Two new mutations: Delineating the mystery of SARS-CoV-2 genome isolated from a Bangladeshi person

Dr. ABMM Khademul Islam, Associate Professor,
Genetic Engineering and Biotechnology, University of Dhaka

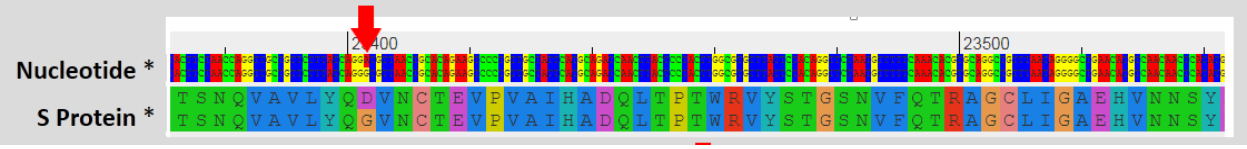
SARS-CoV-2 has so far infected more than 41,00,000 people in 187 countries and claimed over 2,85,000 deaths but no drug or vaccine is yet available. In Bangladesh over 16,500 people got infected and out of which 250 died. Lock-down can provide temporary solution but we need a sustainable solution for this. Although there are three (A,B,C) SARS-CoV-2 variants but still we don't know which one is prevailing in our country, how and through which route it has been transmitted here; if it has acquired any mutations by now and how deadly has it become. Also, we do not know why some people are affected more, showing serious symptoms while others remain asymptomatic. We do not know clearly why and how this coronavirus created havoc in some countries whereas others are mildly affected. In the modern era, problems in biological sciences are tackled by a bottom up approach – where we do genome sequence of the relevant organism and associate it with other metadata to address the problem and find solutions. For the same reason so far 80 countries have deposited more than 24,000 genome sequences of this virus, which includes even countries like Nepal and Vietnam where the coronavirus problem is comparatively less severe. Since the first the first cases were reported on 7 March 2020 by the country's epidemiology institute IEDCR, we have been repeatedly advocating the need of genome sequencing of this virus. We also ensured that Bangladesh has made substantial advancement in science and technology, especially with the special attention of the Prim-minister of the country in this sector, now we are able to do genome sequencing by Next Generation Sequencing (NGS) in our country. There are some institutes and private organizations where NGS machines are available and virus genome sequencing can be done, and we also have expert and experienced Bioinfomaticians who can perform complete genome sequence analysis. However, finally the icebreaking work has been done by the Child Health Research Foundation (CHRF). Dr. Senjuti Shaha, Dr. Samir Kumar Shaha along with their team from CHRF collected samples from a coronavirus infected 22 years old female person and arrange to do whole genome sequencing of the virus using Illumina iSeq 100 NGS platform. As soon as the news of deposition of genome sequence data become available on 12th May, Tuesday afternoon, we sought to extract this sequence and information form the public repository GISAID, CNCB, and started to explore it. Lead by me, at the Department of Genetic Engineering and Biotechnology, University of Dhaka, the Epigenetic and Bioinformatics team on nCoV research has done basic analysis of the genome. My team member Mr. Abdullah Al Kamran Khan was with me in this analysis. We compared the sequence with that of first reported coronavirus genome sequence from Wuhan, China – which is globally considered as 'reference'. Strikingly, we have found that this genome is very similar (99.7% similarity) to that of reference SARS-CoV-2 isolated from Wuhan. There are changes only in 9 place and these changes are single nucleotide change (SNP). No deletion or insertion/addition of any large sequence compared to the original reference.

However, to our surprise we found that this genome has acquired two new mutations which have not been seen among the viruses reported so far. Hence, we observed it closely. At position 1163 (genes orf1ab) a new mutation from A to T has been detected. Previously at the same position

nucleotide A to C in one virus, and nucleotide A to G change in another genome reported. Also, there is a brand new mutation at 17019 position was detected in our Bangladeshi isolated virus which has not been reported so far. This means that these are the new changes that the virus has acquired after entering in Bangladesh. Out of 9, other 7 mutations are very common in sequenced viruses so far. Further studies are required to know what trouble or benefit these new mutations have brought to us.

Position	Reference base	Mutated base	Mutation type	Protein: amino acid change	Mutation Frequency (10984)
241	C	T	Upstream gene variant	Non-coding	7080
1163	A	T	Missense	Orf1ab: 300I>F	1
3037	C	T	Synonymous	Orf1ab: No change	7104
14408	C	T	Missense	Orf1ab: 4715P>L	7120
17019	G	T	Missense	Orf1ab: 5585E>D	1
23403	A	G	Missense	S: 614D>G	7145
28881	G	A	Missense	N: 203R>K	1735
28882	G	A	Synonymous	N: No change	1731
28883	G	C	Missense	N: 204G>R	1730

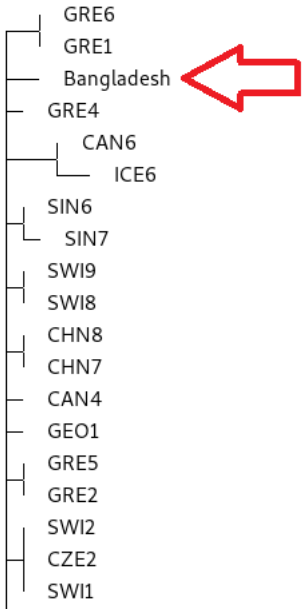
Interestingly, of these 9 mutations, it contains a mutation (Single Nucleotide Mutation or SNP) in its Spike protein. There is non-silent (non-synonymous), amino acid changing (Aspartate to Glycine) mutation at the 614th position of the Spike protein (D614G). This is of particular interest because probably due to this mutation it could quickly spread in the European and American population and out competed the original virus from China. This creates an additional serine protease (Elastase) cleavage site near the Open Reading Frame (ORF) S1 and S2 junction of the Spike protein.



The interesting aspect is that in human, a single nucleotide mutation (deletion of C nucleotide, delC) (rs35074065 variant site) in the TMPRSS2 receptor gene, facilitates the entry of SARS-CoV-2 with D614G mutation to the cell very effectively. Dr Hemayet Ullah from Howard University, USA, also informed us that this delC mutation is very common in the American and European population but very rare in the East Asian/Asian population - hence the change of amino acid aspartic acid to glycine in the S protein of the virus may be helpful for Asian countries but more infective in the American and European population. We do see a less severe effect in Asian countries compare to that in the European and American areas. Any deleterious mutation from the perspective of an organism gets

lost by natural selection and hope that later in time more virulent mutation does not appear in Asian countries. Several research papers are also available on this mutation.

To understand the origin, we have constructed phylogenetic tree (UPGMA and Neighbor-Joining) in MEGA with default parameters, with representative sequences from 60 other countries and the reference sequence, totaling 350 sequences. Phylogenetic tree shows that this Bangladeshi SARS-CoV-2 genome isolate seems more closer to European cluster, most likely the person gest infected from someone who returned from European country or maybe she herself returned from any European country. We are fine tuning the phylogenetic tree, and also in process of making phylogenetic tree with selected high quality 10,000 sequences from 80 countries to better explain the origin and route of transmission of this particular virus.



It is imperative to understand that to understand the pattern of infection in Bangladesh, only one genome sequence is not enough. We need sequence of at least 100 isolates. We have made a proposal to ICT ministry in response to their “Call for Nation (Hakathon)”. In this study proposal we aim to create a dataset by combining 100 coronavirus genomes from Bangladeshi patients and integrate this genome information with patient’s personal/clinical/treatment/diagnostic etc. information. This information will be analyzed extensively by computational methods to do clustering, phylogenetic and pharmacogenomics studies, and will compare data with other world-wide available data to make a concrete information-base that will help pharmaceutical industries to produce appropriate drugs and vaccine for our population. Also, the ICT ministry will be able to announce that Bangladesh has uncovered the genome mystery of the coronavirus circulating in Bangladesh and trace back the transmission. This project will be a multicenter research where essential help from ICT/Bangladesh Govt., and help of IEDCR through Bangladesh Govt. will be required to get patients’ samples and relevant clinical data. We will carry out sequencing (Next Generation Sequencing) of the viral genome and other analyses with our own resources in Bangladesh. If ICT/Govt. support us, it is also possible to do further research in future where in

addition to the viral genome we can sequence genome of some individuals who were infected and developed the disease as well as healthy individuals who did not develop the disease. This may also let us know the factors (if any) that conferred resistance to them. Our team consists of relevant experts who are well experienced in doing similar projects at home and abroad, also all members have their own young, energetic and well-trained working group.