Journal of Biochemistry and Biophysics

Volume 3 | Issue 1

Research Article Open Access

Genetic Variation of SARS-CoV-2 Circulating Worldwide and its Association for Altering Disease Fatality

Dasgupta R*

Glen Classic, Hiranandani Gardens, Powai, Mumbai, India

***Corresponding author:** Dasgupta R, 4NBIO, 2502, Glen Classic, Hiranandani Gardens, Powai, Mumbai, India, Tel: 229004030428, E-mail: rxdasg@gmail.com

Received Date: October 19, 2020 **Accepted Date:** January 27, 2021 **Published Date:** January 29, 2021

Citation: Dasgupta R (2021) Genetic Variation of SARS-CoV-2 Circulating Worldwide and its Association for Altering Disease Fatality. J Biochem Biophy 3(1): 101

Abstract

The emergence of SARS-CoV-2 has resulted in > 36,361,054 infections and > 1,056,186 deaths worldwide. Using publicly available genome sequences of patient samples from different geographical regions, a study has been conducted to co-relate mutational frequency with disease transmission and fatality rate. Seven hundred genome sequences were randomly chosen from different countries. The regions of the genome encoding structural proteins Spike (S), Nucleocapsid (N), envelop (E) and Membrane (M) proteins and ORF8 were studied here. Through Insilco approach, this study showed that several evolutionary conserved amino acid residues underwent mutations. Some of these mutations are common in multiple geographies. This study highlights that the mutational rate is inversely proportional to disease fatality and favours disease transmission. The changes in the conserved residues have significant implication on the stability of the proteins and subsequent interaction, which are essential for virus propagation. This provides a better understanding of the genetic variation in SARS-CoV-2 across the countries and its association with reducing disease fatality.

Keywords: COVID-19; SARS-CoV2; Disease fatality; Sequence alignment; Mutation; Structural protein

Introduction

The COVID-19 pandemic caused by a novel 2019 SARS coronavirus, known as SARS-CoV-2, is rapidly spreading worldwide after one month of the initially identified case on December 2019, in Wuhan city, China [1]. The genome sequence study has revealed that SARSCoV-2 is a member of the genus Beta-coronavirus and belongs to the subgenus Sarbecovirus that includes SARS-CoV while MERS-CoV belongs to a separate subgenus, Merbecovirus [2,3]. Epidemiological data suggests that SARS-CoV-2 had spread widely from the city of Wuhan in China [4] after its zoonotic transmission originating from bats via the Malayan pangolins [5]. It has spread over 200 countries and infected millions of people worldwide. As the number of the positive cases increasing drastically, the World Health Organization (WHO) raised the importance of understanding genetic changes through mutation that could have occurred in the SARS-CoV-2. The SARS-CoV-2 genome is composed of approximately 30,000 nucleotides [5]. The genome includes a variable number (from 6 to 11) of open reading frames (ORFs) [6]. The first ORF (ORF1ab) representing approximately 67% of the entire genome encodes 16 non-structural proteins (nsps), while the remaining ORFs encode accessory proteins including ORF8 and structural proteins. The four major structural proteins are the spike surface glycoprotein (S), small envelope protein (E), membrane protein (M), and nucleocapsid (N) protein [6].

Searching for mutations and their evolutionary conservation while the virus continues to spread, 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, this study attempted to understand worldwide common variants, varying mutational frequency from region to region and it's association for reducing disease fatality.

Methods

In the present report we have randomly chosen 700 genome sequences (mostly sourced from NCBI virus (Severe acute respiratory syndrome coronavirus 2 data hub) and very few Global Initiative on Sharing All Influenza, GISAID (https://www.gisaid.org/)) from Covid-19 patient samples worldwide. There are few criteria that have been used while randomly selecting the genome sequences from Covid-19 patient samples. These are the length of the sequences, number of available sequences, isolation source and time frame. Only complete sequences are used for this study and time frame is Jan, 2020 to July, 2020. The isolation source is oronasopharynx. The minimum number of sequences is ten and maximum is seventy. This study excludes many geographical

locations due to unavailability of sufficient number of sequences (at least ten). GISAID was not readily accessible to author during this study. Here author studied sixty sequences for China. The study is based on multiple sequence alignments using CLUSTALW with default parameters. Wuhan isolate, SARS-CoV-2 sequence NC_045512.2 (length 29903 nt) was used as a reference sequence and for sequence comparison. By comparing the protein sequences of four SARS-CoV-2 structural and one nonstructural (ORF8) protein base pair differences were identified. Briefly, Multiple Sequence Alignment (MSA) of each of the amino acid sequences of SARS-CoV-2 proteins with their respective reference sequences of SARS-CoV-2_Wuhan (accession numbers YP_009724390, YP_009724392, YP_009725302, YP_009724393, YP_009724397, and YP_009724396 for S, E, M, N and ORF8 respectively) was performed using CLUSTALW with default parameters.

First, sequence alignment was conducted between spike (S) proteins of SARS-CoV2 accession # YP_009724390.1 (coded by NC_045512.2 reference sequence, region: 21563..25384), MERS-CoV accession # AFS88936.1 (coded by JX869059.2: 21456..25517), SARS-CoV accession # AAP30030.1 (coded by AY278488.2, region: 21473..25240), Pangolin coronavirus accession # QIA48614.1(coded by MT040333.1, region: 21540..25343), and Bat CoV RaTG13 accession # QHR63300.2 (coded by MN996532.2, region: 21560..25369). Similar study was conducted for other structural proteins (envelop protein, membrane protein, nucleocapsid protein) and ORF8. The aligned view of each of these helped to understand evolutionary relationship and identify conserved residues (Figures 4, 5, 6 and 7). Later it was used as reference to understand whether the mutated amino acid residues of the proteins studied here from patient samples are evolutionary conserved.

For envelop (E) protein sequence alignment was conducted between MERS-CoV accession # AFS88941.1 (coded by JX869059.2, region: 27590..27838), SARS-CoV accession # AAP30033.1 (coded by AY278488.2, region: 26098..26328), Pangolin coronavirus accession # QIG55947.1 (coded by MT121216.1, region: 26079..26306), Bat coronavirus RaTG13 accession # QHR63302.1 (coded by MN996532.2, region: 26230..26457), SARS-CoV2 REFSEQ accession # YP_009724392.1 (coded by NC_045512.2, region: 26245..26472).

Similarly, for membrane protein alignment was conducted between MERS-CoV accession # AFS88942.1 (coded by JX869059.2, region: 27853..28512), SARS-CoV accession # AAP30034.1 (coded by AY278488.2, region: 26379..27044), Pangolin coronavirus accession # QIG55948.1 (coded by MT121216.1, region: 26357..27025), Bat coronavirus RaTG13 accession # QHR63303.1 (coded by MN996532.2, region: 26508..27173), SARS-CoV2 REFSEQ accession # YP_009724393.1 (coded by NC_045512.2, region: 26523..27191). Alignment view has not been provided for M (membrane) protein.

Sequence alignment was conducted for ORF8 between MERS-CoV accession # AFV09328.1 (coded by JX869059.2, region: 28762..29100), SARS-CoV accession # AAP30036.1 (coded by AY278488.2, region: 27254..27622), Pangolin coronavirus accession # QIG55952.1 (coded by MT121216.1, region: 27728..28045), Bat coronavirus RaTG13 accession # QHR63307.1 (coded by MN996532.2, region: 27875..28240), SARS-CoV2 REFSEQ accession # YP_009724396.1 (coded by NC_045512.2, region: 27894..28259).

For nucleocapsid protein, sequence alignment was conducted between MERS-CoV accession # AFS88943.1(coded by JX869059.2, region: 28566..29807), SARS-CoV accession # AAP30037.1(coded by AY278488.2, region: 28101..29369), Pangolin coronavirus accession # QIG55953.1(coded by MT121216.1, region: 28060..29319), Bat coronavirus RaTG13 accession # QHR63308.1(coded by MN996532.2, region: 28255..29514), SARS-CoV2 REFSEQ accession # YP_009724397.2 (coded by NC_045512.2, region: :28274..29533).

Further, the amino acid residues and regions which had undergone mutations in sequences from patient samples are highlighted in Figures 4,5,6 and 7 for S, E, M, N and ORF8 respectively. Amino acid residues were highlighted considering three different scenarios 1) mutations for evolutionary conserved residues with high statistical significance i.e. p -value is 0.05 or more (green colour Figures 4,5,6 and 7), 2) mutations for evolutionary conserved residues with p -value in less than 0.05 (turquoise colour Figures 4,5,6 and 7), 3) mutations for evolutionary less- conserved or non-conserved residues (grey colour Figures 4,5,6 and 7).

The occurrence of mutation has been measured manually with respect to reference sequence.

Mutational rate has been calculated by counting total number of missense mutations (variants) in all structural proteins (Spike, Nucleocapsid, Envelope and Membrane proteins) and ORF8 (considered in this study) in each genome followed by dividing the total number of genome sequences undergone mutation.

Mutation Rate = (Number of unique mutations / Number of SARS-Cov2 sequences) X 1000

The regions or countries in which sufficient genome sequences (minimum 10 and maximum 60 genome sequences) were not available during this study has been excluded.

Statistical Significance of the variants has been checked by setting p-value as 0.05 or 5 percent. It implies the confidence level is 95 percent. Anything below 0.05 is considered statistically less significant.

Disease transmission rate of each country was calculated by the number of infected persons per one million people (sourced from https://www.worldometers.info/coronavirus/) multiplied by hundred.

Fatality rate of each country was measured using cumulative death (https://covid19.who.int/) divided by total number of affected cases and then multiplied by hundred.

Results

Several different missense mutations have been observed in all structural proteins (except membrane protein, M) and ORF8 protein across the globe. These are recorded in tables (Table 1 for S protein, Table 2 for N protein, Table 3 for E protein and Table 4 for ORF8 protein, supplementary Table 5 for the mutated sequence details). There are ~150 sequences within the seven hundred sequences studied here did not show any mutation. These are excluded from supplementary Table 5.

Mutation →	L5F	S12F	R21T	Y28H	I39V	H49Y	L54F	T76I	R78M	Y145Stop	H146Y/R	W152L	S162I	L176F	N189D	I197V	G261D	A288T	Q314R	R407I	Y452F	A522V	T572I	E583D	D614G	V622F	Q677H	S884F	A930V	T940A	V1122L	G1124V	R1185H	M1237I
Country ↓																																		
Australia	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-
Bangladesh	+	-	-	-	+	+	-	+	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	+	-
Brazil	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-
Chile	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-
China	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-
Czech Republic	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-
Denmark	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	+	-	-	-	-	-	-	-	-	-
Egypt	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-
France	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-
Germany	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-
Greece	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	+	-	-	-
India	+	-	-	+	-	-	+	-	+	+	-	+	+	-	-	-	-	-	-	+	-	-	+	+	+	-	+	-	+	-	-	+	-	-
Iran	-	-	-	-	-	-	-	-	-	-	+	-	-	-	+	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-
Iraq	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	+	-	-	-	-	-	-	-	-	-
Italy	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-
Jamaica	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-
Japan	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-
Jordan	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-
Mexico	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-
Morocco	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-
Netherland	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	+	-	-	-	+	-	-	-	-	-	-	-	-	-
Philippines	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-
Poland	-	-	-	-	-	-	-	-	-	-	-	-	-	+		-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-
Puerto Rico	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-
Russia	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-
Spain	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-
Sri Lanka	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-
Taiwan	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	+	-	-	-	-	-	-
Tunesia	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-
UK	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-
USA	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-	-	-	-	+

Table 1: The different types of missense mutations in S (Spike) protein for different countries (vertically) are placed in tabular form. Single letter amino acid code has been used and placed horizontally in table; + is used for presence and - is used for absence of respective mutation

Mutation→	P6T	A12G	P13L	D22Y/N	S33I	G35W	P67S	T121I	A134V	L139F	R149L	S180T	S183Y	S186F	S194L	S197L	S202N	R203K	G204R	T205I	S235F	T282I	P344S	D347H	D348Y	D371Y
Country↓																										
Bangladesh	-	-	-	-	-	-	-	-	-	-	-	+		-	-	-	-	+	+	-	-	+	-	-	-	-
Brazil	-	-	-	-	-	-	-	-	-	-	-	-		-	-	-	-	+	+	-	-	-	-	-	-	-
Chile	-	-	-	-	-	-	-	-	-	-	-	-		-	-	+	-	+	+	-	-	-	-	-	-	-
China	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-	-	-	-
Czech Republic	-	-	-	-	-	-		-	-	-	-	-		-	-	-	-	+	+	-	-	-	-	-	-	-
Denmark	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-	-	-	-
Germany	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-	-	-	-
Greece	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-	-	-	-
India	+	-	+	+	+	-	-	-	+	+		-	-		+	-	+	+	+	-	-	-	+	-	+	-
Iran	-	-	-	-	-	-	-	-	-	-	+	-	-	+	+		-	-	-	-	-	-	-	-	-	-
Iraq	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-
Israel	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-	-	-	-
Italy	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-	-	-	-
Jamaica	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Japan	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-	-	-	-
Mexico	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-
Morocco	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-
Philippines	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-	-	-	-
Poland	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-	-	-	-
Puerto Rico	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-	-	-	-
Russia	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-	-	-	-
Russia	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-	-	-	-
Russia	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-	-	-	-
Russia	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-	-	-	-
Sri Lanka	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-	-	-	-
Taiwan	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-	-	-	-
Tunesia	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-
UK	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-	-	-	-
USA	-	-	-	-	-	+	+	-	-	-	-	-	+	-	-	-	-	+	+	-	+	-	-	-	-	+

Table 2: The different types of missense mutations in N (nucleocapsid) protein for different countries (vertically) are placed in tabular form. Single letter amino acid code has been used and placed horizontally in table; + is used for presence and - is used for absence of respective mutation

Mutation→	F3I	A14S	S24L	P38R	I39V	V62L	A65V	L84S
Country ↓								
Bangladesh	-	-	-	+	+	-	-	-
India	-	+	-	-	-	-	-	+
USA	+	-	+	-	-	+	+	+
China	-	-	-	-	-	-	-	+
Ghana	-	-	-	-	-	-	-	+
Taiwan	-	-	+	-		+	-	-
Italy	-	-	-	-	-	-	-	+

Table 3: The different types of missense mutations in ORF8 protein for different countries are placed in tabular form. Single letter amino acid code has been used; + is used for presence and - is used for absence of respective mutation

Mutation→	V5A	L21F	I33T	L49M	V62F
Country↓	-	-	-	-	-
Brazil	+	-	-	-	-
Philippines	-	+	-	-	-
Greece	-	-	-	+	-
India	-	-	-	-	+
Iran	-	-	+	-	-

Table 4: The different types of missense mutations in E (envelope) protein for different countries are placed in tabular form. Single letter amino acid code has been used; + is used for presence and - is used for absence of respective mutation

Many of these amino acid residues are evolutionary conserved (Figures 4,5,6 and 7) through bat coronavirus RaTG13, pangolin coronavirus, human SARS-CoV, MERS-CoV and SARS-Cov2, some of these amino acid residues are evolutionary conserved (Figures 4,5,6 and 7) through RaTG13, pangolin coronavirus, and human SARS-Cov2, and again quite few regions (amino acid residues) are evolutionary conserved (Figures 4,5,6 and 7) through RaTG13, pangolin coronavirus, human SARS-CoV, and SARS-Cov2 as observed by multiple sequence alignment. This study classifies the mutations in 3 different cases 1) mutations for evolutionary conserved residues with high statistical significance i.e. p -value is 0.05 or more (green colour Figures 1,2,3 and 4), 2) mutations for evolutionary conserved residues with p -value in less than 0.05 (turquoise colour Figures 1,2,3 and 4), 3) mutations for evolutionary less- conserved or non-conserved residues (grey colour Figures 1,2,3 and 4).

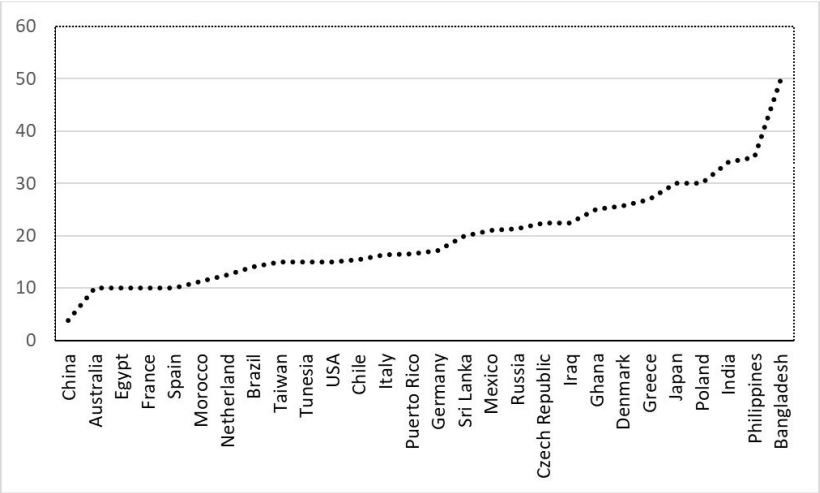


Figure 1: Mutational rate has been plotted with respect to country. It has been calculated by counting number of missense mutations in all structural proteins and ORF8 in each followed by dividing the total number of genome sequences undergone mutation. We excluded regions in which sufficient genome sequences (minimum 10 and maximum 60 genome sequences) are not available during this study

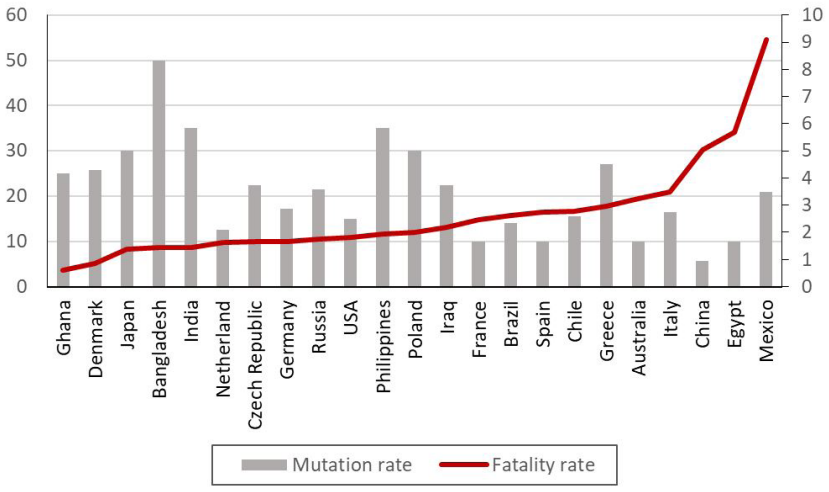


Figure 2: Mutational and fatality rates have been plotted. Fatality rate has been calculated using cumulative death (https://covid19.who.int/) divided by total number of affected cases and then multiplied by hundred

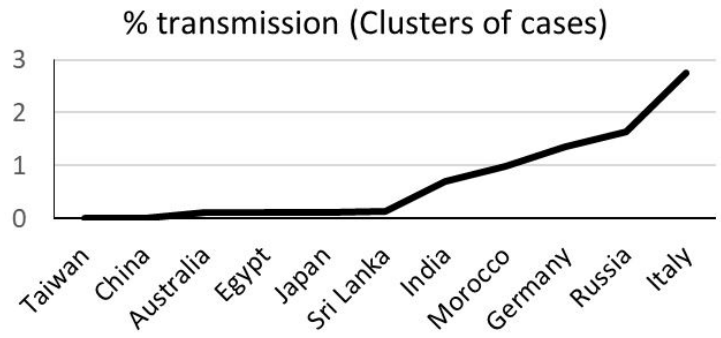


Figure 3a: Disease transmission rate has been plotted with respect to countries where transmission type is clusters of cases

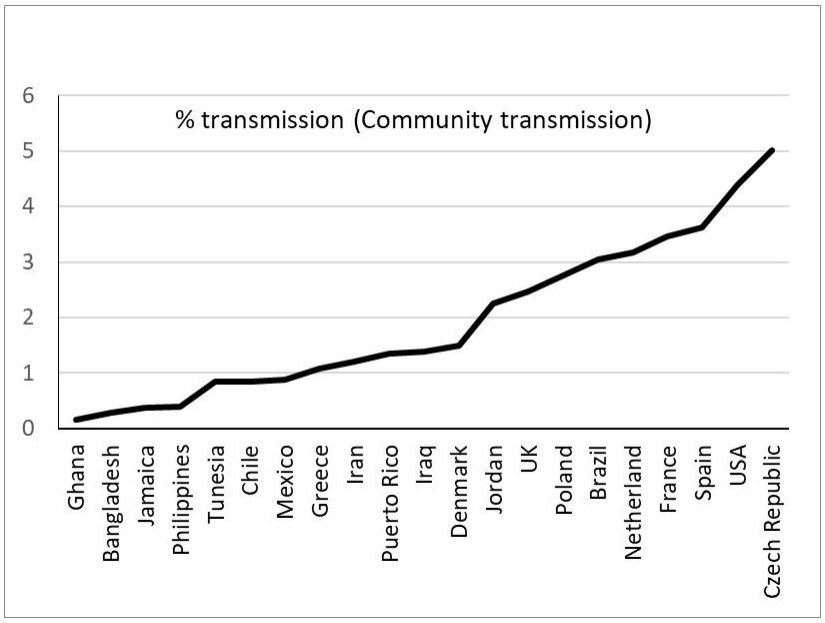


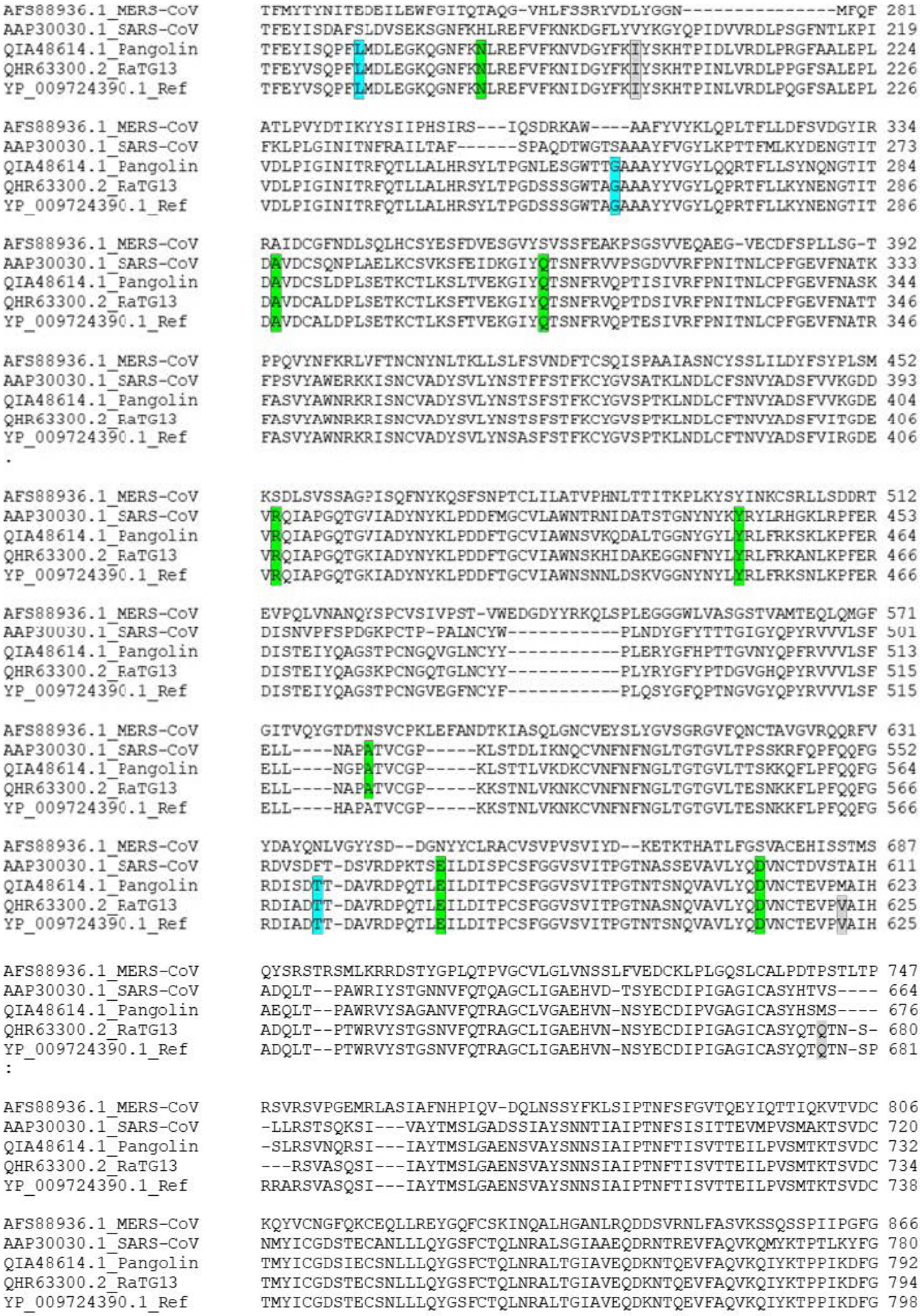
Figure 3b: Disease transmission rate has been plotted with respect to countries where transmission type is community transmission

```
AFS88936.1_MERS-CoV      MIHSVFLLMFLLTPTESYVDVGPDSVKSACIEVDIQQTFFDKTWPRPIDVSKADGIIYPQ 60
AAP30030.1_SARS-CoV      ----MFIFLLFLTL-TSG------SDLDRCTTFDDVQA-----PNYTQHTSSMRGVYYPD 44
QIA48614.1_Pangolin      ----MFVFLFVLPL-VSS------Q----CVNLTTRTG-----IPPGYTNSSTRGVYYPD 40
QHR63300.2_RaTG13        ----MFVFLVLLPL-VSS------Q----CVNLTTRTQ-----LPPAYTNSSTRGVYYPD 40
YP_009724390.1_Ref       ----MFVFLVLLPL-VSS------Q----CVNLTTRTQ-----LPPAYTNSFTRGVYYPD 40

AFS88936.1_MERS-CoV      GRTYSNITITYQGLF-PYQGDHGDMYVYSAGHATGTTPQKLFVANYSQDVKQFANGFVVR 119
AAP30030.1_SARS-CoV      EIFRSDTLYLTQDLFLPFYSNVT---GFHTIN-------HT----FDNPVIPFKDGIYFA 90
QIA48614.1_Pangolin      KVFRSSILHLTQDLFLPFFSNVT---WFNTINYQG--GFKK----FDNPVLPFNDGVYFA 91
QHR63300.2_RaTG13        KVFRSSVLHLTQDLFLPFFSNVT---WFHAIHVSGTNGIKR----FDNPVLPFNDGVYFA 93
YP_009724390.1_Ref       KVFRSSVLHSTQDLFLPFFSNVT---WFHAIHVSGTNGTKR----FDNPVLPFNDGVYFA 93

AFS88936.1_MERS-CoV      IGAAANSTGTVIISPSTSATIRKIYPAFMLGSSVGNFSDGKMGRFFNHTLVLLPDGCGTL 179
AAP30030.1_SARS-CoV      ATE-----------------KSNVVRGWVFGSTMNNKSQ---------SVIIINNSTNVV 124
QIA48614.1_Pangolin      STE-----------------KSNIIRGWIFGTTLDARTQ---------SLLIVNNATNVV 125
QHR63300.2_RaTG13        STE-----------------KSNIIRGWIFGTTLDSKTQ---------SLLIVNNATNVV 127
YP_009724390.1_Ref       STE-----------------KSNIIRGWIFGTTLDSKTQ---------SLLIVNNATNVV 127

AFS88936.1_MERS-CoV      LRAF--YCILEPRSGNHCPAGNSYTSFATYHTPATDCSDGNYNRNASLNSFKEYFNLRNC 237
AAP30030.1_SARS-CoV      IRACNFELCDNPFFAVSKPMGT-----------------------QTHTMIFDNAFNC 159
QIA48614.1_Pangolin      IKVCEFQFCTDPFLGVYYHNNN--------------------KTWVENEFRVYSSANNC 164
QHR63300.2_RaTG13        IKVCEFQFCNDPFLGVYYHKNN--------------------KSWMESEFRVYSSANNC 166
YP_009724390.1_Ref       IKVCEFQFCNDPFLGVYYHKNN--------------------KSWMESEFRVYSSANNC 166
```


AFS88936.1_MERS-CoV TFMYTYNITEDEILEWFGITQTAQG-VHLFSSRYVDLYGGN---------------MFQF 281
AAP30030.1_SARS-CoV TFEYISDAFSLDVSEKSGNFKHLREFVFKNKDGFLYVYKGYQPIDVVRDLPSGFNTLKPI 219
QIA48614.1_Pangolin TFEYISQPFLMDLEGKQGNFKNLREFVFKNVDGYFKIYSKHTPIDLVRDLPRGFAALEPL 224
QHR63300.2_RaTG13 TFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRDLPPGFSALEPL 226
YP_009724390.1_Ref TFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRDLPQGFSALEPL 226
AFS88936.1_MERS-CoV ATLPVYDTIKYYSIIPHSIRS---IQSDRKAW----AAFYVYKLQPLTFLLDFSVDGYIR 334
AAP30030.1_SARS-CoV FKLPLGINITNFRAILTAF------SPAQDTWGTSAAAYFVGYLKPTTFMLKYDENGTIT 273
QIA48614.1_Pangolin VDLPIGINITRFQTLLALHRSYLTPGNLESGWTTGAAAYYVGYLQQRTFLLSYNQNGTIT 284
QHR63300.2_RaTG13 VDLPIGINITRFQTLLALHRSYLTPGDSSSGWTAGAAAYYVGYLQPRTFLLKYNENGTIT 286
YP_009724390.1_Ref VDLPIGINITRFQTLLALHRSYLTPGDSSSGWTAGAAAYYVGYLQPRTFLLKYNENGTIT 286
AFS88936.1_MERS-CoV RAIDCGFNDLSQLHCSYESFDVESGVYSVSSFEAKPSGSVVEQAEG-VECDFSPLLSG-T 392
AAP30030.1_SARS-CoV DAVDCSQNPLAELKCSVKSFEIDKGIYQTSNFRVVPSGDVVRFPNITNLCPFGEVFNATK 333
QIA48614.1_Pangolin DAVDCSLDPLSETKCTLKSLTVEKGIYQTSNFRVQPTISIVRFPNITNLCPFGEVFNASK 344
QHR63300.2_RaTG13 DAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTDSIVRFPNITNLCPFGEVFNATT 346
YP_009724390.1_Ref DAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATR 346
AFS88936.1_MERS-CoV PPQVYNFKRLVFTNCNYNLTKLLSLFSVNDFTCSQISPAAIASNCYSSLILDYFSYPLSM 452
AAP30030.1_SARS-CoV FPSVYAWERKKISNCVADYSVLYNSTFFSTFKCYGVSATKLNDLCFSNVYADSFVVKGDD 393
QIA48614.1_Pangolin FASVYAWNRKRISNCVADYSVLYNSTSFSTFKCYGVSPTKLNDLCFTNVYADSFVVKGDE 404
QHR63300.2_RaTG13 FASVYAWNRKRISNCVADYSVLYNSTSFSTFKCYGVSPTKLNDLCFTNVYADSFVITGDE 406
YP_009724390.1_Ref FASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDE 406
.
AFS88936.1_MERS-CoV KSDLSVSSAGPISQFNYKQSFSNPTCLILATVPHNLTTITKPLKYSYINKCSRLLSDDRT 512
AAP30030.1_SARS-CoV VRQIAPGQTGVIADYNYKLPDDFMGCVLAWNTRNIDATSTGNYNYKYRYLRHGKLRPFER 453
QIA48614.1_Pangolin VRQIAPGQTGVIADYNYKLPDDFTGCVIAWNSVKQDALTGGNYGYLYRLFRKSKLKPFER 464
QHR63300.2_RaTG13 VRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSKHIDAKEGGNFNYLYRLFRKANLKPFER 466
YP_009724390.1_Ref VRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFER 466
AFS88936.1_MERS-CoV EVPQLVNANQYSPCVSIVPST-VWEDGDYYRKQLSPLEGGGWLVASGSTVAMTEQLQMGF 571
AAP30030.1_SARS-CoV DISNVPFSPDGKPCTP-PALNCYW----------PLNDYGFYTTTGIGYQPYRVVVLSF 501
QIA48614.1_Pangolin DISTEIYQAGSTPCNGQVGLNCYY----------PLERYGFHPTTGVNYQPFRVVVLSF 513
QHR63300.2_RaTG13 DISTEIYQAGSKPCNGQTGLNCYY----------PLYRYGFYPTDGVGHQPYRVVVLSF 515
YP_009724390.1_Ref DISTEIYQAGSTPCNGVEGFNCYF----------PLQSYGFQPTNGVGYQPYRVVVLSF 515
AFS88936.1_MERS-CoV GITVQYGTDTNSVCPKLEFANDTKIASQLGNCVEYSLYGVSGRGVFQNCTAVGVRQQRFV 631
AAP30030.1_SARS-CoV ELL----NAPATVCGP-----KLSTDLIKNQCVNFNFNGLTGTGVLTPSSKRFQPFQQFG 552
QIA48614.1_Pangolin ELL----NGPATVCGP-----KLSTTLVKDKCVNFNFNGLTGTGVLTTSKKQFLPFQQFG 564
QHR63300.2_RaTG13 ELL----NAPATVCGP-----KKSTNLVKNKCVNFNFNGLTGTGVLTESNKKFLPFQQFG 566
YP_009724390.1_Ref ELL----HAPATVCGP-----KKSTNLVKNKCVNFNFNGLTGTGVLTESNKKFLPFQQFG 566
AFS88936.1_MERS-CoV YDAYQNLVGYYSD--DGNYYCLRACVSVPVSVIYD--KETKTHATLFGSVACEHISSTMS 687
AAP30030.1_SARS-CoV RDVSDFT-DSVRDPKTSEILDISPCSFGGVSVITPGTNASSEVAVLYQDVNCTDVSTAIH 611
QIA48614.1_Pangolin RDISDTT-DAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQDVNCTEVPMAIH 623
QHR63300.2_RaTG13 RDIADTT-DAVRDPQTLEILDITPCSFGGVSVITPGTNASNQVAVLYQDVNCTEVPVAIH 625
YP_009724390.1_Ref RDIADTT-DAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQDVNCTEVPVAIH 625
AFS88936.1_MERS-CoV QYSRSTRSMLKRRDSTYGPLQTPVGCVLGLVNSSLFVEDCKLPLGQSLCALPDTPSTLTP 747
AAP30030.1_SARS-CoV ADQLT--PAWRIYSTGNNVFQTQAGCLIGAEHVD-TSYECDIPIGAGICASYHTVS---- 664
QIA48614.1_Pangolin AEQLT--PAWRVYSAGANVFQTRAGCLVGAEHVN-NSYECDIPVGAGICASYHSMS---- 676
QHR63300.2_RaTG13 ADQLT--PTWRVYSTGSNVFQTRAGCLIGAEHVN-NSYECDIPIGAGICASYQTQTN-S- 680
YP_009724390.1_Ref ADQLT--PTWRVYSTGSNVFQTRAGCLIGAEHVN-NSYECDIPIGAGICASYQTQTN-SP 681
:
AFS88936.1_MERS-CoV RSVRSVPGEMRLASIAFNHPIQV-DQLNSSYFKLSIPTNFSFGVTQEYIQTTIQKVTVDC 806
AAP30030.1_SARS-CoV -LLRSTSQKSI---VAYTMSLGADSSIAYSNNTIAIPTNFSISITTEVMPVSMAKTSVDC 720
QIA48614.1_Pangolin -SLRSVNQRSI---IAYTMSLGAENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDC 732
QHR63300.2_RaTG13 ---RSVASQSI---IAYTMSLGAENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDC 734
YP_009724390.1_Ref RRARSVASQSI---IAYTMSLGAENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDC 738
AFS88936.1_MERS-CoV KQYVCNGFQKCEQLLREYGQFCSKINQALHGANLRQDDSVRNLFASVKSSQSSPIIPGFG 866
AAP30030.1_SARS-CoV NMYICGDSTECANLLLQYGSFCTQLNRALSGIAAEQDRNTREVFAQVKQMYKTPTLKYFG 780
QIA48614.1_Pangolin TMYICGDSIECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFG 792
QHR63300.2_RaTG13 TMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFG 794
YP_009724390.1_Ref TMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFG 798

```
AFS88936.1_MERS-CoV      GDFNLTLLEPVSISTGSRSARSAIEDLLFDKVTIADPGYMQGYDDCMQQGPASARDLICA 926
AAP30030.1_SARS-CoV      GF-NFSQILPDP---LKPTKRSFIEDLLFNKVTLADAGFMKQYGECL--GDINARDLICA 834
QIA48614.1_Pangolin      GF-NFSQILPDP---SKPSKRSFIEDLLFNKVTLADAGFIKQYGDCL--GDIAARDLICA 846
QHR63300.2_RaTG13        GF-NFSQILPDP---SKPSKRSFIEDLLFNKVTLADAGFIKQYGDCL--GDIAARDLICA 848
YP_009724390.1_Ref       GF-NFSQILPDP---SKPSKRSFIEDLLFNKVTLADAGFIKQYGDCL--GDIAARDLICA 852

AFS88936.1_MERS-CoV      QYVAGYKVLPPLMDVNMEAAYTSSLLGSIAGVGWTAGLSSFAAIPFAQSIFYRLNGVGIT 986
AAP30030.1_SARS-CoV      QKFNGLTVLPPLLTDDMIAAYTAALVSGTATAGWTFGAGAALQIPFAMQMAYRFNGIGVT 894
QIA48614.1_Pangolin      QKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVT 906
QHR63300.2_RaTG13        QKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVT 908
YP_009724390.1_Ref       QKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVT 912

AFS88936.1_MERS-CoV      QQVLSENQKLIANKFNQALGAMQTGFTTTNEAFQKVQDAVNNNAQALSKLASELSNTFGA 1046
AAP30030.1_SARS-CoV      QNVLYENQKQIANQFNKAISQIQESLTTTSTALGKLQDVVNQNAQALNTLVKQLSSNFGA 954
QIA48614.1_Pangolin      QNVLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVVNQNAQALNTLVKQLSSNFGA 966
QHR63300.2_RaTG13        QNVLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVVNQNAQALNTLVKQLSSNFGA 968
YP_009724390.1_Ref       QNVLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVVNQNAQALNTLVKQLSSNFGA 972

AFS88936.1_MERS-CoV      ISASIGDIIQRLDVLEQDAQIDRLINGRLTTLNAFVAQQLVRSESAALSAQLAKDKVNEC 1106
AAP30030.1_SARS-CoV      ISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSEC 1014
QIA48614.1_Pangolin      ISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSEC 1026
QHR63300.2_RaTG13        ISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSEC 1028
YP_009724390.1_Ref       ISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSEC 1032

AFS88936.1_MERS-CoV      VKAQSKRSGFCGQGTHIVSFVVNAPNGLYFMHVGYYPSNHIEVVSAYGLCDAANPTNCIA 1166
AAP30030.1_SARS-CoV      VLGQSKRVDFCGKGYHLMSFPQAAPHGVVFLHVTYVPSQERNFTTAPAICHEGKA---YF 1071
QIA48614.1_Pangolin      VLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHEGKA---HF 1083
QHR63300.2_RaTG13        VLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKA---HF 1085
YP_009724390.1_Ref       VLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKA---HF 1089

AFS88936.1_MERS-CoV      PVNGYFIKTNNTRIVDEWSYTGSSFYAPEPITSLNTKYVAPQVTYQN-ISTNLPPPLLGN 1225
AAP30030.1_SARS-CoV      PREGVFVFN-----GTSWFITQRNFFSPQIITTDNTFVSGNCDVVIGIINNTVYDPLQ-- 1124
QIA48614.1_Pangolin      PREGVFVSN-----GTHWFITQRNFYEPQIITTDNTFVSGSCDVVIGIVNNTVYDPLQ-- 1136
QHR63300.2_RaTG13        PREGVFVSN-----GTHWFVTQRNFYEPQIITTDNTFVSGSCDVVIGIVNNTVYDPLQ-- 1138
YP_009724390.1_Ref       PREGVFVSN-----GTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQ-- 1142

AFS88936.1_MERS-CoV      STGIDFQDELDEFFKNVSTSIPNFGSLTQINTTLLDLTYEMLSLQQVVKALNESYIDLKE 1285
AAP30030.1_SARS-CoV      PELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQE 1184
QIA48614.1_Pangolin      PELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESPIDLQE 1196
QHR63300.2_RaTG13        PELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQE 1198
YP_009724390.1_Ref       PELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQE 1202

AFS88936.1_MERS-CoV      LGNYTYYNKWPWYIWLGFIAGLVALALCVFFILCCTGCGTNCMGKLKCNRCCDRYEEYDL 1345
AAP30030.1_SARS-CoV      LGKYEQYIKWPWYVWLGFIAGLIAIVMVTILLCCMTSCCSCLKGACSCGSCCKF-DEDDS 1243
QIA48614.1_Pangolin      LGKYEQYIKWPWYIWLGFIAGLIAIIMVTIMLCCMTSCCSCLKGCCSCGSCCKF-DEDDS 1255
QHR63300.2_RaTG13        LGKYEQYIKWPWYIWLGFIAGLIAIIMVTIMLCCMTSCCSCLKGCCSCGSCCKF-DEDDS 1257
YP_009724390.1_Ref       LGKYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKF-DEDDS 1261

AFS88936.1_MERS-CoV      EPHKVHVH---- 1353
AAP30030.1_SARS-CoV      EPVLKGVKLHYT 1255
QIA48614.1_Pangolin      EPVLKGVKLHYT 1267
QHR63300.2_RaTG13        EPVLKGVKLHYT 1269
YP_009724390.1_Ref       EPVLKGVKLHYT 1273
```

Figure 4: Multiple sequence alignment between S proteins of SARS-CoV2 accession # YP_009724390.1, MERS-CoV accession # AFS88936.1, SARS-CoV accession # AAP30030.1, Pangolin coronavirus accession # QIA48614.1, and Bat CoV RaTG13 accession # QHR63300. Amino acid residues which had undergone mutations (as shown in table 1) were highlighted as follows 1) mutations for evolutionary conserved residues with high statistical significance i.e. p -value is 0.05 or more (green colour), 2) mutations for evolutionary conserved residues with p -value in less than 0.05 (turquoise colour), 3) mutations for evolutionary less- conserved or non-conserved residues (grey colour)

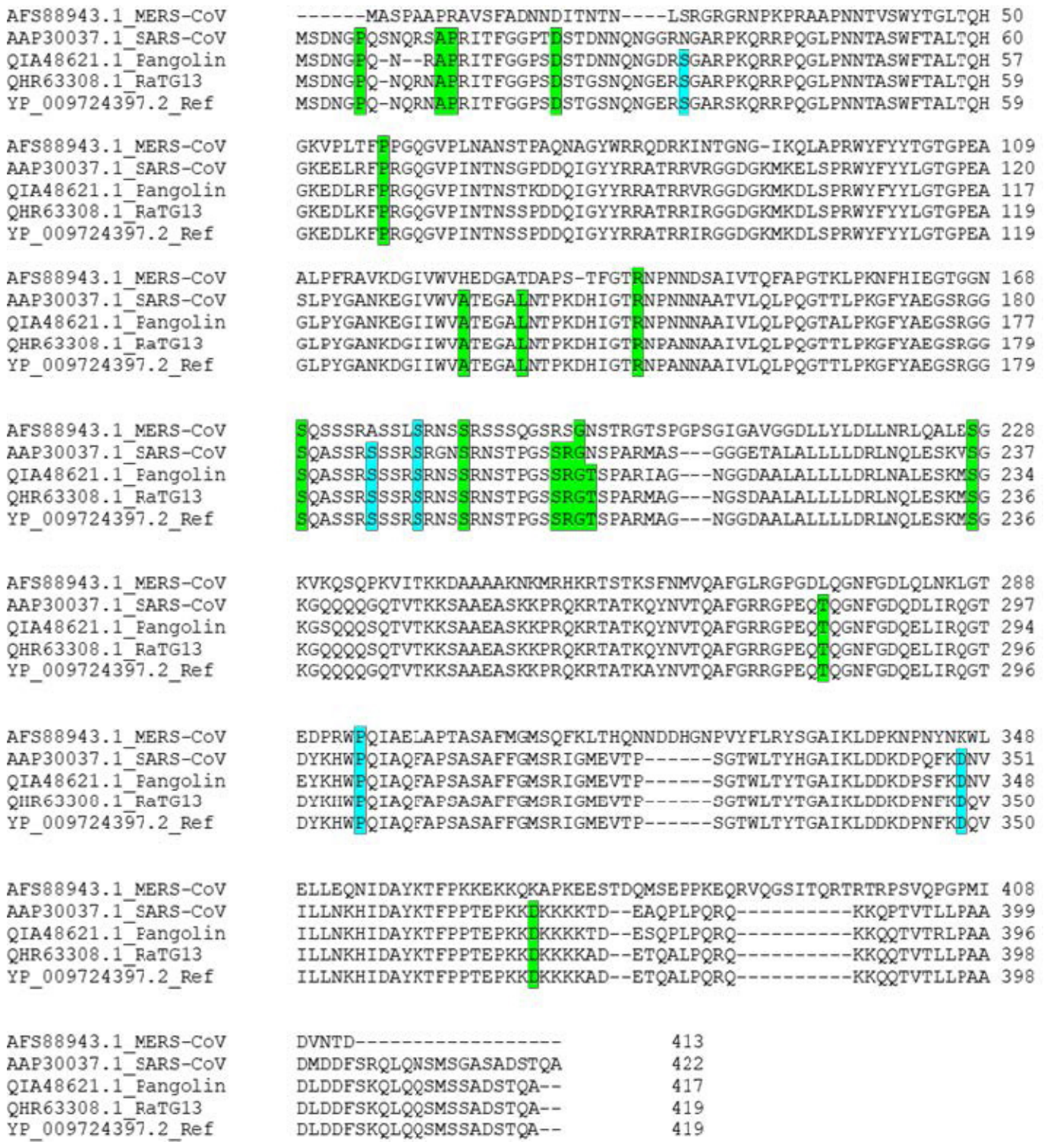


Figure 5: Multiple sequence alignment between N (nucleocapsid) proteins MERS-CoV accession # AFS88943.1, SARS-CoV accession # AAP30037.1, Pangolin coronavirus accession # QIG55953.1, Bat coronavirus RaTG13 accession # QHR63308.1, SARS-CoV2 REFSEQ accession # YP_009724397.2. Amino acid residues which had undergone mutations (as shown in table 2) were highlighted as follows 1) mutations for evolutionary conserved residues with high statistical significance i.e. p -value is 0.05 or more (green colour), 2) mutations for evolutionary conserved residues with p -value in less than 0.05 (turquoise colour), 3) mutations for evolutionary less- conserved or non-conserved residues (grey colour)

Figure 6: Multiple sequence alignment between ORF8 proteins between MERS-CoV accession # AFV09328.1, SARS-CoV accession # AAP30036.1, Pangolin coronavirus accession # QIG55952.1, Bat coronavirus RaTG13 accession # QHR63307.1, SARS-CoV2 REFSEQ accession # YP_009724396.1. Amino acid residues which had undergone mutations (as shown in table 3) were highlighted as follows 1) mutations for evolutionary conserved residues with high statistical significance i.e. p -value is 0.05 or more (green colour), 2) mutations for evolutionary conserved residues with p -value in less than 0.05 (turquoise colour), 3) mutations for evolutionary less- conserved or non-conserved residues (grey colour)

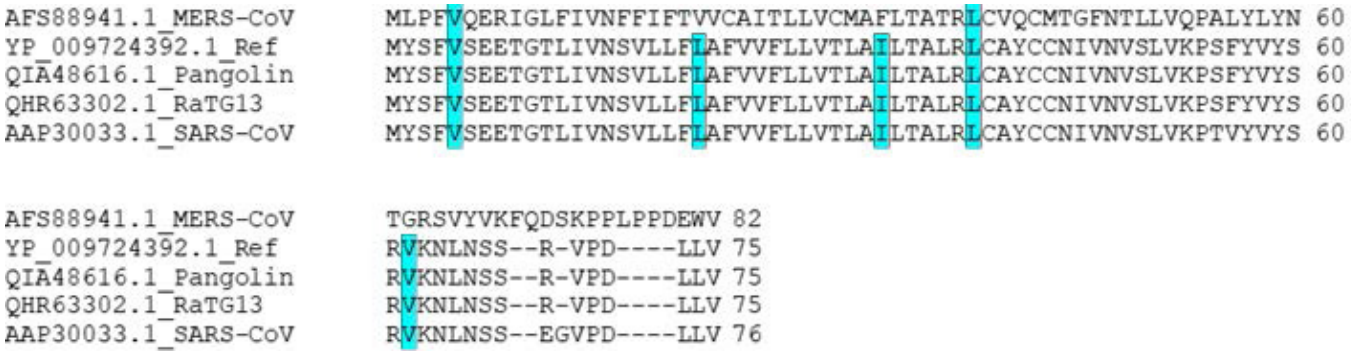


Figure 7: Multiple sequence alignment between E (envelop) proteins between MERS-CoV accession # AFS88941.1, SARS-CoV accession # AAP30033.1, Pangolin coronavirus accession # QIG55947.1, Bat coronavirus RaTG13 accession # QHR63302.1, SARS-CoV2 REFSEQ accession # YP_009724392.1. Amino acid residues which had undergone mutations (as shown in table 4) were highlighted as follows 1) mutations for evolutionary conserved residues with high statistical significance i.e. p -value is 0.05 or more (green colour), 2) mutations for evolutionary conserved residues with p -value in less than 0.05 (turquoise colour), 3) mutations for evolutionary less- conserved or non-conserved residues (grey colour)

S protein contains maximum number of different missense mutations (Table 1). All the countries studied here, have common D614G mutation for S protein. Similarly, N protein has few mutations which are common in many countries (Table 2). ORF8 (Table 3) and E protein (Table 4) have very few mutations across the globe. M protein doesn't have any mutation for the genome sequences studied here.

L5F substitution in S protein was found in some genome sequences from the patient samples of Bangladesh, India, Italy (very few), Philippines and USA. This (L5F mutation) is mostly coupled with D614G substitution. In addition, L54F substitution was found in few sequences of patient samples from India which is coupled with D614G mutation.

For N (nucleocapsid) protein, S194L, S202N, R203G, G204R mutations were abundantly found in sequences from multiple countries. Again, like S protein, India has maximum number of different types of substitution in N protein (Table 2). T205I mutation is only noted in all N protein sequences from Iraq. P67S or S235F are found in many sequences (N protein) from USA.

For ORF8, very few variants were found in the sequences studied here. USA has maximum types of missense mutations in ORF8 (table 3). L84S mutation is common among the sequences of patient samples from China, Ghana, India, Italy, Tunisia and USA.

This study reports very few mutations in E protein (Table 4). It did not find any common variants from multiple countries.

Next, mutation rate was calculated and plotted with respect to different countries (Figure 1). It shows that Bangladesh, India and Philippines have the highest mutational frequency whereas Australia, Italy, Egypt have least.

Disease transmission rate was calculated and plotted against countries (Figures 3a and b). Two disease transmission types i.e. Community transmission and cluster of cases, are plotted separately. Italy shows highest transmission rate for countries in which transmission type is cluster of cases. Czech Republic has highest transmission rate among countries with community transmission.

Then fatality rate was calculated using WHO cumulative death (https://covid19.who.int/) and plotted against mutational rate. This study reports that they are inversely related (Figure 2) for the countries with community transmission.

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. This study reports certain amino acid variations in structural proteins and ORF8 and its possible implication for altering structure, function, infectivity and fatality.

Coronavirus entry into host cells is mediated by the transmembrane spike (S) protein that forms homotrimers protruding from the viral surface [7]. S protein 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, known as furin cleavage site and 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 [8]. This study reports that amino acid at position 614 (a residue close to furin cleavage site) underwent mutation (D614G) which is common across all geographies. Earlier study reported that 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 [9]. There are several other miss-sense mutations in S protein (Table 1). All these mutations could be classified as stabilizing and destabilizing based on the free-energy changes. L5F mutation in sequences from Bangladesh, India, Italy (very few), Philippines and USA patient samples and L54F mutation in sequences from India, are mostly coupled with D614G substitution. Both of the amino acids (leucine and phenylalanine) 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 [10]. There are multiple residues in receptor binding domain (RBD) which play important role in binding to ACE2 (angiotensin converting enzyme 2) receptor [11]. This study reports multiple point mutations in neighbouring amino acids of ACE2 binding residues. All these potentially play important role for increasing person to person transmission by altering the affinity to ACE2, stability of protein and interaction with neighbouring molecules.

The nucleocapsid (N) protein is an important structural protein for the coronaviruses. It is highly abundant in the viruses. Its function involves, entering into 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 [12]. Three disordered regions were reported on 1) N terminal (residues 1-43), 2) linker peptide, and 3) C terminal end (residues 365-419) [12]. This study reports multiple point mutations in linker region (S194L, S202N, R203K and G204R) and these are found to be common for many countries (Table 2). Earlier report also pointed out that these positions are evolutionary conserved [13]. It was also reported that multiple disordered regions facilitate the N protein to transiently bind to different partners and maintain a correct conformation [12]. The mutated residues reported here, surround with GSK3 phosphorylation 'SRGTS' (amino acid position 202-206) and CDK phosphorylation 'SPAR' (amino acid position 206-209) motifs [14]. Most possibly the substitutions alter the binding affinity to attain the stable conformation and simultaneously contribute towards enhancing transmission rate.

This study reports very few missense substitutions in ORF8. L84S is one of the variants for ORF8 that was found in patient samples of multiple countries such as Ghana, Italy, USA, India and China. Earlier study showed that ORF8 along with N and ORF3b are potent interferon antagonist, in the early stages of SARS-CoV-2 infection [15]. It hinders the host's antiviral response and then benefit virus replication by delaying the release of IFNs [15]. This residue is not evolutionary conserved as shown in figure 6. For L84S, Leu is non-polar hydrophobic whereas Ser is polar amino acid. Ser residue undergoes phosphorylation and possibly play an important role for reducing anti-interferon activity and hence the disease severity.

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. 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) [16]. We found very few genetic variation (L21F, I33T, L49M, V62F) in E protein and none of these substitutions are common worldwide. L21F potentially interfere the oligomerisation of SARS-CoV2 as it was shown earlier that V25F hampers the oligomerisation of SARS-CoV E [16]. V62F mutation in C- terminal domain can potentially interfere the interaction with target proteins, thereby altering the host-cell processes required for viral infection [12].

This analysis does not find any substitution in Membrane (M) protein in the 700 genome sequences from patient samples of different geographical regions. This implies that M protein is comparably stable and supports our earlier study [13].

We noticed D614G for S and S194L, S202N, R203K, G204R for N proteins are common mutations for all most all countries. This study reports one D614G in S, one R203K, G204R in N and two L84S in ORF8 proteins of the sequences (out of 60 genome sequences) from patient samples of China.

While looking into the total affected cases and fatality rate from WHO data (https://covid19.who.int/), China shows comparatively less transmission rate (Figure 3a) but more fatality (Figure 2).

To understand the possible implication of these mutations, first we plotted mutational frequency (rate) with respect to different countries (Figure 1). It shows that Bangladesh, India and Philippines have the highest mutation rate. Then we tried to relate mutation rate with fatality rate. Very interestingly, the result shows that Bangladesh, India and Philippines have less fatality whereas Italy, Mexico, France have less mutation but more fatality. In other words, the fatality rate is decreasing with increasing mutation rate or vice versa (Figure 2). So, we hypothesize that disease fatality is inversely proportional to mutation rate. This finding is consistent with earlier studies on NSP6, S protein and RNA-dependent RNA polymerase (RdRP) [17,18].

Secondly, Czech Republic is experiencing community transmission with highest transmission rate followed by USA among the countries studied here (Figure 3b) but the fatality rate is low compared to Mexico, France, Egypt, Spain, Netherland etc (Figure 2). If we try to corelate this finding with mutation rate, it shows that Czech Republic, USA have comparably more mutation and less fatality. For S protein, D614G variant is common in among the countries studied here. This study reports least mutation rate of the sequences from the patient samples of China. As the disease is originated from China, it is probably an indication that the virus is getting mutated after multiple passages.

D614G mutation is abundantly found in patient samples across the globe. It possibly creates favourable environment for enhanced disease transmission. This finding is consistent with earlier report that states a single point mutation in S gene leading to an amino acid substitution at codon 614 from an aspartic acid 614 into glycine (D614G) resulted in greater infectivity compared to the wild type SARS-CoV2 [19].

Conclusion

This study reports several missense mutations for SARS-CoV2 genome sequences studied here of patients from diverse geographical-locations. This study concludes that D614G variant potentially creates favourable environment for rapid disease transmission and disease fatality decreases with increasing mutation rate. Within a very short time frame, the virus evolved rapidly. There are many variations those were evolved within the country as we found quite a few country specific mutations. Although the clinical significance of the observed mutations is not yet available, our findings lay the groundwork to understand the impact of SARS-CoV2 mutations on disease severity. This study also warrants the importance of sequencing the whole genome of SARS-CoV-2 after several passages and key mutations should be used for the effective drug designing and treatment options. All together our findings make us optimistic that the disease severity will diminish as we move along with time and more genomic variations.

Declaration of Interests

The author declares no competing financial interests.

Supplementary

References

1. Farkas C, Fuentes-Villalobos F, Garrido JL, Haigh J, Barría MI (2020) Insights on early mutational events in SARSCoV-2 virus reveal founder effects across geographical regions. PeerJ 8: e9255.

2. Lu R, Zhao X, Li J, Niu P, Yang B (2020) Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet 395: 565-74.

3. Zhu N, Zhang D, Wang W, Li X, Yang B et al. (2020) A Novel Coronavirus from Patients with Pneumonia in China, 2019. New Eng J Med 382: 727-33.

4. Chinazzi M, Davis JT, Ajelli M, Gioannini C, Litvinova M (2020) The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. Science 368: 395-400.

5. Zhang T, Wu Q, Zhang Z (2020) Probable pangolin origin of SARS-CoV-2 associated with the COVID-19 outbreak. Curr Biol 30: 1578.

6. Song Z, Xu Y, Bao L, Zhang L, Yu P, et al. (2020) From SARS to MERS, Thrusting Coronaviruses into the Spotlight. Viruses 11: 59.

7. Tortorici MA, Veesler D (2019) Structural insights into coronavirus entry. Adv Virus Res 105: 93-116.

8. Walls AC, Park Y-J, Tortorici MA, Wall A, McGuire AT, et al. (2020) Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. Cell 181: 281-92.

9. Becerra-Flores M, Cardozo T (2020) SARS-CoV-2 viral spike G614 mutation exhibits higher case fatality rate. Int J Clin Pract 74: e13525.

10. Benvenuto D, Angeletti S, Giovanetti M, Bianchi M, Pascarella S, et al. (2020) Evolutionary analysis of SARS-CoV-2: how mutation of Non-Structural Protein 6 (NSP6) could affect viral autophagy. J Infect 81: e24-e27.

11. Dasgupta R (2020) Comparative genomics of receptor binding domains of S protein and host receptor interaction in COVID-19 patient. Int J Creative Res Thoughts 8: 2571-5.

12. Zeng W, Liu G, Ma H, Zhao D, Yang Y, et al. (2020) Biochemical characterization of SARS-CoV-2 nucleocapsid protein. Biochem Biophys Res Commun 527: 618-23.

13. Dasgupta R (2020) Mutational Analysis Of Structural Proteins In Sars-Cov-2 Viral RNA Of Covid-19 In India. RJLBPCS 6: 10.26479/2020.0605.01.

14. Surjit M, Kumar R, Mishra RN, Reddy MK, Chow VTK, et al. (2005) The severe acute respiratory syndrome coronavirus nucleocapsid protein is phosphorylated localizes in the cytoplasm by 14-3-3-mediated translocation. J Virol 79: 11476-86.

15. Li JY, Liao CH, Wang Q, Tan YJ, Luo R, et al. (2020) The ORF6, ORF8 and nucleocapsid proteins of SARS-CoV-2 inhibit type I interferon signaling pathway. Virus Res 286: 198074.

16. Schoeman D, Fielding BC (2019) Fielding. Coronavirus envelope protein: current knowledge. Virol J 16: 69.

17. Jalali S, Bhadra B, Dasgupta S (2020) Significance of Mutation Rate of Structural and Non-Structural Proteins of SARS-Cov-2 Showed with Lower Death Rate of COVID-19. Sci J Biol 3: 17-22.

18. Patra SK (2020) Perspective on Accelerating the Mutation Rate of SARS-CoV-2for a Better Way of COVID-19 Treatment; Enhanced Mutation Therapy of COVID-19. AIJR Preprints 62: 1-3.

19. Kim S, Lee JH, Lee S, Shim S, Nguyen TT, et al. (2020) The Progression of SARS Coronavirus 2 (SARS-CoV2): Mutation in the Receptor Binding Domain of Spike Gene. Immune Netw 20: e41.