AIJR Preprints
Section: Coronavirus
Article Id: 202, Version: 1, 2020
URL: https://preprints.aijr.org/index.php/ap/preprint/view/202
{Click on above link to see the latest available version of this article}

NOT PEER-REVIEWED

Version 1: Received: 07 August 2020 / Approved: 14 August 2020 / Online: 16 August 2020

Mutations in structural proteins of SARS-CoV-2 and potential implications for the ongoing outbreak of infection in India

Rimjhim Dasgupta

4NBIO, 2502, Glen Classic, Hiranandani Gardens, Powai, Mumbai 400076

Abstract

SARS-CoV-2 has spread in India very quickly from its first reported case on 30 January 2020 in Thrissur, Kerala. With the drastic increasing number of positive cases around the world WHO raised the importance in the assessment of the risk of spread and understanding genetic modifications that could have occurred in the SARS-CoV-2. Using available genome sequence in NCBI repository from the samples of different locations in India, we identified the regions (hotspots) of the viral genome with high rates of mutation. We analysed four regions of the genome encoding structural proteins Spike (S), Nucleocapsid (N), envelop (E) and Membrane (M) proteins. Through computational biology approach, we identified multiple substitution mutations in S and N proteins whereas there is only one substitution in E protein and none in M protein. We showed most of these amino acid residues are evolutionary conserved. The changes in the conserved residues may have significant implication on the stability of the proteins and subsequent interaction with other elements, which are essential for virus propagation. This provides a basis for a better understanding of the genetic variation in SARS-CoV-2 circulating in the India, which might provide important clues for identifying potential therapeutic targets, development of efficient vaccines, antiviral drugs and diagnostic assays for controlling COVID-19.

Keywords: COVID-19, SARS-CoV2, sequence alignment, mutation

1 Introduction

The pandemic Corona Virus Disease 2019 (COVID-19) caused by the RNA coronavirus severe acute respiratory syndrome coronavirus 2 (SARS-CoV2), has spread over 200 countries and infected millions of people worldwide. The first case in India was reported on Jan 30 in Thrissur, Kerala and two other cases were reported by Feb 3, 2020. All of them were the students studying at a university in Wuhan, China and came home (Kerala) for vacation. As the number of the positive cases increasing drastically the World Health Organization (WHO) raised the importance of understanding genetic modification that could have occurred in the SARS-CoV-2 and whether the variations in genome sequences is impacting person to person transmission and mortality rate. Searching for mutations and their evolutionary conservation while the virus continues to spread within the country can offer opportunities for a better understanding of virus evolution, biopathology, and transmission. Having this motivation, considering Wuhan based genome NC_045512.2 as reference, we report on multiple missense mutations on structural proteins Spike (S), nucleocapsid (N) and envelope (E) in many Indian isolates and absent in the Chinese ones.

How to Cite:

Rimjhim Dasgupta, “Mutations in structural proteins of SARS-CoV-2 and potential implications for the ongoing outbreak of infection in India”. *AIJR Preprints*, 202, version 1, 2020.

2 Methods

Wuhan isolate, SARS-CoV-2 sequence NC_045512.2 (length 29903 nt) was used as a reference sequence and for sequence comparisons. For our study, we considered the sequences that were available in GISAID and NCBI. We first compared 20 available sequences from Wuhan. They showed 100% sequence similarities. In the present report we have focussed on sequence alignments, we have used NCBI BLAST, and CLUSTAL OMEGA. ~150 genome sequences (sourced from Global Initiative on Sharing All Influenza, GISAID, https://www.gisaid.org/ and NCBI) from different regions of India were considered for this study.

3 Results

We have taken attempt to conduct genomic comparison by considering ~150 SARS-Cov2 sequences from patient samples of different locations in India. Since COVID19 started from Wuhan, China, we started our analyses from sequences from Wuhan. We first compared 20 sequences of patient samples from Wuhan. It showed 100% sequence similarities (data not shown). We considered NC_045512.2 as reference for our analysis.

Our analysis revealed that several sites are possessing non-synonymous substitutions and associated amino acid changes. We analysed four regions of the genome encoding structural proteins, those are Spike (S), Nucleocapsid (N), envelop (E) and Membrane (M) proteins. Multiple non-synonymous substitutions were found in S and N proteins. The list of mutations detected in this set of sequences from ~150 samples are provided in Table 1 and table 2.

Total 25 different missense mutations were found in S protein. Among them 8 are in N terminal domain (NTD) of S1 subunit, 3 (R407I, T572I, E583D) are in receptor binding domain (RBD), 1 is in C terminal of S1 subunit and 13 is in S2 domain. We found L54F, R78M, D614G mutations in 9%, 4% and 90% of sequences respectively. L54F and R78M mutations were detected only in some patients in Gujarat. We noticed Y145Stop, R407I mutations in one sequence from Kerala (QHS34546.1). S162I substitution (polar to nonpolar) was observed in one sequence (QJY40469.1). 2% of the sequences have T572I substitutions. All these also have D614G substitutions.

Table 1: The list of mutations is placed in tabular form. Sequence ID is the accession number followed by sample location (patient sample location), mutated residue (single letter code for amino acid has been used) and region (region in protein sequence)

Sequence ID	Sample location	Mutation	Region
QIA98583.1	Kerala	A930V	S2 domain
QHS34546.1	Kerala	Y145Stop, R407I, D614G,	Junction of S1 & S2
QJQ39968.1	Bangalore	D614G	Junction of S1 & S2
QJQ39980.1	Bangalore	D614G	Junction of S1 & S2
QJQ39992.1	Bangalore	D614G	Junction of S1 & S2
QJF77858.1	Hyderabad	D614G	Junction of S1 & S2
QJF77882.1	Hyderabad	D614G	Junction of S1 & S2
QKQ30054.1	Ahmedabad	D614G	Junction of S1 & S2
QKQ30066.1	Ahmedabad	D614G	Junction of S1 & S2
QKQ30090.1	Ahmedabad	D614G	Junction of S1 & S2
QKQ30102.1	Ahmedabad	D614G	Junction of S1 & S2
QKQ30114.1	Ahmedabad	D614G	Junction of S1 & S2

QKQ30150.1	Ahmedabad	D614G	Junction of S1 & S2
QKQ30246.1	Ahmedabad	D614G	Junction of S1 & S2
QKQ30258.1	Ahmedabad	D614G	Junction of S1 & S2
QKQ30270.1	Ahmedabad	D614G	Junction of S1 & S2
QKQ63388.1	Ahmedabad	D614G	Junction of S1 & S2
QJQ28345.1	Ahmedabad	D614G	Junction of S1 & S2
QJQ28357.1	Ahmedabad	D614G	Junction of S1 & S2
QJQ28369.1	Ahmedabad	D614G	Junction of S1 & S2
QJQ28381.1	Ahmedabad	D614G	Junction of S1 & S2
QJQ28405.1	Ahmedabad	D614G	Junction of S1 & S2
QKY60165.1	Bharuch	D614G	Junction of S1 & S2
QJW00363.1	Dahegam	D614G	Junction of S1 & S2
QJW00375.1	Dahegam	D614G	Junction of S1 & S2
QJW00387.1	Dahegam	D614G	Junction of S1 & S2
QKV25909.1	Ahmedabad	D614G	Junction of S1 & S2
QKV27563.1	Dhanera	D614G	Junction of S1 & S2
QKV25945.1	Vadodara	D614G	Junction of S1 & S2
QKV25957.1	Vadodara	D614G	Junction of S1 & S2
QKV25981.1	Vadodara	D614G	Junction of S1 & S2
QKV26029.1	Vadodara	D614G	Junction of S1 & S2
QKV26041.1	Vadodara	D614G	Junction of S1 & S2
QKV26065.1	Vadodara	D614G	Junction of S1 & S2
QKV26089.1	Vadodara	D614G	Junction of S1 & S2
QKG91214.1	Himatnagar	D614G	Junction of S1 & S2
QKI28661.1	Himatnagar	D614G	Junction of S1 & S2
QJS39639.1	Hyderabad	D614G	Junction of S1 & S2
QJY40421.1	Jamnagar	D614G	Junction of S1 & S2
QJY40445.1	Jamnagar	D614G	Junction of S1 & S2
QJY40505.1	Junagadh	D614G	Junction of S1 & S2
QKQ29946.1	Kadi	D614G	Junction of S1 & S2
QKQ29958.1	Kadi	D614G	Junction of S1 & S2
QKQ30042.1	Kalol	D614G	Junction of S1 & S2

QKQ30174.1	Kalol	D614G	Junction of S1 & S2
QKQ30198.1	Kalol	D614G	Junction of S1 & S2
QKH78810.1	Khedbrahma	D614G	Junction of S1 & S2
QKI28637.1	Khedbrahma	D614G	Junction of S1 & S2
QJY40385.1	Kodinar	D614G	Junction of S1 & S2
QKY64792.1	Mandvi	D614G	Junction of S1 & S2
QJW00447.1	Mansa	D614G	Junction of S1 & S2
QJQ28429.1	Mansa	D614G	Junction of S1 & S2
QJW39904.1	Modasa	D614G	Junction of S1 & S2
QKY74628.1	Palanpur	D614G	Junction of S1 & S2
QKY59941.1	Palanpur	D614G	Junction of S1 & S2
QJW39916.1	Prantij	D614G	Junction of S1 & S2
QJT43680.1	Prantij	D614G	Junction of S1 & S2
QKJ68485.1	Rajkot	D614G	Junction of S1 & S2
QKY60049.1	Surat	D614G	Junction of S1 & S2
QKY60085.1	Surat	D614G	Junction of S1 & S2
QJY40529.1	Una	D614G	Junction of S1 & S2
QJY40577.1	Una	D614G	Junction of S1 & S2
QJY51288.1	Vadodara	D614G	Junction of S1 & S2
QJY51348.1	Vadodara	D614G	Junction of S1 & S2
QJY51384.1	Vadodara	D614G	Junction of S1 & S2
QKY65277.1	Bhuj	D614G	Junction of S1 & S2
EPI_ISL_430465	Kolkata	D614G	Junction of S1 & S2
EPI_ISL_430467	Kolkata	D614G	Junction of S1 & S2
EPI_ISL_430468	Kolkata	D614G, G1124V	Junction of S1 & S2, S2
EPI_ISL_430464	Kolkata	D614G, G1124V	Junction of S1 & S2, S2
QKQ30078.1	Ahmedabad	E583D, D614G	RBD, Junction of S1 & S2
QKQ30222.1	Ahmedabad	E583D, D614G	RBD, Junction of S1 & S2
QJW00327.1	Himatnagar	E583D, D614G	RBD, Junction of S1 & S2
QJQ27854.1	Bangalore	K187X, D614G	NTD, Junction of S1 & S2
QKV25969.1	Vadodara	L54F, D614G	NTD, Junction of S1 & S2
QKJ84943.1	Himatnagar	L54F, D614G	NTD, Junction of S1 & S2

QKJ84979.1	Himatnagar	L54F, D614G	NTD, Junction of S1 & S2
QKY64614.1	Kheda	L54F, D614G	NTD, Junction of S1 & S2
QKY64640.1	Kheda	L54F, D614G	NTD, Junction of S1 & S2
QKY74640.1	Palanpur	L54F, D614G	NTD, Junction of S1 & S2
QKY60213.1	Savli	L54F, D614G	NTD, Junction of S1 & S2
QKY60225.1	Savli	L54F, D614G	NTD, Junction of S1 & S2
QKY60025.1	Surat	L54F, D614G	NTD, Junction of S1 & S2
QKY60073.1	Surat	L54F, D614G	NTD, Junction of S1 & S2
QKY60201.1	Vadodara	L54F, D614G	NTD, Junction of S1 & S2
QKY60237.1	Vadodara	L54F, D614G	NTD, Junction of S1 & S2
QKY60264.1	Vadodara	L54F, D614G	NTD, Junction of S1 & S2
QKI28601.1	Jamnagar	L5F, T572I, D614G	NTD, RBD, Junction of S1 & S2
QJY40469.1	Jamnagar	L5F, S162I, D614G	NTD, RBD, Junction of S1 & S2
QJY40409.1	Una	Q677H	S2 domain
QJW69139.1	Modasa	R78M, D614G, Q784X, V785X, K786X, Q787X, I788X, Y789X,	NTD, S2 domain
QJW00315.1	Modasa	R78M, D614G	NTD, Junction of S1 & S2
QJW39892.1	Modasa	R78M, D614G	NTD, Junction of S1 & S2
QJW39928.1	Modasa	R78M, D614G	NTD, Junction of S1 & S2
QJT43704.1	Modasa	R78M, D614G	NTD, Junction of S1 & S2
QJT43716.1	Modasa	R78M, D614G	NTD, Junction of S1 & S2
QJY77055.1	Jamnagar	D614G, K811X, P812X, S813X, K814X, R815X	S2 domain
QKQ30126.1	Ahmedabad	T572I, D614G,	RBD, Junction of S1 & S2
QKQ30186.1	Ahmedabad	T572I, D614G,	RBD, Junction of S1 & S2
QKV27587.1	Nadiad	W152L, D614G	NTD, Junction of S1 & S2
QJF77846.1	Hyderabad	Y28H	NTD

We conducted multiple sequence alignment of S proteins to analyse the evolutionary conservation of amino acid residues for many of these mutations (Figure 1). Our sequence alignment result shows that Leucine at position 5 is evolutionary conversed through bat, pangolin, SARS-Cov, SARS-CoV2, Leucine at 54F is conserved in RatG13, pangolin, MERS-CoV, SARS-CoV and SARS-CoV2, Arginine in position 78 is conserved in bat and SARS-CoV2, Tyrosine at 146 is conserved in bat, pangolin and SARS-CoV2, Arginine at position 408 is conserved in bat, pangolin, SARS-Cov, SARS-CoV2, Threonine at 572 is conserved in bat, pangolin and SARS-CoV2, Aspartic acid at position 614 is conversed in bat, pangolin, SARS-Cov, SARS-CoV2, Glycine at 1124 is conversed in bat, pangolin, SARS-Cov, SARS-CoV2.

```
AFS88936.1_MERS-CoV      MIHSVFLLMFLLTPTESYVDVGPDSVKSACIEVDIQQTFFDKTWPRPIDVSKADGIIYPQ 60
AAP30030.1_SARS-CoV      ----MFIFLLFL-TLTSGS------DLDRCTTFDDVQ-----APNYTQHTSSMRGVYYPD 44
QIA48614.1_Pangolin      ----MFVFLFVL-PLVSSQ----------CVNLTTRT-----GIPPGYTNSSTRGVYYPD 40
QHR63300.2_RaTG13        ----MFVFLVLL-PLVSSQ----------CVNLTTRT-----QLPPAYTNSSTRGVYYPD 40
QHS34546.1               ----MFVFLVLL-PLVSSQ----------CVNLTTRT-----QLPPAYTNSFTRGVYYPD 40
YP_009724390.1_Ref       ----MFVFLVLL-PLVSSQ----------CVNLTTRT-----QLPPAYTNSFTRGVYYPD 40
QKY60264.1               ----MFVFLVLL-PLVSSQ----------CVNLTTRT-----QLPPAYTNSFTRGVYYPD 40
EPI_ISL_430464           ----MFVFLVLL-PLVSSQ----------CVNLTTRT-----QLPPAYTNSFTRGVYYPD 40
QJW00315.1               ----MFVFLVLL-PLVSSQ----------CVNLTTRT-----QLPPAYTNSFTRGVYYPD 40
QKQ30126.1               ----MFVFLVLL-PLVSSQ----------CVNLTTRT-----QLPPAYTNSFTRGVYYPD 40
QKI28601.1               ----MFVFFVLL-PLVSSQ----------CVNLTTRT-----QLPPAYTNSFTRGVYYPD 40
                                  5

AFS88936.1_MERS-CoV      GRTYSNITITYQGLF-PYQGDHGDMYVYSAGHATGTTPQKLFVANYSQDVKQFANGFVVR 119
AAP30030.1_SARS-CoV      EIFRSDTLYLTQDLFLPFYSNV---TGFHTIN-------HT----FDNPVIPFKDGIYFA 90
QIA48614.1_Pangolin      KVFRSSILHLTQDLFLPFFSNV---TWFNTINYQG--GFKK----FDNPVLPFNDGVYFA 91
QHR63300.2_RaTG13        KVFRSSVLHLTQDLFLPFFSNV---TWFHAIHVSGTNGIKR----FDNPVLPFNDGVYFA 93
QHS34546.1               KVFRSSVLHSTQDLFLPFFSNV---TWFHAIHVSGTNGTKR----FDNPVLPFNDGVYFA 93
YP_009724390.1_Ref       KVFRSSVLHSTQDLFLPFFSNV---TWFHAIHVSGTNGTKR----FDNPVLPFNDGVYFA 93
QKY60264.1               KVFRSSVLHSTQDFFLPFFSNV---TWFHAIHVSGTNGTKR----FDNPVLPFNDGVYFA 93
EPI_ISL_430464           KVFRSSVLHSTQDLFLPFFSNV---TWFHAIHVSGTNGTKR----FDNPVLPFNDGVYFA 93
QJW00315.1               KVFRSSVLHSTQDLFLPFFSNV---TWFHAIHVSGTNGTKM----FDNPVLPFNDGVYFA 93
QKQ30126.1               KVFRSSVLHSTQDLFLPFFSNV---TWFHAIHVSGTNGTKR----FDNPVLPFNDGVYFA 93
QKI28601.1               KVFRSSVLHSTQDLFLPFFSNV---TWFHAIHVSGTNGTKR----FDNPVLPFNDGVYFA 93
                                  54                      78

AFS88936.1_MERS-CoV      LRA--FYCILEPRSGNHCPAGNSYTSFATYHTPATDCSDGNYNRNASLNSFKEYFNLRNC 237
AAP30030.1_SARS-CoV      IRACNFELCDNPFFAVSKPMG----------------------T----QTHTMIFDNAFNC 159
QIA48614.1_Pangolin      IKVCEFQFCTDPFLGVYYHNN----------------------NKTWVENEFRVYSSANNC 164
QHR63300.2_RaTG13        IKVCEFQFCNDPFLGVYYHKN----------------------NKSWMESEFRVYSSANNC 166
QHS34546.1               IKVCEFQFCNDPFLGVY-HKN----------------------NKSWMESEFRVYSSANNC 165
YP_009724390.1_Ref       IKVCEFQFCNDPFLGVYYHKN----------------------NKSWMESEFRVYSSANNC 166
QKY60264.1               IKVCEFQFCNDPFLGVYYHKN----------------------NKSWMESEFRVYSSANNC 166
EPI_ISL_430464           IKVCEFQFCNDPFLGVYYHKN----------------------NKSWMESEFRVYSSANNC 166
QJW00315.1               IKVCEFQFCNDPFLGVYYHKN----------------------NKSWMESEFRVYSSANNC 166
QKQ30126.1               IKVCEFQFCNDPFLGVYYHKN----------------------NKSWMESEFRVYSSANNC 166
QKI28601.1               IKVCEFQFCNDPFLGVYYHKN----------------------NKSWMESEFRVYSSANNC 166
                                        146

AFS88936.1_MERS-CoV      RFVYDAYQNLVGYYSD--DGNYYCLRACVSVPVSVIYD--KETKTHATLFGSVACEHISS 684
AAP30030.1_SARS-CoV      QFGRDVSDFTD-SVRDPKTSEILDISPCSFGGVSVITPGTNASSEVAVLYQDVNCTDVST 608
QIA48614.1_Pangolin      QFGRDISDTTD-AVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQDVNCTEVPM 620
QHR63300.2_RaTG13        QFGRDIADTTD-AVRDPQTLEILDITPCSFGGVSVITPGTNASNQVAVLYQDVNCTEVPV 622
QHS34546.1               QFGRDIADTTD-AVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQDVNCTEVPV 621
YP_009724390.1_Ref       QFGRDIADTTD-AVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQDVNCTEVPV 622
QKY60264.1               QFGRDIADTTD-AVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPV 622
EPI_ISL_430464           QFGRDIADTTD-AVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPV 622
QJW00315.1               QFGRDIADTTD-AVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPV 622
QKQ30126.1               QFGRDIADITD-AVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPV 622
QKI28601.1               QFGRDIADITD-AVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPV 622
                                                          614

AFS88936.1_MERS-CoV      CIAPVNGYFIKTNNTRIVDEWSYTGSSFYAPEPITSLNTKYVAPQVTYQN-ISTNLPPPL 1222
AAP30030.1_SARS-CoV      -YFPREGVFVFN-----GTSWFITQRNFFSPQIITTDNTFVSGNCDVVIGIINNTVYDPL 1123
QIA48614.1_Pangolin      -HFPREGVFVSN-----GTHWFITQRNFYEPQIITTDNTFVSGSCDVVIGIVNNTVYDPL 1135
QHR63300.2_RaTG13        -HFPREGVFVSN-----GTHWFVTQRNFYEPQIITTDNTFVSGSCDVVIGIVNNTVYDPL 1137
QHS34546.1               -HFPREGVFVSN-----GTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPL 1140
YP_009724390.1_Ref       -HFPREGVFVSN-----GTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPL 1141
QKY60264.1               -HFPREGVFVSN-----GTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPL 1141
EPI_ISL_430464           -HFPREGVFVSN-----GTHWFVTQRNFYEPQIITTDNTFVSVNCDVVIGIVNNTVYDPL 1141
QJW00315.1               -HFPREGVFVSN-----GTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPL 1141
QKQ30126.1               -HFPREGVFVSN-----GTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPL 1141
QKI28601.1               -HFPREGVFVSN-----GTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPL 1141
                                                    1124
```

Figure 1: Alignment of S proteins (MERS-CoV accession # AFS88936.1; SARS-CoV accession # AAP30030.1; Pangolin coronavirus accession # QIA48614.1; Bat coronavirus RaTG13 accession # QHR63300.2; SARS-CoV2 REFSEQ: accession NC_045512.2 accession # YP_009724390.1; SARS-CoV2 accession # QHS34546.1 of Kerala patient sample; SARS-CoV2 accession # QKY60264.1 of Vadodara patient sample; SARS-CoV2 accession # EPI_ISL_430464 Kolkata patient sample; SARS-CoV2 accession # QJW00315.1 Modasa patient sample; SARS-CoV2 accession # QKQ30126.1 Ahmedabad patient sample; SARS-CoV2 QKI28601.1 Bayad patient sample); Mutated conserved amino acid positions are highlighted in green and mutated amino acids are in pink highlight; The numbers below the alignments denote the residue position of corresponding to SARS-CoV-2 S protein

Total 23 different missense mutations were found in nucleocapsid (N) protein. One sample from Gujarat (QKI28681.1) has P6T mutation, nonpolar to polar amino acid. P13L mutation was found in 10% sequences. D22Y and D22N were found in 2 sequences and both also have additional mutations. S33I mutation was found in 2 sequences and these do not have any additional mutation in N protein. 4% sequences have S194L (polar to non-polar) mutation and among them 6% also has S202N (both are polar but Ser is polar neutral whereas Arg is hydrophilic) mutation. Two mutations, R203K and G204R were found in ~6% of ~150 sequences. Sample locations are Gujarat, Bangalore, Pune, Delhi and Kolkata.

We also conducted multiple sequence alignment of N proteins to analyse the evolutionary conservation of amino acid residues for many of these mutations. Our result shows that Proline at position 6, 13 are conversed in bat, pangolin, SARS-Cov, SARS-CoV2; Aspartic acid at position 22 is conversed in bat, pangolin, MERS-CoV, SARS-Cov, SARS-CoV2; serine at position 194, Glycine at position 204 are conserved in are conversed in bat, pangolin, MERS-CoV, SARS-Cov, SARS-CoV2; Arginine is conversed in bat, pangolin, SARS-Cov, SARS-CoV2 (Figure 2)

```
AFS88943.1_MERS-CoV      ------MASPAAPRAVSFADNNDITNTN----LSRGRGRNPKPRAAPNNTVSWYTGLTQH 50
AAP30037.1_SARS-CoV      MSDNGPQSNQRSAPRITFGGPTDSTDNNQNGGRNGARPKQRRPQGLPNNTASWFTALTQH 60
QIG55953.1_Pangolin      MSDNGPQ-NQRNAPRITFGGPSDSAGSNQNGERSGARPKQRRPQGLPNNTASWFTALTQH 59
QHR63308.1_RaTG13        MSDNGPQ-NQRNAPRITFGGPSDSTGSNQNGERSGARPKQRRPQGLPNNTASWFTALTQH 59
QKY74648.1               MSDNGPQ-NQRNAPRITFGGPSDSTGSNQNGERSGARSKQRRPQGLPNNTASWFTALTQH 59
QKY60069.1               MSDNGPQ-NQRNAPRITFGGPSDSTGSNQNGERSGARSKQRRPQGLPNNTASWFTALTQH 59
QJY40417.1               MSDNGPQ-NQRNALRITFGGPSDSTGSNQNGERSGARSKQRRPQGLPNNTASWFTALTQH 59
QKY60093.1               MSDNGPQ-NQRNAPRITFGGPSNSTGSNQNGERSGARSKQRRPQGLPNNTASWFTALTQH 59
YP_009724397.2_Ref       MSDNGPQ-NQRNAPRITFGGPSDSTGSNQNGERSGARSKQRRPQGLPNNTASWFTALTQH 59
QKI28681.1               MSDNGTQ-NQRNAPRITFGGPSDSTGSNQNGERSGARSKQRRPQGLPNNTASWFTALTQH 59
                                 6        13        22

AFS88943.1_MERS-CoV      SQSSSRASSLSRNSSRSSSQGSRSGNSTRGTSPGPSGIGAVGGDLLYLDLLNRLQALESG 228
AAP30037.1_SARS-CoV      SQASSRSSSRSRGNSRNSTPGSSRGNSPARMA---SGGGETALALLLLDRLNQLESKVSG 237
QIG55953.1_Pangolin      SQASSRSSSRSRNSSRNTTPGSSRGTSPARMA---GNGGDAALALLLLDRLNQLESKMSG 236
QHR63308.1_RaTG13        SQASSRSSSRSRNSSRNSTPGSSRGTSPARMA---GNGSDAALALLLLDRLNQLESKMSG 236
QKY74648.1               SQASSRSSSRSRNSLRNSTPGSSGRTSPARMA---GNGGDAALALLLLDRLNQLESKMSG 236
QKY60069.1               SQASSRSSSRSRNSLRNSTPGSSNRGTSPARMA---GNGGDAALALLLLDRLNQLESKMSG 236
QJY40417.1               SQASSRSSSRSRNSSRNSTPGSSRGTSPARMA---GNGGDAALALLLLDRLNQLESKMSG 236
QKY60093.1               SQASSRSSSRSRNSLRNSTPGSSRGTSPARMA---GNGGDAALALLLLDRLNQLESKMSG 236
YP_009724397.2_Ref       SQASSRSSSRSRNSSRNSTPGSSRGTSPARMA---GNGGDAALALLLLDRLNQLESKMSG 236
QKI28681.1               SQASSRSSSRSRNSSRNSTPGSSRGTSPARMA---GNGGDAALALLLLDRLNQLESKMSG 236
                                        194      202-204
```

Figure 2: Alignment of N proteins (MERS-CoV accession # AFS88943.1; SARS-CoV accession # AAP30037.1; Pangolin coronavirus accession # QIG55953.1; Bat coronavirus RaTG13 accession # QHR63308.1; SARS-CoV2 REFSEQ: accession NC_045512.2 accession # YP_009724397.2; SARS-CoV2 accession # QKY74648.1 of Palanpur patient sample; SARS-CoV2 accession # QJY40417.1 of Una patient sample; SARS-CoV2 accession # QKY60093.1 of Surat patient sample; SARS-CoV2 accession # QKY60069.1 of Surat patient sample; SARS-CoV2 accession # QKI28681.1 of Khedbrahma patient sample; Mutated conserved amino acid positions are highlighted in green and mutated amino acids are in pink highlight; The numbers below the alignments denote the residue position of corresponding to SARS-CoV-2 N protein

SARS- CoV2 envelope (E) protein is a small, integral membrane protein involved in several aspects of the virus' life cycle, such as assembly, budding, envelope formation, and pathogenesis. One substitution (V62F) was found at the C terminal end in one sample from Gujarat (Figure 3)

```
                  62
QKV26079.1    RFKNLNSSRVPDLLV 75
NC_045512_E RVKNLNSSRVPDLLV 75
```

Figure 3: Alignment of C terminal end of E protein shows substitution mutation V62F (SARS-CoV2 accession # QKV26079.1 of Nadiad patient sample with Reference sequence NC_045512.2 accession # YP_009724392.1); Mutated residue in pink highlight; The numbers above the alignment denote the residue position of corresponding to SARS-CoV-2 E protein

The important point to note here that most of the patient samples where mutations were identified either have international travel history or direct contact to the person travelled.

Table 2: The list of mutations is placed in tabular form. Sequence ID is the accession number followed by sample location (patient sample location), mutated residue (single letter code for amino acid has been used) and region (region in protein sequence)

Sequence ID	Sample location	Mutation	Region
QKI28681.1	Khedbrahma	P6T	NTD
QJU70565.1	Hyderabad	P13L	NTD
QJQ27850.1	Bangalore	P13L	NTD
QJQ27874.1	Bangalore	P13L	NTD
QJQ27886.1	Bangalore	P13L	NTD
QJF77878.1	Hyderabad	P13L	NTD
QJF11844.1	Bangalore	P13L	NTD
QJF11856.1	Bangalore	P13L	NTD
QJF11868.1	Bangalore	P13L	NTD
QJR84533.1	Ahmedabad	P13L	NTD
QJY40405.1	Botad	P13L	NTD
QJS39659.1	Hyderabad	P13L	NTD
QJT43688.1	Prantij	P13L	NTD
QJQ28437.1	Mansa	P13R	NTD
QJY77063.1	Jamnagar	P13L	NTD
QJY40417.1	Una	P13L, A146S	NTD, RNA binding domain
QKQ30158.1	Ahmedabad	D22Y, S194L	NTD, Linker peptide
QKY60093.1	Surat	D22N, L139F, S194L	NTD, RNA binding domain, Linker peptide
QJQ28377.1	Ahmedabad	S33I	NTD
QJQ28389.1	Ahmedabad	S33I	NTD
QKV26073.1	Vadodara	A134V	RNA binding domain
QJF77854.1	Hyderabad	P344S	dimerization domain (DD)
QJW69147.1	Modasa	D399X, L400X, D401X, D402X, F403X, S404X, K405X, Q406X, L407X,	C terminal end
QKY60069.1	Surat	S194L, S202N	Linker peptide
QJY40429.1	Jamnagar	S194L, D348Y	Linker peptide, dimerization domain (DD)
QKI28669.1	Himatnagar	S194L	Linker peptide
QKY74648.1	Palanpur	S194L, R203G, G204R	Linker peptide
QJY40585.1	Una	R203K, G204R	Linker peptide
QJY40537.1	Una	R203K, G204R	Linker peptide
QKY74636.1	Palanpur	R203K, G204R	Linker peptide
QKY64800.1	Mandvi	R203K, G204R	Linker peptide
QJF11880.1	Bangalore	R203K, G204R	Linker peptide
QJF11832.1	Bangalore	R203K, G204R	Linker peptide
QJF11820.1	Bangalore	R203K, G204R	Linker peptide
QJH92187.1	Pune	R203K, G204R	Linker peptide
EPI_ISL_430468	Kolkata	R203K, G204R	Linker peptide
EPI_ISL_430464	Kolkata	R203K, G204R	Linker peptide
EPI_ISL_430465	Kolkata	R203K, G204R	Linker peptide

4 Discussion

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is being intensively studied, particularly its evolution, in the increasingly available sequences with classical phylogenetic tree representation. We report here certain protein mutations which could have potential structural and/or functional implication.

4.1 Mutations in Spike protein and potential implication

Coronavirus entry into host cells is mediated by the transmembrane spike (S) glycoprotein that forms homotrimers protruding from the viral surface (Tortorici and Veesler, 2019).
S comprises two functional subunits, responsible for binding to the host cell receptor (S1 subunit) and fusion of the viral and cellular membranes (S2 subunit). For many Coronaviruses, S is cleaved at the boundary between the S1 and S2 subunits, which
remain non-covalently bound in the prefusion conformation. This region is reported to be the most potent and indispensable for viral attachment and entry into host system (Walls et al., 2020). The miss-sense mutations in S protein those we report are mostly single point mutations with few double and triple mutations. All these mutations could be classified as stabilizing and destabilizing based on the free-energy changes.

We found L5F, L54F substitutions in 1.5% and 9% sequences considered in this study. Both amino acids are non-polar, but phenylalanine has a benzoic ring in the side chain which may stiffen the secondary structure by means of aromatic-aromatic, hydrophobic or stacking interactions. Earlier report in Non- Structural Protein 6 (NSP6) by amino acid change stability (ACS) analysis showed that this (leucine to phenylalanine) leads to a lower stability of the protein structure (Benvenuto et al. 2020). We noticed R78M mutation in 4% of the sequences studied. Arginine (R) is basic, hydrophilic whereas methionine (M) is nonpolar hydrophobic. Other mutations in N terminal domain (NTD) includes Y145Stop, W152L, S162I, K187X. Previous study (Walls et al., 2020) showed a high thermal factor for the NTD (through the cryo-EM structure of the S protein). As these mutations are found at NTD, this observation puts an open question, whether the virus adopts viability through mutations, stabilizing the flexible NTD.

The receptor-binding domain or RBD of the Spike protein of SARS-CoV-2 lies between amino acids 330 and 583. We found R407I, T572I, E583D mutations in RBD. The high-temperature factor of RBD was reported earlier (Walls et al., 2020) which is associated with its dynamic nature leading to the conformational switch between close and open states. All three substitutions we report here, lie in the RBD, which could be corelated with the alteration in receptor binding affinity. Arginine (R407I) is a positively charged amino acid and Isoleucine is a hydrophobic amino acid. While positively charged amino acid could be more exposed, hydrophobic amino acids secure themselves away from the outer aqueous environment. For T572I, Threonine is polar and Isoleucine is a hydrophobic amino acid. Recent report suggests that S protein binds to ACE2 (angiotensin converting enzyme 2) through Leu455, Phe486, Gln493, Asn501, and Tyr505 (Liu et al. 2020). This substitution mutation of neighbouring residues may impact host receptor interaction by altering the protein conformation. All together these may potentially impact on the binding affinity to host receptor.

D614G substitution was observed in ~90% sequences studied in our analysis. It occurs either as a single mutation or coupled with other mutations (L5F/S162I/D614G, L54F/D614G, R78M/D614G, W152L/D614G, E583D/D614G, Y145Stop/R407I/D614G). Laha et al (2020) showed that the change from Aspartic acid to Glycine alters the electro-static potential of the surface of the protein. This change creates a favourable environment in a hydrophobic pocket of the S protein. Another report (Flores and Cardozo 2020) revealed that the amino acid at position 614 occurs at an internal protein interface of the viral spike, and the presence of G at this position destabilises a specific conformation, within which the key host receptor binding site is more accessible. It makes G614 is a more pathogenic. In this study we have seen that L5F, L54F, R78M mutations are individually coupled with D614G. Earlier (Benvenuto et al. 2020) ACS analysis showed that leucine to phenylalanine leads to a lower stability of the protein structure. There is possibility that D614G coupled with Leucine to Phenylalanine mutation brings more favourable environment to enhance accessibility to host receptor and makes the virus more infectious.

Another substitution Q677H was observed in one sequence of patient sample from Una. This position lies at the junction of S1 and S2 subunit in the vicinity of furin cleavage site, which plays an important role in virus entry. Glutamine is polar whereas Histidine is a basic hydrophilic amino acid. It was reported earlier that glutamine to histidine in H1N1 had a severe impact in virus entry replication and cross infectivity to other species (Tiwari and Mishra, 2020). Whether both these mutations (D614G, Q677H) have resulted in the evolution of a more stable and transmissible viral subtype needs investigation.

Mutation A930V was found in the sequence (QIA98583.1) from Kerala sample which falls in the S2 subunit. Considering the nature of valine being destabilizing causing distortion (Laha et al 2020), this mutation might have implications in viral membrane fusion. The residue alanine at position 930 stabilizes the protein due to its hydrophobic nature. On substitution with Valine in the same position, it can potentially change the affinity of the molecule toward its receptor. Secondly this substitution was observed only in QIA98583.1 (patient from Kerala who was studying in Wuhan). Interestingly this sequence doesn't have D614G mutation. We have not found D614G or any other mutations (we reported here) in any Chinese sequences studied in this study whereas many of these including D614G are reported in sequences from other countries (Flores and Cardozo 2020). Probably this is the indication that the virus is getting mutated after multiple passages. Our multiple sequence alignment shows that the amino acid residues which had undergone mutations, are evolutionary conserved (Figure 1). This implies the significance of these residues in respective positions. Any alteration could impact the structural and functional conformation of the protein which makes it more infectious.

Although the clinical significance of the observed mutations is not readily available, our findings in Indian patients lay the ground work for India to understand the impact of SARS-CoV2 mutations on disease severity, host immune response, vaccine development and serological response.

4.2 Mutations in Nucelocapsid (N) protein and potential structural and functional implication

The nucleocapsid protein is an important structural protein for the coronaviruses. It is highly abundant in the viruses. Its function involves entering the host cell, binding to the RNA, and forming the ribonucleoprotein core. It consists of RNA binding domain (RBD; residues 44-180) in the N-terminal region (N) of the protein, linker peptide (residues 181-246), the dimerization domain (DD; residues 247-364) in the C-terminal region (Zeng et al. 2020). Three disordered regions were reported on 1) N terminal (residues 1-43), 2) linker peptide, and 3) C terminal end (residues 365-419).

We found P6T, P13L, D22Y, D22N, S33I in first disordered region, A134V, L139F, A146S in RBD, S194L, S202N, R203G, G204R in linker (second disordered region), P344S, D348Y in dimerization domain and D399X, L400X, D401X, D402X, F403X, S404X, K405X, Q406X, L407X in 3rd disordered region . R203K/G and G204R mutations were observed in ~6% sequences. Sometimes it is coupled with S194L. Triple base mutations of 2881-2883 GGG/AAC was detected in coding region that results in two consecutive amino acid changes, R203K and G204R. The sites of these mutations are located in the SR-rich intrinsically disordered region (Chang et al. 2014). Moreover, this region surrounds with GSK3 phosphorylation 'SRGTS' (amino acid position 202-206) and CDK phosphorylation 'SPAR' (amino acid position 206-209) motifs (Surjit et al. 2005). When Ser202 is phosphorylated it incorporates a large negative group tethered to the sidechain of Ser, as seen in many other substrates of kinases. Arg203 is a part of GSK3 phosphorylation motif and its sidechain could potentially contribute to charge neutralization at phosphorylated-Ser202.

Earlier structural study (Zeng et al. 2020) reports that multiple disordered regions facilitate the N protein to transiently bind to different partners and maintain a correct conformation. Since we report multiples mutations in these disordered regions, possibly it impacts the binding of different partners to attain the correct conformation. Our protein alignment study shows that most of the positions in which point mutations occurred are conserved in bat, pangolin, SARS-Cov and MERS-CoV (Figure 2). It implies the structural and functional significance of these residues for protein stability and interaction with binding partners.

4.3 Mutation in envelope (E) protein and its potential implication

The envelope (E) protein is a small, integral membrane protein involved in several aspects of the virus' life cycle, such as assembly, budding, envelope formation, and pathogenesis. Recent studies have reported its structural motifs and topology. The report also showed its function as an ion-channelling viroporin, and it interacts with both other coronaviruses (MERS-CoV, SARS-CoV) proteins and host cell proteins (Schoeman and Fielding 2019). The SARS-CoV E protein consists of three domains, i.e. the amino (N)-terminal domain (residues ~1-8), the transmembrane domain (TMD, residues 9 -38)), and the carboxy (C)-terminal domain (residues 39- 75) (Schoeman and Fielding 2019). We found one substitution (V62F) at the C-terminus in one sequence of patient sample from Gujarat. Earlier study showed that V25F hampers the oligomerisation of SARS-CoV E, to some extent. It was also reported that SARS-CoV E protein contain a binding motif known as the postsynaptic density protein 95 (PSD95)/Drosophila disc large tumour suppressor (Dlg1)/zonula occludens-1 protein (zo-1) (PDZ)-binding motif (PBM), located in the last four amino acids of the C terminus (Schoeman and Fielding 2019). The PBM motif can bind to the C-terminus of target proteins such as the cellular adapter proteins involved in host-cell processes important for viral infection. We identified the similar motif 'DLLV' (residues: 72-75) at the C-terminus in all the sequences of SARS-COV2 analysed in this study (Figure 3). V62F can potentially interfere the interaction with target proteins thereby altering the host-cell processes required for viral infection.
This analysis does not find any mutation in Membrane (M) protein in the sequences from ~150 patient samples of different regions in India. This implies that M protein is comparably stable but needs further investigation by increasing sample size and covering more locations.

5 Conclusion

All together our findings provide leads which might benefit outbreak tracking, development of therapeutic and prophylactic strategies against the infection. These mutations in the conserved amino acids, likely developed during virus spread, could affect virus host receptor interaction, stability, intracellular survival. This report provides important insights for functional validation to understand the molecular basis of differential disease severity. Genome follow-up of SARS-CoV-2 spread is urgently needed in order to identify mutations that could significantly modify virus pathogenicity. This study also warrants the importance of sequencing the whole genome of SARS-CoV-2 after several passages and key mutations should be noted for the effective drug designing and treatment options such as antiviral and immune therapy. The high mortality rate, along with their ease of transmission, underpins the need for more research into SARS-Cov2 molecular biology which can aid in the production of effective anti-coronaviral agents.

References

1. Becerra-Flores M, Cardozo T. SARS-CoV-2 viral spike G614 mutation exhibits higher case fatality rate. *Int J Clin Pract*. 00: e13525, 2020. https://doi. org/10.1111/ijcp.13525
2. Domenico Benvenuto, Ayse Banu Demir, Marta Giovanetti, Martina Bianchi, Silvia Angeletti, Stefano Pascarella, Roberto Cauda, Massimo Ciccozzi, Antonio Cassone. Evolutionary analysis of SARS-CoV-2: how mutation of Non-Structural Protein 6 (NSP6) could affect viral autophagy, *Journal of Infection* 81, e24–e27, 2020 https://doi.org/10.1016/j.jinf.2020.03.058
3. Chang C-k, Hou M-H, Chang C-F, Hsiao C-D and HuangT-HThe SARS coronavirus nucleocapsid protein – Forms and functions. *Antivir. Res*. 103, pp. 39–50, 2014
4. Dewald Schoeman and Burtram C. Fielding. Coronavirus envelope protein: current knowledge. *Virology Journal* 16,69, 2019
5. Sayantan Laha, Joyeeta Chakraborty, Shantanab Das, Soumen Kanti Manna, Sampa Biswas, Raghunath Chatterjee. Characterizations of SARS-CoV-2 mutational profile, spike protein stability and viral transmission. *Infection, Genetics and Evolution* 85, 104445, 2020. https://doi.org/10.1016/j.meegid.2020.104445
6. Liu Z, Xiao X, Wei X, et al. Composition and divergence of coronavirus spike proteins and host ACE2 receptors predict potential intermediate hosts of SARS-CoV-2. *J Med Virol*. 92, pp.595–601, 2020. https://doi.org/10.1002/jmv.25726
7. Manish Tiwari, Divya Mishra. Investigating the genomic landscape of novel coronavirus (2019-nCoV) to identify non-synonymous mutations for use in diagnosis and drug design. *Journal of Clinical Virology*, 128, 104441, 2020
8. Surjit M, Kumar R, Mishra RN, Reddy MK, Chow VTK. The severe acute respiratory syndrome coronavirus nucleocapsid protein is phosphorylated localizes in the cytoplasm by 14-3-3-mediated translocation. *J. Virol*. 79, pp. 11476–11486, 2005
9. Tortorici, M.A., and Veesler, D. Structural insights into coronavirus entry. *Adv. Virus Res*. 105, pp. 93–116. 2019.
10. Alexandra C. Walls, Young-Jun Park, M. Alejandra Tortorici, Abigail Wall, Andrew T. McGuire, David Veesler. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein; *Cell* 180, pp. 281–292, 2020. https://doi.org/10.1016/j.cell.2020.02.05827.
11. Weihong Zeng, Guangfeng Liu, Huan Ma, Dan Zhao, Yunru Yang, Muziying Liu, Ahmed Mohammed, Changcheng Zhao, Yun Yang, Jiajia Xie, Chengchao Ding, Xiaoling Ma, Jianping Weng, Yong Gao, Hongliang He, Tengchuan Jin, Biochemical characterization of SARS-CoV-2 nucleocapsid protein. *Biochemical and Biophysical Research Communications* ,527, pp. 618-623, 2020