 doi:10.20944/preprints202007.0251.v1

RNA-dependent RNA polymerase and spike protein mutant variants of SARS-CoV-2 predominate in severely affected COVID-19 patients

Subrata K Biswas[a*], Sonchita R Mudi[b]

[a]Department of Biochemistry and Molecular Biology, Bangabandhu Sheikh Mujib Medical University (BSMMU), Dhaka, Bangladesh; [b]Department of Biochemistry, Kumudini Women's Medical College, Mirzapur, Tangail, Bangladesh

[*]**Corresponding author**

Subrata Kumar Biswas, MBBS, MD, PhD

Associate Professor, Department of Biochemistry and Molecular Biology

Bangabandhu Sheikh Mujib Medical University (BSMMU)

Shahbag, Dhaka – 1000, Bangladesh

Phone: +880-1715-966663, Fax: +880-2-8631179

Email: su.biswas@yahoo.com; su.biswas@bsmmu.edu.bd

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 July 2020 doi:10.20944/preprints202007.0251.v1

ABSTRACT

The severity of coronavirus disease 2019 (COVID-19), caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), greatly varies from patient to patient. In the present study, we explored and compared mutation profiles of SARS-CoV-2 isolated from mildly affected and severely affected COVID-19 patients in order to explore any relationship between mutation profile and disease severity. Genomic sequences of SARS-CoV-2 were downloaded from GISAID database. With the help of Genome Detective Coronavirus Typing Tool, genomic sequences were aligned with the Wuhan seafood market pneumonia virus reference sequence and all the mutations were identified. Distribution of mutant variants was then compared between mildly and severely affected groups. Among the numerous mutations detected, 14,408C>T and 23,403A>G mutations resulting in RNA-dependent RNA polymerase (RdRp) P323L and spike protein D614G mutations, respectively, were found predominantly in severely affected group (>82%) compared with mildly affected group (<46%, p<0.001). The 241C>T mutation in the non-coding region of the genome was also found predominantly in severely affected group. The 3,037C>T, a silent mutation, also appeared in relatively high frequency in severely affected group. We concluded that RdRp P323L and spike protein D614G mutations predominate in severely affected COVID-19 patients. Further studies will be required to explore whether these mutations have any impact on the severity of COVID-19.

 doi:10.20944/preprints202007.0251.v1

Introduction

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the viral pathogen that causes coronavirus disease 2019 (COVID-19), has infected millions of people worldwide just in six months [1]. Although majority of the SARS-CoV-2 infected individuals recover after developing mild to moderate symptoms, more than 500,000 people have already been died due to severe form of COVID-19 [1]. Severity of COVID-19 has been found to greatly vary from patient to patient, but it is so far not entirely clear what is responsible for the variable severity of COVID-19 in the population [1-2]. We recently hypothesized that genetic variation in SARS-CoV-2 may explain the variable severity of COVID-19 [2]. To the best of our knowledge, this hypothesis has not yet been accepted or rejected by scientific investigations, but some studies suggested that D614G mutation in the spike protein may contribute to increased infectivity or transmissibility of SARS-CoV-2 leading to increased severity of COVID-19 [3-4]. In fact, spike protein mediates viral entry into the host cell through binding of the virus with host cell receptor angiotensin-converting enzyme-2 (ACE2) [5]. The D614G mutation of the spike protein was observed sometimes in late January 2020 both in Europe and in China, but then this mutation spread first in the Europe and gradually globally [3-4]. Thus the distribution of spike protein D614G mutation has temporal and geographical variation. In the present study, we explored and compared mutation profiles of SARS-CoV-2 isolated from mildly affected and severely affected COVID-19 patients in order to explore relationship between mutation profile and disease severity.

Methods

Genomic sequences of SARS-CoV-2 were downloaded from GISAID database (https://www.gisaid.org). Although there were more than 45,000 SARS-CoV-2 genomic

 doi:10.20944/preprints202007.0251.v1

sequences deposited in the GISAID website by June 12, 2020, only 2,443 complete (>29,000 base pair length) sequences had 'patient status' information available. Out of 2,443 sequences we manually searched and selected those sequences for which patient status information clearly indicated either mild or severe disease. In this way, we selected 102 sequences to be included in the present study: 46 in the mildly affected group with patient status 'mild' / 'asymptomatic' / 'not hospitalized', and 56 in the severely affected group with patient status 'severe' / 'ICU' / 'deceased'. Sequences with patient status described with ambiguous words like 'released', 'hospitalized', 'alive', 'live', 'unknown' etc. were excluded due to uncertainty whether the patients were mildly or severely affected. Mutation profile was determined using the Genome Detective Coronavirus Typing Tool (available at https://www.genomedetective.com/app/typingtool/cov), a web-based bioinformatics pipeline that can accurately identify changes at nucleotides, coding regions and proteins using a novel dynamic aligner to allow tracking new viral mutations [6]. With the help of this coronavirus typing tool, each of the 102 SARS-CoV-2 sequences was aligned with the Wuhan seafood market pneumonia virus reference sequence NC_045512.3, and all the nucleotide and amino acid sequence variations were identified comparing with the reference sequence. Each of the mutations was counted for mildly and severely affected groups, and expressed in number and percentage. Distribution of selected mutant variants was compared between the mildly affected and severely affected groups by chi-squared test using SPSS version 21.0. A p value of less than 0.01 was considered statistically significant.

Results

The SARS-CoV-2 genomic sequences that we included in the present study were sequenced from viral isolates collected in the USA (number; mild, severe: 3, 3), Mexico (0, 3), Brazil (4,

 doi:10.20944/preprints202007.0251.v1

1), Austria (0, 15), Russia (0, 13), Belgium (5, 8), Hungary (1, 0), Spain (1, 3), Turkey (1, 0), Bosnia and Herzegovina (0, 1), India (21, 6), Sri Lanka (0, 1), Japan (3, 0), Indonesia (0, 1), Lebanon (0, 1), Kuwait (1, 0) and Nigeria (6, 0). The viral isolates were collected between February 03 and May 27, 2020. There were 29 men, 13 women and 4 with gender information unavailable in the mildly affected group (n=46), and 31 men and 25 women in the severely affected group (n=56). Age distribution was 17 to 98 years for mildly affected group and 17 to 93 years for severely affected group.

A comprehensive list of all mutations compared to the Wuhan reference sequence NC_045512.3 in mildly affected and severely affected groups is presented in Supplementary Table 1. In the coding region, there were 103 mutations in the mildly affected group with 37 silent and 66 missense mutations. In the severely affected group, there were 111 mutations with 40 silent and 71 missense mutations. In the non-coding region, there were 2 and 8 mutations in the mildly affected and severely affected groups, respectively, in the 5´ un-translated region (UTR); whereas, there were 9 and 15 mutations in the 3´ UTR in the mildly affected and severely affected groups, respectively. However, majority of the mutations appeared in low frequency, i.e., the mutations were found only in a few cases of mildly and severely affected groups (Supplementary Table 1), and therefore, those mutations were unlikely to be related to the severity of COVID-19.

Any mutation with a frequency of 5 or more in either mildly affected or severely affected groups are presented in Table 1. In the open reading frame (ORF)1ab of the SARS-CoV-2 genome, the most frequent mutation identified was 14,408C>T at the nucleotide level. This mutation results in a missense mutation P4715L at the amino acid level of ORF1ab

 doi:10.20944/preprints202007.0251.v1

polyprotein that ultimately appears as P323L mutation in the RNA-dependent RNA polymerase (RdRp) enzyme. This mutation was predominantly occurred in severely affected group (82.1%) compared with mildly affected group (45.7%; p<0.001) (Table 1). In ORF1ab, 11,083G>T mutation at the nucleotide level caused L3606F mutation at the amino acid level and involved non-structural protein (nsp) 6. This mutation was found mainly in the mildly affected group (28.3%) compared with severely affected group (1.8%, p<0.001). Another mutation 5,700C>A in ORF1ab caused missense mutation A1812D at the amino acid level. This mutation involved nsp3 and was found in 15.2% of mildly affected group but not found in severely affected group. Among the silent mutations present in ORF1ab, 3,037C>T mutation was found more commonly in severely affected group (64.3%) compared with mildly affected group (45.7%, p=0.06). Other silent mutations in ORF1ab appeared in relatively low frequency in mildly and severely affected groups (Table 1).

In the spike protein, 23,403A>G mutation at the nucleotide level resulted in D614G mutation at the amino acid level, and it was predominantly found in severely affected group (85.7%) compared with mildly affected group (45.7%, p<0.001) (Table 1). In ORF3a, 25,563G>T mutation at the nucleotide level resulted in Q57H mutation at the amino acid level, and this mutation was more prevalent in severely affected group (26.8%) compared with mildly affected group (10.9%, p=0.08). In ORF8, 28,144T>C mutation at the nucleotide level resulted in L84S mutation at the amino acid level. This mutation was found in 26.1% of mildly affected group and in 5.4% of severely affected group (p=0.008). Other mutations affecting spike, membrane and nucleocapsid proteins of SARS-CoV-2 genome appeared in low frequency (Table 1).

 doi:10.20944/preprints202007.0251.v1

Among all the mutations in the non-coding region of SARS-CoV-2 genome, the 241C>T mutation in the 5´ UTR appeared most predominantly in severely affected group (85.7%) compared with mildly affected group (45.7%, $p<0.001$) (Table 1). The 29,742G>A, 29,827A>T and 29,830G>T mutations in the 3´ UTR appeared at a frequency of 17.4, 34.8 and 43.5%, respectively, in mildly affected group; but none of these mutations was found in the severely affected group (Table 1).

Of note, in the mildly affected group, four most common mutations (241C>T, 3,037C>T, 14,408C>T and 23,403A>G) coincided. In the severely affected group, however, 241C>T and 23,403A>G coincided, and 3,037C>T and 14,408C>T occurred in subsets of them. Among these four predominant mutations, 14,408C>T and 23,403A>G are the most important ones as they alter amino acid sequence in the RdRp and spike protein, respectively. There was temporal and geographical variation in the distribution of 23,403A>G mutation that caused D614G mutation in the spike protein of SARS-CoV-2 [3-4]. In Table 2, we showed collection period of viral isolates in month and frequency of D614G mutation in mildly affected and severely affected groups. Although the percentage of D614G mutation gradually increased from February towards May in both groups, there was more D614G mutation in severely affected group compared with mildly affected group in March and April when most of the viral isolates were collected for sequencing (Table 2). This finding suggests that increased spike protein D614G mutation in severely affected group was unlikely to be due to temporal variation in the distribution of the mutation. However, in this study, majority of samples of mildly affected group were from India whereas those of severely affected group were from Europe, as described above. To explore whether this fact contributed to the increased spike protein D614G mutation in severely affected group, we showed frequency of D614G mutation in mildly affected and severely affected groups for India and Belgium in

 doi:10.20944/preprints202007.0251.v1

Table 3. Of note, the percentage of D614G mutation was found higher in severely affected group compared with mildly affected group for both India and Belgium (Table 3). But such comparisons were not possible for other countries included in this study because very small number of samples was found in either mildly affected or severely affected group for countries other than India and Belgium, as described above.

Discussion

The severity of COVID-19 greatly varies from patient to patient. Majority of the patients either remain asymptomatic or develop mild to moderate symptoms. However, some COVID-19 patients who develop severe disease die even after hospitalization and intensive care [1]. Why the disease severity differs so much from one person to another is one of the mysteries scientists are still trying to solve [1]. The present study was designed to explore whether genetic variation in SARS-CoV-2 may explain variable severity of COVID-19. Mutation profiles of SARS-CoV-2 isolated from mildly affected and severely affected COVID-19 patients were explored and compared. Among numerous mutations observed in this study, two missense mutations, 14,408C>T and 23,403A>G, affecting RdRp and spike protein genes, respectively, were found most predominantly in the severely affected group compared with mildly affected group. Along with these two mutations, 241C>T in the 5´ UTR and a silent mutation 3,037C>T in the ORF1ab were predominantly found in severely affected group (the later not significantly), however, these mutations do not alter amino acid sequence in a protein. Many other mutations that were found in low frequency in the present study are unlikely to exert an effect on the severity of COVID-19. Thus the ability of RdRp and spike protein mutations on the severity of COVID-19 needs to be considered.

 doi:10.20944/preprints202007.0251.v1

The spike protein of SARS-CoV-2 is responsible for binding with host cell receptor ACE2, and thus it allows entry of the virus into the host cell [5]. In fact, the spike protein of SARS-CoV-2 has 10 to 20 folds higher affinity for ACE2 receptor than the corresponding spike protein of SARS-CoV [7]. Thus the spike protein is potentially related to the infectivity of SARS-CoV-2. The 23,403A>G mutation in the genome of SARS-CoV-2 causes replacement of aspartic acid with glycine at position 614 (D614G) of the spike protein. This D614G spike protein mutation appeared sometimes in late January 2020 and then it has spread initially in Europe and then all over the world [3-4]. Several ways have been proposed through which spike protein D614G mutation may increase the infectivity of SARS-CoV-2 [3]. However, computer-based structural analysis of spike protein with D614G mutation suggested that the mutation is unlikely to alter its interaction with human ACE2 receptor [4]. But Korber et al. [3] found that patients infected with spike protein D614G mutant form of SARS-CoV-2 had higher viral loads since fewer PCR cycles were needed for their diagnosis. Furthermore, in cell culture experiment, spike protein D614G mutation was found to infect ACE2 expressing cells more efficiently, and this increased infectivity was found to correlate with less shedding of S1 domain of spike protein and more incorporation of spike protein in the virion [8]. In spite of these facts, previous studies were unable to explore an association between the spike protein D614G mutation and disease severity due to relative lack of clinical data of the patients included in their studies [3-4].

As we found in the present study, previous studies also identified that the spike protein D614G mutation frequently accompanies a silent mutation 3,037C>T and a missense mutation 14,408C>T in ORF1ab [3]. The 14,408C>T mutation in ORF1ab replaces a proline with leucine at position 4715 (P4715L) of ORF1ab polyprotein which actually appears as a replacement of proline with leucine at position 323 (P323L) of RdRp enzyme. The RdRp

 doi:10.20944/preprints202007.0251.v1

enzyme of SARS-CoV-2 catalyzes replication of viral RNA and it possesses proof-reading capability (9). Thus a critical mutation in RdRp gene has the potential to alter viral replication capability with fidelity, and thereby a mutation in RdRp may contribute to infectivity of the virus and severity of the disease. The presence of 14,408C>T mutation in SARS-CoV-2 genome that causes RdRp P323L mutation was found to be associated with overall increase in mutation rate in the viral genome [9]. Although the 14,408C>T (RdRp P323L) mutation was predominantly found in severely affected patients in the present study, further studies will be required to elucidate whether this RdRp mutation has any significant impact on the viability and infectivity of SARS-CoV-2 and severity of COVID-19.

The vast majority of genomic sequences of SARS-CoV-2 available at GISAID database do not contain patient status information. Even many of the sequences that contain patient status information use such ambiguous words to describe the information that do not reflect the severity status of the patient. That's why we were unable to include large number of sequences to analyze in the present study. For the same reason, we were unable to include sequences obtained from a defined geographical location and within a defined time period. Thus temporal and geographical variation in the distribution of mutation may have some influence on our findings. Although temporal effect was excluded, geographical effect cannot be completely excluded in the present study. In spite of these limitations, it may be concluded that the spike protein D614G mutation and RdRp P323L mutation predominate in severely affected COVID-19 patients. Further studies will be required to explore whether RdRp P323L mutation or spike protein D614G mutation or the combination of both mutations can exert an impact on the severity of COVID-19.

Acknowledgement

We gratefully acknowledge the authors, originating and submitting laboratories of the sequences

from GenBank and GISAID's hCoV-19 Database on which this research is based.

Competing interests

The authors claim no competing interest.

Data availability

The mutation profiles of the genomic sequences that support the findings of this study are available from the corresponding author upon reasonable request.

References

1. Callaway E, Ledford H, Mallapaty S. Six months of coronavirus: the mysteries scientists are still racing to solve. *Nature*. 2020; 583: 178-179. doi: 10.1038/d41586-020-01989-z
2. Biswas SK, Mudi SR. Genetic variation in SARS-CoV-2 may explain variable severity of COVID-19. *Med Hypotheses.* 2020; 143:109877. doi: 10.1016/j.mehy.2020.109877
3. Korber B, Fischer WM, Gnanakaran S, et al. Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. *bioRxiv*. 2020. doi: https://doi.org/10.1101/2020.04.29.069054
4. Isabel S, Grana-Miraglia L, Gutierrez JM, et al. Evolutionary and structural analyses of SARS-CoV-2 D614G spike protein mutation now documented worldwide. *bioRxiv*. 2020. doi: https://doi.org/10.1101/2020.06.08.140459
5. Hoffmann M, Kleine-Weber H, Schroeder S, et al. SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell.* 2020;181:271-280.e8. doi: 10.1016/j.cell.2020.02.052.
6. Cleemput S, Dumon W, Fonseca V, et al. Genome Detective Coronavirus Typing Tool for rapid identification and characterization of novel coronavirus genomes. Bioinformatics, 2020; 36:3552-3555. doi: 10.1093/bioinformatics/btaa145.

7. Wrapp D, Wang N, Corbett KS, et al. Cryo-EM structure of the 2019-nCoV spike in the perfusion conformation. *Science*. 2020;367:1260-1263. doi: 10.1126/science.abb2507.
8. Zhang L, Jackson CB, Mou H, et al. The D614G mutation in the SARS-CoV-2 spike protein reduces S1 shedding and increases infectivity. Preprint. *bioRxiv*. 2020;2020.06.12.148726. doi:10.1101/2020.06.12.148726
9. Pachetti M, Marini B, Benedetti F, et al. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. *J Transl Med*. 2020;18:179. doi:10.1186/s12967-020-02344-6

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 July 2020 doi:10.20944/preprints202007.0251.v1

Table 1. Selected nucleotide and amino acid variation in SARS-CoV-2 genomic sequence in mildly affected and severely affected COVID-19 patients

Gene/genomic region	Nucleotide variation	Amino acid variation	Number (%) of mutations	
			Mild (n=46)	Severe (n=56)
5´ UTR	241C>T	N/A	21 (45.7)	48 (85.7)
ORF1ab	313C>T	---	6 (13.0)	0 (0.0)
	3037C>T	---	21 (45.7)	36 (64.3)
	5700C>A	A1812D (nsp3)	7 (15.2)	0 (0.0)
	8782C>T	---	12 (26.1)	3 (5.4)
	11083G>T	L3606F (nsp6)	13 (28.3)	1 (1.8)
	14408C>T	P4715L (RdRp)	21 (45.7)	46 (82.1)
	14805C>T	---	5 (10.9)	2 (3.6)
	15324C>T	---	1 (2.2)	5 (8.9)
	15957G>T	---	0 (0.0)	5 (8.9)
	18877C>T	---	1 (2.2)	11 (19.6)
	20268A>G	---	3 (6.5)	8 (14.3)
S (Spike)	22468G>T	---	8 (17.4)	0 (0.0)
	23403A>G	D614G	21 (45.7)	48 (85.7)
	24197G>T	A879S	0 (0.0)	5 (8.9)
ORF3a	25563G>T	Q57H	5 (10.9)	15 (26.8)
M (Membrane)	26735C>T	---	0 (0.0)	9 (16.1)
ORF8	28144T>C	L84S	12 (26.1)	3 (5.4)
N (Nucleocapsid)	28854C>T	S194L	0 (0.0)	5 (8.9)
	28878G>A	S202N	8 (17.4)	0 (0.0)
	28881G>A, 28882G >A	R203K	9 (19.6)	14 (25.0)
	28883G>C	G204R	9 (19.6)	13 (23.2)
3´ UTR	29742G>A	N/A	8 (17.4)	0 (0.0)
	29827A>T	N/A	16 (34.8)	0 (0.0)
	29830G>T	N/A	20 (43.5)	0 (0.0)

Any mutation with a frequency of 5 or more in either mildly affected or severely affected group is included in this Table. UTR, un-translated region, ORF, open reading frame; N/A, not applicable; ---, silent mutation; nsp, non-structural protein; RdRp, RNA-dependent RNA polymerase; Mild, mildly affected group; Severe, severely affected group

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 July 2020 doi:10.20944/preprints202007.0251.v1

Table 2. Distribution of spike protein D614G mutation with collection time in mildly affected and severely affected COVID-19 patients

Month	Mildly affected group		Severely affected group	
	Number of samples	Number (%) of D614G mutation	Number of samples	Number (%) of D614G mutation
February	7	1 (14.3)	0	0 (0.0)
March	24	11 (45.8)	22	15 (68.2)
April	15	9 (60.0)	28	27 (96.4)
May	0	0 (0.0)	6	6 (100.0)
Total	46	21 (45.7)	56	48 (85.7)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 July 2020 doi:10.20944/preprints202007.0251.v1

Table 3. Frequency of spike protein D614G mutation in mildly affected and severely affected groups for India and Belgium

Country	Mildly affected group		Severely affected group	
	Number of samples	Number (%) of D614G mutation	Number of samples	Number (%) of D614G mutation
India	21	9 (42.8)	6	6 (100.0)
Belgium	5	3 (60.0%)	8	8 (100.0)