mutations	Mutation density	dN/dS
ORF1ab	21291	110	44	66	0.52%	0.49
ORF1a	13219	84	32	52	0.64%	0.63
S	3823	17	5	12	0.44%	0.79
ORF3a	829	13	4	9	1.57%	0.89
E	229	0	0	0	0.00%	NA
M	670	7	5	2	1.04%	0.23
ORF6	187	3	0	6	1.60%	17.8
ORF7a	367	1	1	0	0.27%	1.08
ORF7b	133	0	0	0	0.00%	NA
ORF8	367	9	5	4	2.45%	0.57
N	1261	17	3	14	1.35%	1.55
ORF10	118	3	3	0	2.54%	0

(dN/dS > 1 is positive selection, dN/dS = 1 neutral selection, dN/dS < 1 negative selection, dN/dS = 0 is conserved region).

<https://doi.org/10.1371/journal.pone.0245584.t001>

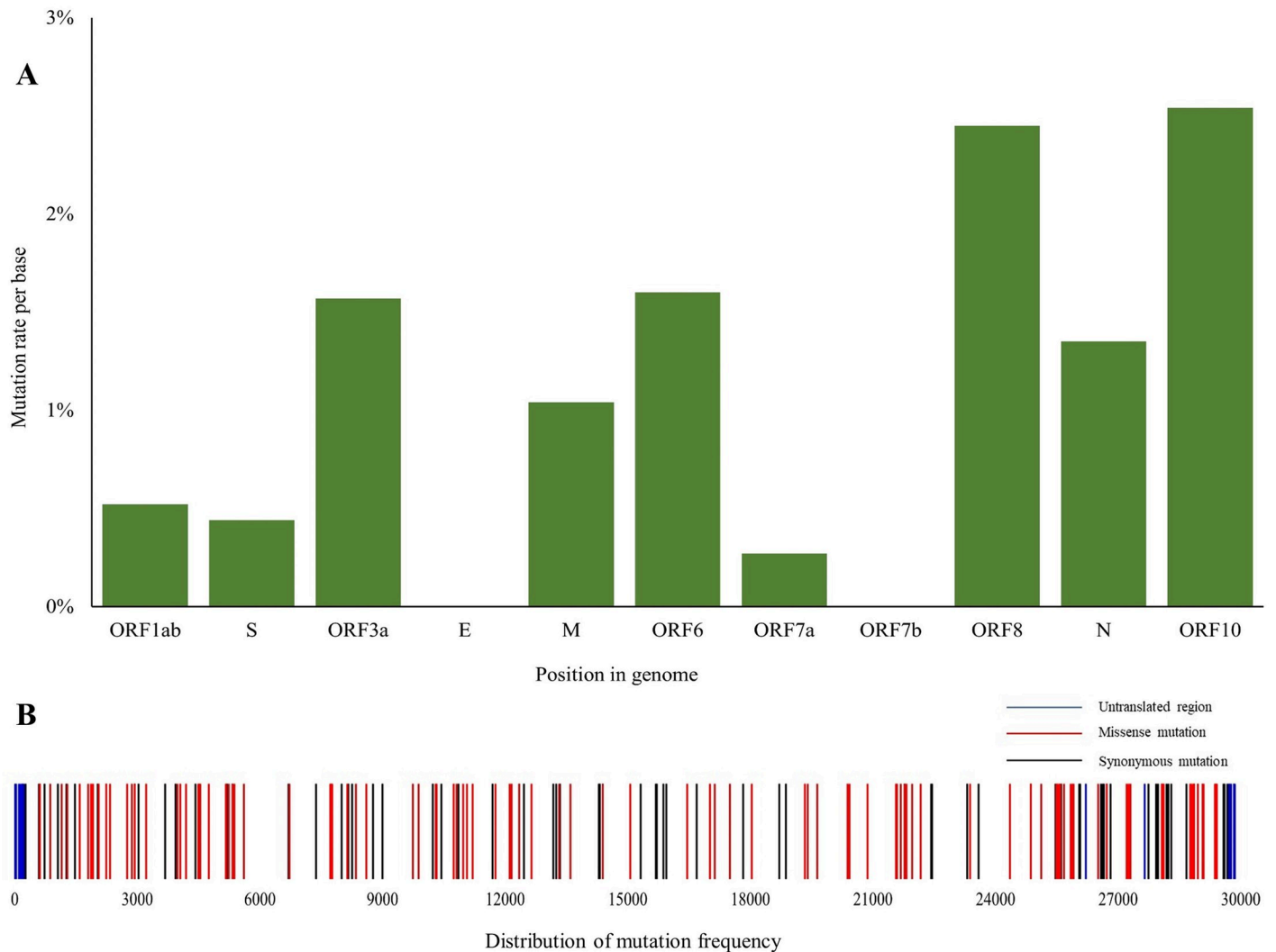


Fig 4. Mutations present in SARS-CoV-2 genome sequences reported from Bangladesh. (A) Mutation rate per base at coding regions, and (B) Frequency of distribution of mutations.

<https://doi.org/10.1371/journal.pone.0245584.g004>

Table 2. High-frequency mutations present in SARS-CoV-2 sequences from Bangladesh, and their predicted effect on the stability of protein structures.

Position	Gene	Feature	Amino acid change	Frequency	ddG (kcal/mol)
241 C > T	5' UTR	Non-coding	-	58/64	
1163 A > T	ORF1ab	missense	I300F	44/64	-0.756
3037 C > T	ORF1ab	Synonymous	-	58/64	
14408 C > T	ORF1ab	Missense	P4715L	57/64	1.463
23403 A > G	S	Missense	D614G	57/64	-0.45
28881 G > A	N	Missense	R203L	49/64	-3.4
28882 G > A	N	Synonymous	-	49/64	
28883 G > C	N	Missense	G204A	49/64	-4.3

(ddG < 0: decrease stability, ddG > 0: increase stability).

<https://doi.org/10.1371/journal.pone.0245584.t002>

Among 110 missense mutations, 99 were predicted to destabilize the corresponding proteins while only 11 mutations predictably increase stability. Ten protein stabilizing mutations were present in ORF1 with 1 mutation in ORF3a. On the other hand, destabilizing mutations were distributed among most of the ORFs (Table 3). None of the mutations in structural proteins were predicted to increase stability.

Taken together these results suggest that the SARS-CoV-2 isolated in Bangladesh has been evolving rapidly, and that 90% of the missense mutations are destabilizing the genes encoding virulence as well as structural proteins. This could be the reason that the infection and mortality rate due to SARS-CoV-2 in Bangladesh is apparently significantly lower than that of the United States of America, as well as European and Latin American countries.

In summary, mutation analysis revealed point mutations as well as deletion of base pairs. Deletions of the base pairs were associated with missing non-structural proteins and predictably affected certain viral properties since ORF7a protein is the growth factor of the coronavirus family, induce apoptosis, and promotes viral encapsulation [32–34] while ORF8 is associated with viral adaptation by playing role in host-virus interaction [35, 36]. Furthermore, we have found that some genes are under positive selection pressure indicating that the virus is fast-evolving presumably to evade host cell's innate immunity; which should be taken into special consideration prior to vaccine development or other treatment strategies.

Finally, a missense mutation at 1163A>T changing the amino acid isoleucine to phenylalanine in Nsp2 protein was found uniquely among 44 isolates in Bangladesh but absent among all the selected sequences from other countries. Nsp2 is a methyltransferase like domain that interacts with PHB and PHB2, and modulates the host cell survival strategy by affecting cellular differentiation, mitochondrial biogenesis, and cell cycle progression to escape from innate immunity [36, 37]. In considering death ratio against the infection rate in Bangladesh in view of the function of this non-structural protein, we assume that this unique mutation might turn out to be vital for the observed low death rate in this country. Therefore possible effects of this missense, point mutation might warrant further studies, and should be monitored closely.

Table 3. Mutations present in different ORFs in the genome sequences of SARS-CoV-2 isolates reported from Bangladesh.

ORF	Unique Mutation in SARS-Cov-2 in Bangladesh	Remarks
ORF1ab	93	A mutation at 1163A>T (Ile300Phe) in ORF1 operon which lower the structural stability of the protein was found existing in 44 strains of Bangladesh but absent in representative countries
S	13	A highly prevalent mutation 23403A>G (Asp614Gly) which ease the viral transmission, dominating throughout the world was found to be present in 57 of 64 sequences in Bangladesh.
ORF3a	9	A mutation at 25563G>T (Gln57His) was found highly prevalent among selected sequences from the USA was also present in Bangladeshi sequences.
M	4	Unique mutations were synonymous
ORF6	3	3 missense mutations were absent in sequences from representative countries.
ORF7a	1	Synonymous mutation
ORF8	7	3 missense mutations but their frequency was lower
N	10	Mutations at 28881G>A (Arg203Lys), 28882G>A (Arg203ARG), 28883G>C (Gly204Arg) were harbored in 49 of 64 sequences which were found highly prevalent among viruses in European and North American countries
ORF10	3	All mutations were synonymous and found this gene conserved

<https://doi.org/10.1371/journal.pone.0245584.t003>

Supporting information

S1 Table. Number of sequences from different countries.

(DOCX)

S2 Table. Sequence list from Bangladesh.

(DOCX)

S3 Table. Sequence list from selected countries.

(DOCX)

S4 Table. Mutations in Bangladeshi sequences. Full mutation analysis report. Mutations' location, type of mutation, their probable effect, global presence scenario.

(XLSX)

Acknowledgments

We acknowledge the contribution of different research groups who generated the primary sequence data which were included in the present analysis. Research groups who sequenced the isolates in Bangladesh include the Child Health Research Foundation, National Institute of Laboratory Medicine and Referral Center, National Institute of Biotechnology, DNA Solution Ltd., COVID-19 Laboratory Centre for Advanced Research in Sciences (CARS), Genomic Research Lab—BCSIRand, and Tejgaon College bmb laboratory.

Author Contributions

Conceptualization: Tushar Ahmed Shishir, Iftekhar Bin Naser, Shah M. Faruque.

Data curation: Iftekhar Bin Naser, Shah M. Faruque.

Formal analysis: Tushar Ahmed Shishir.

Supervision: Iftekhar Bin Naser.

Writing – original draft: Tushar Ahmed Shishir, Shah M. Faruque.

Writing – review & editing: Iftekhar Bin Naser, Shah M. Faruque.

References

1. WHO. Coronavirus disease COVID-2019—Situation Report 169. World Heal Organ. 2020; <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports>
2. Lefkowitz EJ, Dempsey DM, Hendrickson RC, Orton RJ, Siddell SG, Smith DB. Virus taxonomy: The database of the International Committee on Taxonomy of Viruses (ICTV). *Nucleic Acids Res.* 2018; 46(D1):D708–17. <https://doi.org/10.1093/nar/gkx932> PMID: 29040670
3. Zhong NS, Zheng BJ, Li YM, Poon LLM, Xie ZH, Chan KH, et al. Epidemiology and cause of severe acute respiratory syndrome (SARS) in Guangdong, People's Republic of China, in February, 2003. *Lancet.* 2003; 362(9393):1353–1358. [https://doi.org/10.1016/s0140-6736\(03\)14630-2](https://doi.org/10.1016/s0140-6736(03)14630-2) PMID: 14585636
4. De Wit E, Van Doremalen N, Falzarano D, Munster VJ. SARS and MERS: recent insights into emerging coronaviruses. *Nature Reviews Microbiology.* 2016; 14(8):523–534. <https://doi.org/10.1038/nrmicro.2016.81> PMID: 27344959
5. Yin Y, Wunderink R. MERS, SARS and other coronaviruses as causes of pneumonia. *Respirology.* 2017; 23(2):130–137. <https://doi.org/10.1111/resp.13196> PMID: 29052924
6. Webmeter. Coronavirus Age, Sex, Demographics (COVID-19)—Worldometer. 2020. www.worldometers.info
7. Rahaman Khan MH, Hossain A. COVID-19 Outbreak Situations in Bangladesh: An Empirical Analysis. *medRxiv.* 2020. Forthcoming

8. Cui J, Li F, Shi Z. Origin and evolution of pathogenic coronaviruses. *Nature Reviews Microbiology*. 2019; 17(3):181–192. <https://doi.org/10.1038/s41579-018-0118-9> PMID: 30531947
9. Denison MR, Graham RL, Donaldson EF, Eckerle LD, Baric RS. Coronaviruses: An RNA proofreading machine regulates replication fidelity and diversity. *RNA Biology*. 2011; 8(2):270–279. <https://doi.org/10.4161/rna.8.2.15013> PMID: 21593585
10. Yang Y, Liu C, Du L, Jiang S, Shi Z, Baric RS, et al. Two Mutations Were Critical for Bat-to-Human Transmission of Middle East Respiratory Syndrome Coronavirus. *J Virol*. 2015; 89(17):9119–9123. <https://doi.org/10.1128/JVI.01279-15> PMID: 26063432
11. Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. The proximal origin of SARS-CoV-2. *Nature Medicine*. 2020; 89(17):9119–9123. <https://doi.org/10.1038/s41591-020-0820-9> PMID: 32284615
12. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data—from vision to reality. *Euro-surveillance*. 2017; 22(13).
13. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, et al. A new coronavirus associated with human respiratory disease in China. *Nature*. 2020; 579(7798):265–269. <https://doi.org/10.1038/s41586-020-2008-3> PMID: 32015508
14. Shean RC, Makhsous N, Stoddard GD, Lin MJ, Greninger AL. VAPiD: A lightweight cross-platform viral annotation pipeline and identification tool to facilitate virus genome submissions to NCBI GenBank. *BMC Bioinformatics*. 2019; 20(1). <https://doi.org/10.1186/s12859-019-2606-y> PMID: 30674273
15. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol Biol Evol*. 2013; 30(4):772–780. <https://doi.org/10.1093/molbev/mst010> PMID: 23329690
16. Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol*. 2015; 32(1):268–274. <https://doi.org/10.1093/molbev/msu300> PMID: 25371430
17. Sagulenko P, Puller V, Neher RA. TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evol*. 2018; 4(1). <https://doi.org/10.1093/ve/vex042> PMID: 29340210
18. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res*. 2019; 47(W1):W256–W259. <https://doi.org/10.1093/nar/gkz239> PMID: 30931475
19. GISAID. Clade and lineage nomenclature aids in genomic epidemiology studies of active hCoV-19 viruses [Internet]. 2020 [cited 2020 Jun 21]. <https://www.gisaid.org/references/statements-clarifications/clade-and-lineage-nomenclature-aids-in-genomic-epidemiology-of-active-hCoV-19-viruses/>
20. Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, et al. Tracking changes in SARS-CoV-2 Spike: evidence that D614G increases infectivity of the COVID-19 virus. *Cell*. 2020. <https://doi.org/10.1016/j.cell.2020.06.043> PMID: 32697968
21. Zhang L, Jackson CB, Mou H, Ojha A, Rangarajan ES, Izard T, et al. The D614G mutation in the SARS-CoV-2 spike protein reduces S1 shedding and increases infectivity. *bioRxiv*. 2020; Forthcoming
22. Balaban M, Moshiri N, Mai U, Jia X, Mirarab S. TreeCluster: Clustering biological sequences using phylogenetic trees. *PLoS One*. 2019; 14(8):e0221068. <https://doi.org/10.1371/journal.pone.0221068> PMID: 31437182
23. Li H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*. 2018; 34(18):3094–3100. <https://doi.org/10.1093/bioinformatics/bty191> PMID: 29750242
24. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009; 25(16):2078–2079. <https://doi.org/10.1093/bioinformatics/btp352> PMID: 19505943
25. Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, et al. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb genomics*. 2016; 2(4). <https://doi.org/10.1099/mgen.0.000056> PMID: 28348851
26. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)*. 2012 Apr–Jun; 6(2):80–92. <https://doi.org/10.4161/fly.19695> PMID: 22728672
27. Leigh JW, Bryant D. POPART: Full-feature software for haplotype network construction. *Methods Ecol Evol*. 2015; 6(9):1110–1116.
28. Kosakovskiy P, Frost SDW. Not so different after all: A comparison of methods for detecting amino acid sites under selection. *Mol Biol Evol*. 2005; 22(5):1208–1222. <https://doi.org/10.1093/molbev/msi105> PMID: 15703242

29. Kosakovsky Pond SL, Frost SDW. Datamonkey: Rapid detection of selective pressure on individual sites of codon alignments. *Bioinformatics*. 2005; 21(10):2531–2533. <https://doi.org/10.1093/bioinformatics/bti320> PMID: 15713735
30. Cao H, Wang J, He L, Qi Y, Zhang JZ. DeepDDG: Predicting the Stability Change of Protein Point Mutations Using Neural Networks. *J Chem Inf Model*. 2019; 59(4):1508–1514. <https://doi.org/10.1021/acs.jcim.8b00697> PMID: 30759982
31. Gunalan V., Mirazimi A. & Tan Y. A putative diacidic motif in the SARS-CoV ORF6 protein influences its subcellular localization and suppression of expression of co-transfected expression constructs. *BMC Res Notes*. 2011; <https://doi.org/10.1186/1756-0500-4-446> PMID: 22026976
32. Hänel K, Willbold D. SARS-CoV accessory protein 7a directly interacts with human LFA-1. *Biol Chem*. 2007; 388(12). <https://doi.org/10.1515/BC.2007.157> PMID: 18020948
33. Schaecher SR, Touchette E, Schriewer J, Buller RM, Pekosz A. Severe Acute Respiratory Syndrome Coronavirus Gene 7 Products Contribute to Virus-Induced Apoptosis. *J Virol*. 2007; 81(20):11054–11068. <https://doi.org/10.1128/JVI.01266-07> PMID: 17686858
34. Taylor JK, Coleman CM, Postel S, Sisk JM, Bernbaum JG, Venkataraman T, et al. Severe Acute Respiratory Syndrome Coronavirus ORF7a Inhibits Bone Marrow Stromal Antigen 2 Virion Tethering through a Novel Mechanism of Glycosylation Interference. *J Virol*. 2015; 89(23):11820–11833. <https://doi.org/10.1128/JVI.02274-15> PMID: 26378163
35. Zhang Y, Zhang J, Chen Y, Luo B, Yuan Y, Huang F, et al. The ORF8 Protein of SARS-CoV-2 Mediates Immune Evasion through Potently Downregulating MHC-I. *bioRxiv*. 2020; Forthcoming
36. Yoshimoto FK. The Proteins of Severe Acute Respiratory Syndrome Coronavirus-2 (SARS CoV-2 or n-COV19), the Cause of COVID-19. *Protein J [Internet]*. 2020; 39(3):198–216. <https://doi.org/10.1007/s10930-020-09901-4> PMID: 32447571
37. Angeletti S, Benvenuto D, Bianchi M, Giovanetti M, Pascarella S, Ciccozzi M. COVID-2019: The role of the nsp2 and nsp3 in its pathogenesis. *J Med Virol*. 2020; 92(6):584–588. <https://doi.org/10.1002/jmv.25719> PMID: 32083328