Comparative genomics and diversity of SARS-CoV-2 suggest potential regional virulence

Sundru Manjulata Devi (manjusundru@gmail.com)
Bioinformatics Section, SVR BIOSCIENCE RESEARCH SERVICES, Salboni, West Bengal, India

Annapurna Pamreddy
Division of Nephrology, Department of Medicine, University of Texas Health, Long School of Medicine, San Antonio, Texas, USA

Balakuntalam Kasinath
Division of Nephrology, Department of Medicine, University of Texas Health, Long School of Medicine, San Antonio, Texas, USA

Kumar Sharma (SharmaK3@uthscsa.edu)
Division of Nephrology, Department of Medicine, University of Texas Health, Long School of Medicine, San Antonio, Texas, USA

Research Article

Keywords: SARS-CoV-2, comparative genomics, diversity, mutational analysis, pathogenesis

Posted Date: May 20th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-29557/v1

Abstract

It is widely known fact about the global pandemic caused by Severe Acute Respiratory Syndrome Coronavirus -2 (SARS-CoV-2) to humans, which imposed immediate lockdown of effected territories in the prevailing provinces. However, few provinces were able to control infection severity with lower death rates. Interestingly three types of genomic features were noticed through comparative genomics in the available genome sequences SARS-CoV-2, due to the insertion/deletions of orf3a, orf6, orf7a and orf7b. Whole genome phylogeny (n=75 genomes) revealed a large diversity within the SARS-CoV-2, and distributed in 6 clusters namely China, Diamond princess, Asian, European, USA and Beijing. This study asserts diversity in the genome with high mutation rate and migration of carriers over the world. Here, we describe the polymorphic loci of Spike glycoprotein and its putative mechanism for pathogenicity, which unveiled the presence of GPI anchor amidation, PPI hotspot, O-linked glycosylation, catalytic site, Iron binding site, signal cleavage, disulphide linkage, sulfation, transmembrane region, and C-terminal signal sites. Mutational changes at spike glycoprotein of South Korea, India, Greece, Spain, Australia, Sweden and Yunnan samples possibly suggest the prevalence of mutated strains with either low or high virulence. The regions at the spike glycoprotein also have high binding capacity to angiotensin converting enzyme 2 (ACE2) suggesting a key link for explaining damage to multiple organs including lungs, kidney and heart. Factors influencing the mutations at the spike glycoprotein region will need to be investigated to understand and neutralize the upsurge of the alarming Pandemic and to control the global spread of the disease.

Introduction

In the history of global infections, COVID–19 (Corona Virus disease) has left its dangerous, uncontrollable outbreak footprint. Towards the end of the year 2019, people of Wuhan, the capital city of Hubei province in China, developed a strange pneumonia-like infection due to an unknown aetiology. It was later recognized to be a part of coronavirus family[1]. This pandemic was spread over 210 countries with over 3,248,685 confirmed cases and 229,399 deaths world-wide till date (01 May, 2020) (https://www.worldmeters.info/coronavirus). As of the date of manuscript preparation, the US now has almost 1/3 of all COVID–19 cases worldwide. With a high mortality rate of about 3–6% across the world, the havoc created by COVID–19 has been massive. The transmission was vast and understanding the genome variation has been in priority. Till date, no specific drugs or vaccines are available to control the infection, and symptomatic treatments to block the viral replication is in early trials. Considering COVID–19 as a major public health emergency, globally several countries have suspended their trade and called off social events to prevent community transmission. Furthermore, to battle the virus, countries world-wide have resorted to self-quarantine and social-isolation as containment strategy for the benefit of the mankind. Medical supplies, protective agents and hand hygiene are the only resort to prevent the transmission dynamics of this deadly disease. Ian M. Jones had suggested that the SARS-CoV–2 mutates rapidly in the respiratory tract[2]. The data sharing among collaborators or investigators had made the analysis more accurate and easy. Moreover, certain strains prevailing at few provinces had shown low

mortality rates. Whereas, countries like Italy, Spain, USA and a few more had high mortality rates indicating the presence of evolved virulent strain when compared to the original strain from Wuhan. Understanding the genetically distinct variants in a phenotype is very important in analysing the pathogenic mechanism of infected hosts.

As RNA viruses tend to evolve rapidly among large populations with short generation times [3], monitoring evolutionary patterns in "real time" is important. This leads to the emergence of many new novel strains of COVID–19, it has become an important aspect to differentiate between the virulent strains of SARS-CoVs in the current scenario. The virus was found more related to the betacoronavirus of bats (RaTG13) and pangolins (*Manis javanica)* with 96.2% and 91.02% homology respectively. The spike gene of SARS-CoV–2 has shown slight variation with polybasic cleavage site (PCS)[4,5]. The PCS of Spike protein gets cleaved by furin leaving its infection to different organs of the host [4–6]. The whole-genome sequence (WGS) data would probably show its evolution and reasons for its mutation rates[1,7]. Laboratories from many countries have deposited over 2400 genome sequences of SARS-CoV–2 at the NCBI, GISAID and Nextstrain databases (https://nextstrain.org/ncov), which allowed us to analyze this novel virus. Acknowledging the importance of the spread and the evolution of the virulent pathogens, the NEXTstrain database provided the necessary information related to the phylodynamics, genomes and the surveillance data [8].

It has been hypothesized that the pathogenesis of disease is possibly due to alveolar damage followed by spleen atrophy, enlarged liver, injury to kidney and neuronal dysfunction in patients[9,10]. The ability of SARS-CoV–2 to interact with the kidneys of host and shed of viral particles through faecal and urine of patients suggests the multiple organ damage with increased severity[11]. Though, the target organ reported was lungs due to the specific binding of SARS-CoV–2 to ACE2 receptors, the presence of other sites responsible for effective binding of the spike protein in other organs remains an enigma. Hence, the present study focused on the variations in the genomes of COVID–19, which were distributed worldwide. Further investigation on the mutated strains pathogenesis in biopsy samples of different human organs will be investigated.

Results And Discussion

Genome diversity and comparative genomics among SARS-CoV–2

In the present investigation, SARS-COV–2 genome sequences were retrieved from NCBI and GISAID database (till April 25th 2020). Among the 520 complete genome data of SARS-CoV–2, the genomes which showed variations in their size and their geographic region were targeted. A total of 75 complete sequences of SARS-CoV–2 that were prevalent in countries like China, USA, France, Australia, Spain, Italy, India, Nepal, Taiwan, South Korea, South Africa, Greece, Sweden, Pakistan, Peru, Brazil, Iraq, Turkey and Israel were collected (Table S1). The knowledge of patient's ethnicity and racial background were not readily available for all the samples. The genome analysis revealed that the SARS-CoV–2 is a 30 Kbp genome with over 10 to 12 genes (Table 1). The largest genome size was noticed in a Shanghai patient

(SH01) of China (Accession number MT121215) with 29945 bp (reported on 2nd Feb 2020). While the smaller genome size of 29852 bp was detected in a USA patient (CA6) (Accession number MT044258) (isolated on 27th Jan 2020). These two samples were compared to Wuhan-Hu–1 (MN908947 or NC045512), which is of 29,903 bp genome size and serves as a reference sample (Fig. 1a). Comparative genomics of these three isolates revealed that the SH01 sample had a deletion of ORF3a, ORF6, ORF7a, and 7b, while CA6 isolate had a deletion of ORF7b. However, the genome size of SH01 was noticed to be larger when compared to the other two samples. Though the function of these genes was not known, their absence had revealed diversity at strain level (Table 1, Fig. 1b).

Most of the coronaviruses (CoVs) of the Coronaviridae family possess two overlapping ORF1a and ORF1b polypeptides and other structural proteins like Spike (S), Envelope (E), membrane (M) and nucleocapsid (N)[12]. Among the samples that were analysed, ORF7b was present only in 8 samples of ncov-FIN, Yunnan–01, WH09/CHN, WIV02, WIV04, WIV05, WIV06 and WIV07. Subsequently, ORF3a, ORF7a, ORF7b and ORF8 were found to be deleted in HU/DP/Kng/19–20, SH01/CHN, WHU01 and WHU02 samples. However, the severity of the infection in these variants associated with the corresponding patient is not yet known, as the case history details are not available. These mutations could make an impact on the immunogenic changes that would either suppress or become more virulent than the wild type strain. The prevalence of more virulent strains may increase the severity of outbreak. However, extensive research has to be conducted to correlate the nature of mutations with the outbreak severity.

Table 1: General features of genes involved in SARS-CoV-2 of three isolates, i.e., Wuhan-Hu-1, CA6 and SH01 (in this study Wuhan-Hu-1 served as a reference sample)

S.NO	Gene name	Description	Gene length (bp)		
			Wuhan-Hu-1	CA6	SH01
1	orf1ab	Polyprotein	21291	21267	21271
2	orf S	Surface glycoprotein	3822	3822	3822
3	orf 3a	Hypothetical protein	828	828	-
4	orf E	Envelope protein	228	228	228
5	orf M	Membrane protein	669	669	669
6	orf 6	Hypothetical protein	186	186	-
7	orf 7a	Hypothetical protein	366	366	-
8	orf 7b	Hypothetical protein	132	-	-
9	orf 8	Hypothetical protein	366	366	-
10	orf N	Nucleocapsid phosphoprotein	1260	1260	1260
11	orf 10	Hypothetical protein	117	117	117

1. absence of genes

A recent work published by Tang and his co-workers[2] suggested the prevalence of two types of COVID–19, named as L and S type, based on their SNPs at ORF1ab and ORF8. L type was prevalent and accounted for about 70% of infected China population during Jan-Feb 2020. In the same study, they have noticed many nonsynonymous mutations in the 103 samples analyzed. However, the factors behind the emergence of L and S type are still ambiguous. In the current study, around 75 samples from different

parts of the world were considered to study their evolutionary patterns. The genome size is diverse and shows many deletions and insertions. In any case, the genetic information indicates that SARS-CoV–2 is not derived from available virus data in a laboratory, as it shares homology towards SARS, betacoronavirus of bats, and pangolins[14].

Phylogenetic evolution among COVID–19 positive samples

Due to the variations in the genome sequences of SARS-CoV–2, a genome phylogeny was constructed to understand the evolution and transmission pattern. The dendrogram suggested six groups (Fig. 2), of which group 1 had, isolates from Wuhan (IPBCAMS-WH–1/2019, WH–2/2019, and WH–3/2019), Shanghai (SH01/CHN), USA (USA-CA2), Australia (AUS/VIC01), South Korea (SNU01) and Sweden (Human/2020/SWE). These eight samples had a close resemblance with Wuhan-Hu–1 (reference isolate) and hence this group is described as Wuhan group. Group 2 is almost a clone where one international conveyance, i.e., Diamond Princess Cruise from Japan, and had over 700 coronavirus cases, with patients from the different parts of the world such as USA, Hongkong, Japan and China, the spread of the infection was vast in other countries. This group 2 is named as Diamond Princess cruise group, and the patients were quarantined in the ship for two weeks. The third European group had patients from Italy, Finland, Brazil, and a few cases from China, Japan, Taiwan and USA were grouped together, suggesting transmission from the Wuhan epicentre. Next is the Asian group (group IV), isolated in subjects from China, Japan, Taiwan, Nepal, India, and Hongkong. Group V is the USA group with the samples clubbed together with the patients from California (CA), Texas (TX), Washington (WA), and Illinois (IL). However, WA had a close cluster indicating the community transfer at USA. Also, a few patients of the USA had a travel history to China and other COVID–19 affected areas. Group VI is the Beijing group of China, which emerged from a patient from Yunnan. Phylogenomics suggests the high rate of either co-infection or recombination between different strains of SARS-CoV–2, showing its diversity. However, multiple samples have to be studied to ascertain sustainable facts, given their uncertain nature of genome variations. Tang et al.,[2] could differentiate the SARS-CoV–2 by phylogenetics into L and S types based on their aggressiveness. It was perceived that L type was more virulent than S type and mainly possessed isolates from Wuhan, France, Australia, Singapore, USA, Hongkong, Taiwan, Japan and other countries. It will be always interesting to study the transmission of these mutations and its pathogenesis in any prevalent area.

Notable changes at Spike glycoprotein

Regardless of critical advances in cutting edge sequencing innovations, which have encouraged the disclosure of thousands of novel zoonotic viruses, methods for downstream evaluation of these novel sequences are deficient. Hence, an approach to determine the functional viromics in a more applicable way to understand the host-protein interactions is obligatory. The Spike (S) protein plays a role in the entry of virus into host cells, by binding to angiotensin converting enzyme 2 (ACE–2). The motif finder programme of S protein in Wuhan-Hu–1 showed 9 Pfam motifs (Fig. S1a) (https://www.genome.jp/tools/motif/). The S protein of CoVs isolated from bats and infected humans

had >98% homology with few mutations. Basically, the S protein had an identical ribosome binding domain (RBD) and an O-linked glycan residue domain with polybasic cleavage site (RRARS) which was analysed through multiple alignment by Geneious Prime programme[15]. In the current investigation, we find that the RBD of all 75 samples is highly conserved with 9 amino acid variations when compared to Bat-RaTG13 (Fig. S1b). Similarly, the O-linked glycan residue domain had an insertion of four nucleotides *PRRA* (Fig. S1b). Later on, the samples which showed mutations at Spike glycoprotein were retrieved from Nextstrain database. Among the 358 samples analysed (data not shown), over 33 samples showed variations and suggested strain variation (Table 2). In samples from Peru (1), Israel (1), Greece (3), Spain (2), France (1), India (10) showed a common mutational site at D614G. However, these samples had variations in other ORFs of their genome, suggesting strain diversity. It was found that most of the strains possessed a unique pattern showing its strain-specificity. However, the immunological aspects of various strains and analysis is still lacking and need to be investigated. In the entire study, the structural genes of SARS-CoV–2 were mutated more rapidly than the non-structural genes.

It has been reported that Human angiotensin converting enzyme II (ACE2) receptor is the binding site for most SARS-CoV[11]. This was supported by another study which asserted that the novel SARS-CoV–2 utilizes the ACE2 to bind and find its entry in to the host cell[4]. ACE2 expression in organs like kidney and heart has been reported, providing a mechanism for the multi-organ dysfunction that can be seen with SARS-CoV–2 infection[16,17]. Interspecies diversity within different bat species harbouring the coronavirus was found[18]. In the same study, a surveillance of bat-CoV's revealed the presence of SARS-like coronavirus, unclassified betacoronavirus and new betacoronavirus species. The co-infection of these CoVs in mineshaft bat species showed potential infection in the host. Further, the RBD of pangolin CoVs are indistinguishable from that of SARS-CoV–2 at 6 of 6 key amino acids examined previously[18,19]. This observation shows that entry of CoV in a host with human-like ACE2 could choose for a RBD with high-affinity[15]. Whether the ACE2 expression in these organs affects the SARS-CoV–2 infectivity remains unclear. Majority of the scientific reports state that acute kidney injury (AKI), abdominal discomfort and cardiac damage are the most commonly reported symptoms of COVID–19[20,21] suggesting that SARS-CoV–2 may have a tropism for these organs. Such recombination's and transmission could likewise choose for the insertion of the polybasic cleavage site (PCS), which is absent in pangolin and bats coronaviruses[12]. These PCS are highly conserved in a particular strain and shows their high pathogenicity, leading to a possible pandemic outbreak with high mortality or morbidity rate[22], as observed in H5N1 virus. A putative recognition motif i.e., *PRRARSV* is present in all the sample analysed and is the active site for furin recognition[14,23]. The natural selection of virus with cleavage site would probably have taken when such virus similar to the existing SARS-CoV had undergone several passages under in vitro cell-line models. It is improbable that the O-linked glycan site would have triggered without immune pressure, which was not present in the cell-lines. The insertion of PRRA amino acids, make the SARS-CoV–2 novel and more pathogenic than SARS and MERS.

Table 2: Mutational changes observed different ORF's of SARS-CoV-2 (mutations at spike glycoprotein (S) was represented in separate column)

Accession Number	Sample ID	Mutational changes detected in different regions of SARS-CoV-2															Muations at S protein
		orf1ab	nsp1	nsp2	nsp3	nsp4	nsp6	nsp10	nsp11	nsp12	nsp15	OrfS	Orf3a	OrfE	OrfM	Orf8	OrfN	
MT350282	BRA/SP02cc/2020																	N74K
MT263074	PER/Peru-10/2020																	D614G
MT233521	Valencia6/2020																	K528del
MT292569	Valencia13/2020																	D614G
MT292575	Valencia16/2020																	D614G
MT276598	ISR/ISR-IT0320																	D614G
MT328032	GRC/10/2020																	D614G
MT328035	GRC/13/2020																	D614G
MT328034	GRC/16/2020																	I197Y
MT328033	GRC/12/2020																	D614G
MT093571	210/human/2020/SWE																	F797C
MT049951	Yunnan-01/human/2020																	Y28N
MT039890	SNU01/South Korea																	S221W
MT007544	Australia/VIC01/2020																	S247R
MT327745	TUR/ERAGEM-001/2020																	V772I
MT320538	FRA/KRA-ROB/2020																	G107del, D614G
MT300186	USA-CA																	D614G
MT324062	ZAF/R03006/2020																	D614G
MT012098	29/human/2020/IND																	Y144del, R408I
MT050493	166/human/2020/IND																	A930V
EPI_ISL_426414	India/GBRC1/2020																	Q271R, D614G
EPI_ISL_426415	India/GBRC1s/2020																	Q271R, D614G
EPI_ISL_430468	India/S2/2020/WestBengal																	D614G, G1124V
EPI_ISL_430464	India/S3/2020/WestBengal																	D614G, G1124V
EPI_ISL_424365	India/3239/2020																	D614G
EPI_ISL_428482	India/nimh-0182/2020																	D614G, C1250F
EPI_ISL_426424	USA/IN_92003/2020																	D614G, L1203F
EPI_ISL_426423	USA/IN_82003/2020																	D614G, L1203F
EPI_ISL_426422	USA/IN_72003/2020																	D614G, L1203F
EPI_ISL_420551	India/777/2020																	D614G
EPI_ISL_429691	BRA/CV35/2020																	Y695S, G832D, H1088N
EPI_ISL_429677	BRA/CV17/2020																	K776T
EPI_ISL_426882	Australia/VIC913/2020																	G446V, G1124V

Prevalence of mutant strains in certain provinces as biological markers

The samples from Brazil (3), France (1), Greece (4), Spain (3), Turkey (1), Peru (1), Israel (1), Sweden (7), India (12), China (3), USA (1), South Korea (1), South Africa (1) and Australia (2) possessed mutational changes at the Spike protein of SARS-CoV–2 and followed by their countries mortality rate was assessed (Table 3). Deletion of an amino acid tyrosine (Y), lysine (K) and Guanine (G) at 144, 528 and 107 positions was noticed in subjects of Indian (MT012098), Spain (MT233521) and France (MT320538), respectively, who had a travel history from Wuhan, China (https://www.covid19india.org/, www.nextstrain.org). Though the spike protein had no variation at the ribosome binding site, the mutations noticed in these 42 samples would either increase or decrease the severity of the outbreak. However, further analysis is required to prove the severity of these samples. The prevalence of these strains in different geographical regions has to be assessed, as these might serve as biomarkers in understanding the antigenic and immunogenic changes.

Table 3: Mutational changes in the spike protein of SARS-CoV-2 with their countries mortality rates

Accession Number	Sample ID	Country	Mutations at S protein	No. of confirmed cases	No. of Deaths	Mortality rate (%)
MT350282	BRA/SP02cc/2020	Brazil	N74K	79695	5513	6.91
MT263074	PER/Peru-10/2020	Peru	D614G	33931	943	2.77
MT233521	Valencia6/2020	Spain	K528del	239639	24543	10.24
MT292569	Valencia13/2020	Spain	D614G	239639	24543	10.24
MT292575	Valencia16/2020	Spain	D614G	239639	24543	10.24
MT276598	ISR/ISR-IT0320	Israel	D614G	15870	219	1.37
MT328032	GRC/10/2020	Greece	D614G	2576	136	5.39
MT328035	GRC/13/2020	Greece	D614G	2576	136	5.39
MT328034	GRC/16/2020	Greece	I197Y	2576	136	5.39
MT328033	GRC/12/2020	Greece	D614G	2576	136	5.39
MT093571	210/human/2020/SWE	Sweden	F797C	21092	2586	12.26
MT049951	Yunnan-01/human/2020	China	Y28N	82862	4633	5.59
MT039890	SNU01/South Korea	South Korea	S221W	10765	247	2.29
MT007544	Australia/VIC01/2020	Australia	S247R	6753	91	1.34
MT327745	TUR/ERAGEM-001/2020	Turkey	V772I	117589	3081	2.62
MT320538	FRA/KRA-ROB/2020	France	G107del, D614G	166420	24087	14.47
MT300186	USA-CA	United States of America	D614G	1067382	61849	5.79
MT324062	ZAF/R03006/2020	South Africa	D614G	5350	103	2.42
MT012098	29/human/2020/IND	India	Y144del, R408I	33610	1079	3.21
MT050493	166/human/2020/IND	India	A930V	33610	1079	3.21
EPI_ISL_426414	India/GBRC1/2020	India	Q271R, D614G	33610	1079	3.21
EPI_ISL_426415	India/GBRC1s/2020	India	Q271R, D614G	33610	1079	3.21
EPI_ISL_430468	India/S2/2020/West Bengal	India	D614G, G1124V	33610	1079	3.21
EPI_ISL_430464	India/S3/2020/West Bengal	India	D614G, G1124V	33610	1079	3.21
EPI_ISL_424365	India/3239/2020	India	D614G	33610	1079	3.21
EPI_ISL_428482	India/nimh-0182/2020	India	D614G, C1250F	33610	1079	3.21

EPI_ISL_426424	USA/IN_92003/2020	India	D614G, L1203F	33610	1079	3.21
EPI_ISL_426423	USA/IN_82003/2020	India	D614G, L1203F	33610	1079	3.21
EPI_ISL_426422	USA/IN_72003/2020	India	D614G, L1203F	33610	1079	3.21
EPI_ISL_420551	India/777/2020	India	D614G	33610	1079	3.21
EPI_ISL_429691	BRA/CV35/2020	Brazil	Y695S, G832D, H1088N	79695	5513	6.91
EPI_ISL_429677	BRA/CV17/2020	Brazil	K776T	79695	5513	6.91
EPI_ISL_426882	Australia/VIC913/2020	Australia	G446V, G1124V	6753	91	1.34
EPI_ISL_429129	Sweden/20-08681	Sweden	D80Y	21092	2586	12.26
EPI_ISL_430859	Sweden/20-08717	Sweden	K1073N	21092	2586	12.26
EPI_ISL_429152	Sweden/20-50179	Sweden	V62F	21092	2586	12.26
EPI_ISL_429157	Sweden/20-50234	Sweden	M1237I	21092	2586	12.26
EPI_ISL_429116	Sweden/20-08143	Sweden	Y917H	21092	2586	12.26
EPI_ISL_411951	Sweden/01/2020	Sweden	F797C	21092	2586	12.26
EPI_ISL_415709	Hanghou/ZJU-01	China	R682Q	82862	4633	5.56
EPI_ISL_421259	Pingxiang/JX151	China	S254F	82862	4633	5.56

While analysing the cases in these provinces, it was noticed that the death rate was low in South Korea, Greece, Brazil, Israel, Peru, Turkey, South Africa and Australia, thus COVID–19 cases curve has declined. However, these states also followed many measures in controlling the outbreak such as early lockdowns, self-isolation, social distancing, hygienic practicesas instructed by their governments. However, in Sweden, and India the COVID–19 cases are being analysed and the graph is up surging due to a hike in the confirmed cases (https://www.worldometers.info/coronavirus/). It can be seen that the death toll is comparatively low in these areas when compared to the other areas such as Wuhan, Italy, Spain, France, United States of America and Germany. This might indicate that the prevalence of mutated strains which might have emerged during coinfection within the provinces, would have either reduced or increased its severity. Furthermore, the pathogenicity probabilistically was assessed in the putative neutral variants. The MutPred Indel software could assess the sites responsible for its virulence (Table 4). The sample Human/2020/SWE from Sweden had not shown any pathogenic sites when compared to Wuhan-Hu–1 (reference strain). However, most of the subjects analysed had putative variants of S protein showing several post-translational mechanisms such as, catalytic site, proteolytic cleavage, Iron binding site, glycosyl-phosphatidylinositol (GPI) anchor amidation, PPI hotspot, sulfation, transmembrane region, copper binding, signal cleavage, cytoplasmic loop, C-terminal signal, and O-linked glycosylation sites,

suggesting probability of more virulence in these samples (P>0.5). Most of the isolates had catalytic site, PPI hotspot and Iron binding as their common mechanism for pathogenesis. However, including these mechanisms many other isolates possessed extra mechanisms for their mode of action. For example, the mechanism of palmitoylation was noted only in a sample from an Indian subject (166/Human/2020/IND). The subjects from Turkey and Brazil had disulfite linkage and sulfation as their mechanism. Considering the prevailing situation in India, the presence of pathogenic variant of spike protein it can be postulated that the rate of COVID–19 cases would increase eventually during the next few days. Hence, every citizen has to be abide to the preventive measures.

Table 4: Pathogenicity prediction with MutPred-Indel model in the putative Spike protein variants of SARS-CoV-2 samples

Sample ID	Country	Site	P-score	Mechanism for pathogenicity
Wuhan-Hu-1 (MN908947) (reference sample)	China	S221	0.515	Catalytic site, Iron binding
Human/2020/SWE	Sweden	-	-	-
29/Human/2020/IND	India	H145	0.385	PPI hotspot, Catalytic site, Iron binding, O-linked glycosylation, C-terminal signal
166/Human/2020/IND	India	S247	0.36415	GPI anchor amidation, PPI hotspot,, Catalytic site, signal cleavage, Iron binding, palmitoylation
Yunnan-01	Yunnan	N28	0.449	Iron binding
SNU-01	South Korea	S221	0.354	Catalytic site, Iron binding
Aus/VIC01	Australia	R247	0.4547	PPI hotspot, Proteolytic cleavage, Copper binding, Catalytic site, Iron binding
TUR/ERAGEM-001/2020	Turkey	V3F	0.35058	Catalytic site, PPI hotspot, Iron binding, Disulfide linkage, Sulfation
BRA/SP02cc/2020	Brazil	K83	0.346	Catalytic site, PPI hotspot, Iron binding, Disulfide linkage, Sulfation
ZAF/R03006/2020	South Africa	G623	0.3458	Catalytic site, Iron binding, PPI hotspot, Disulfide linkage, GPI anchor amidation
GRC/13/2020	Greece	I206	0.35209	PPI hotspot, Catalytic site, Iron binding, Disulfide linkage, GPI anchor amidation
India/GBRC1s/2020	India	R271	0.36268	PPI hotspot, Catalytic site, Iron binding, Disulfide linkage, GPI anchor amidation
FRA/KRA-ROB/2020	France	L1203	0.541	PPI hotshot, Catalytic site, Iron binding
USA/IN_82003/2020	USA	F1203	0.453	Signal cleavage, Iron binding, Transmembrane region, signal helix, PPI hotspot
Valencia6_ESP	Spain	S529	0.370	Iron binding, GPI anchor amidation, Catalytic site, Signal cleavage, C-terminal signal
SWE/20-08681	Sweden	Y80	0.589	PPI hotspot, Iron binding
SWE/20-08717	Sweden	Y80	0.389	Cytoplasmic loop, PPI hotspot, O-linked glycosylation, catalytic site, signal helix
SWE/20-50179	Sweden	V62	0.517	Iron binding, PPI hotspot, catalytic site, cytoplasmic loop, O-linked glycosylation
SWE/20-50234	Sweden	Y80	0.389	PPI hotspot, cytoplasmic loop, C-terminal signal, Signal cleavage, Iron binding
SWE/20-08143	Sweden	Y80	0.390	PPI hotspot, O-linked glycosylation, catalytic site, Iron binding, C-terminal signal
SWE/20-01/2020	Sweden	Y80	0.391	PPI hotspot, Catalytic site, Iron binding, C-terminal signal, signal helix

- no mechanism of pathogenicity was detected

Petit et al. [24]suggested that palmitoylation aids in providing anchoring ability during cell fusion and receptor binding in SARS-CoV, this mechanism noted in COVID−19 sample suggest conformational changes during palmitoylation process leading to signal transduction mechanism at both intra- and-extracellular domains. Sulfation is a process for protein-protein interaction and found to play a role in extracellular extension for high affinity towards binding, leading to the activation of receptors and

stability of proteins by correct protein folding mechanism [25]. Hence, mutational changes in the spike glycoprotein may instigate its conformational changes, which is most likely to prompt the evolving antigenicity[26]. Studies pertaining to the localization of amino acids associated with this protein among different variants of SARS-CoV–2, are readily not available. A recent study on protein-protein interactions (PPI) by Gordon et al.[27] had suggested that the spike protein has the ability to interact with GOLGA7-ZDHHC5 acyl transferase complex and can be a therapeutic target. GPI anchor sites are also found to target host innate defense system, which allows functions in trafficking, cell adhesion and metabolism. It was reported that, Bone marrow stroma antigen 2 (BST2), also called as CD317 or tetherin has a capacity to inhibit the enveloped virus release into the host, hence such sites can be targeted for therapeutics[28]. It will be important to explore these mutational changes. Along these lines, reinforcing SARS-CoV–2 surveillance among different geographical regions can provide scientific evidence for its more pathogenicity and allows in taking preventive and controlling measures in the transmission of disease.

ACE2 expression in human organs targeted in kidney

SARS-CoV–2 infection starts by binding of the viral surface spike protein to the human angiotensin-converting enzyme 2 (ACE2) receptor following modification of the spike protein by transmembrane protease serine 2 (TMPRSS2)[29]. Initially, ACE2 is expressed in the lung (principally Type II alveolar cells7) and seems to be the predominant portal of entry. Considering SARS entry into target human cells, it can be observed that the expression of ACE2 protein is significantly found in the epithelial cells of the lung alveoli and small intestine and endothelial cells of organs including spleen, kidney, liver, lymph nodes, brain[30,31]. Burgeoning data confirm association of COVID–19 infection with increased morbidity and mortality from kidney disease. It is important to investigate whether SARS-CoV–2 replication occurs in these organs contributing to the virus disseminating throughout the body.

High expression of ACE2 was noticed in proximal tubular cells and to a lesser extent in podocytes, however, kidney glomerular endothelial and mesangial cells were not affected[17]. It was perceived that only 6% of patients infected with SARS-CoV experienced Acute Kidney Injury (AKI) during SARS outbreak during 2003 [32]. Furthermore, AKI was identified as a serious complication of SARS, with mortality of 92% in patients [32]. During post-mortem from SARS patients, SARS-CoV viral particles were noticed in renal specimens, suggesting AKI was caused by active replication of SARS-CoV in tubular cells [32]. They suggested that renal impairment was likely associated with multi-organ failure as SARS-CoV was not demonstrable in any of the examined patients. Further, AKI (including cytokine release syndrome and SARS patients) might be a specific pathogenic condition, and might not be due to the active replication of virus at kidneys [32, 33]. An increased viral infection in alveolar cells leads to the production of large amount of cytokines, causing multiple-organ failure. Previously a study had reported that release of interferon-gamma-related cytokine increased the severity of organ damage in SARS patients[34]. Recently, a study described that the human kidney is a specific target for SARS-CoV–2 infection[35]. The difference between the higher renal tropism of SARS-CoV–2 versus SARS-CoV can be assessed by the increased affinity of

SARSCoV–2 for ACE2, contributing towards pronounced infection of the kidney, leading to viral reservoir [36].

In addition, a small survey on COVID–19 patients has revealed that, proteinuria and haematuria are common features that were noticed in 40% of patients post hospitalization [37]. A reduced density of inflammation and edema was observed in CT scan reports of kidneys samples infected with SARS-CoV–2[38]. Furthermore, SARS-CoV–2 seems to be affected more by AKI frequently than subjects infected with SARS-CoV[37]. A very recent study by Yao et al.[39] confirms that SARS-CoV–2 infection damages vessels, kidney and other organs, in addition to the lungs,. Hyaline thrombi are found in small vessels in different organs. It is of utmost importance to investigate pathological changes in autopsy material. Before organ donation is considered in future, it will be important to investigate whether the SARS-CoV–2 has infected the kidneys; the risk of such organ grafts has not been reported as yet. In any case, it has been indicated that SARSCoV–2 has a high tropism for the kidney, where it has been shown to reproduce in practically 30% of COVID–19 patients[40]. Consequently, screening for COVID–19 in kidney donors is probably more important during screening time and need to be quarantined for 14–28 days who possess either symptoms or had a travel history to high-risk regions [31]. A research study demonstrated that more than 66% of patients had died with COVID–19 infection who had diabetes or cardiovascular disease[41]. As a first-line treatment, angiotensin-receptor blockers (ARBs) were given to COVID–19 patients. Certain reports revealed that ARBs were found to express ACE2 by nearly 2 to 5 fold in kidney and heart samples [42–45]. Since SARSCoV–2 has a high tropism for the kidney[35,40], investigating how ARBs affect the infection rate and renal and cardiac injury in COVID–19 infection is of great importance.

Conclusion

Bat coronavirus remains a considerable worldwide risk to general wellbeing of humans. The genomic highlights depicted in the present study might clarify the transmissibility of SARS-CoV–2 in human race, yet its inception is a question. Despite the fact that the nCoV–2019 had close genetic relatedness towards RaTG13 of bat coronavirus, which was isolated in 2013 (however, the genome data of this strain was made open just after the outbreak of COVID–19). The current investigation provided three types of variants in the SARS-CoV–2 genome. The phylogeny showed six clusters which includes Wuhan, Diamond princess, European, Asian, USA and Beijing group. The polybasic cleavage site in the Spike protein of COVID–19 isolates is very conserved and different from bat and pangolins CoV's, suggesting its novel pathogenic nature. Mutations in the spike protein could either reduce or increase the severity of the outbreak. The mechanism of pathogenicity among putative variants of spike surface glycoprotein suggested more virulence in few samples of India, Australia, Greece, South Korea and Yunnan. The available clinical data have confirmed that AKI is one of the main risk factors in the prognosis of COVID–19. Patients with diabetic nephropathy, end-stage renal disease, and, renal transplantation may be at high risk of the SAR-CoV–2. However, spread of SARS-like infections from various intermediate animals may assist in explaining its emergence or outbreaks. The recognizable proof of a potential intermediate host

of SARS-CoV–2, along with their genome sequence data of the virus at early stage, would also be profoundly useful.

Materials And Methods

Information related to daily cases of COVID–19 and SARS-CoV–2 genome data

The genome data of SARS-CoV–2 was retrieved from the public repositories like NCBI data and the global information on COVID–19 cases was obtained through worldometers (https://www.worldometers.info/coronavirus/) and NEXTstrain (https://nextstrain.org/ncov) websites. Totally 75 genomes were considered based on the variation in their genome size, country and divergence. The samples used in the current study are enlisted in Table S1. However, the genome ethnicity and racial inheritance of all the samples are not available.

Comparative genome analysis

Three genomes MT121215 and MT044528 were considered in the study which possessed highest and lowest genome size, and were compared to the reference sample (MN908947). The comparative genome analysis was performed by using Geneious Prime Software Version 2019.2.1[13].

Phylogenetic evolution

To further analyze the evolution of isolates, a phylogenetic tree (n = 75) was constructed using the complete genome data of SARS-CoV–2 by using MEGA-X (Molecular Evolutionary Genetic Analysis) software[46]. The evolutionary history was deduced by using the Neighbor-Joining method with 500 bootstrap replicates[47]. Further, the evolutionary distances were computed using the Maximum Composite Likelihood method. The analysis involved 75 nucleotide sequences. There was a total of 39547 positions in the final dataset.

Bioinformatics tools used in the analysis of spike protein

Multiple sequence analysis of the spike protein among the 75 isolates was performed by using Clustal W programme[48]. The VIPR database is used to analyse the single nucleotide polymorphism (SNP) at spike glycoprotein as described by Pickett et al.[49]. Further Genome Detective Virus Tools was also used to look at the mutational analysis (*https://www.genomedetective.com/app/typingtool/virus/)*. The pfam motifs were analysed by using genome motifs database (https://www.genome.jp/tools/motif/). Further, the ribosome binding region and the polybasic cleavage site was determined as described by Andersen et al.[15].

Pathogenicity prediction in the phenotypes

To further assess the pathogenicity of the variants (putatively neutral) a machine learning-based method software package was employed for the spike protein phenotypes. The MutPred-Indel software assess the probabilistically the pathogenicity of the neutral variants and suggests the features affecting the phenotypes[50].

Declarations

Authors Contribution

SMD performed the bioinformatics analysis on the collected data, planned the work and wrote the manuscript, AP had collected the data and worked on kidney biopsy samples of COVID-19, BK edited the manuscript, and KS monitored the analysis and edited the manuscript.

Acknowledgment

The authors thank all family members and friends for all the support during the crisis period of COVID-19. The work was not supported by any fund, the article was written to analyse the genome data of COVID-19 considering the present awareness.

Conflict of Interest

The author claims no conflict of interest.

References

1. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265-269 (2020).
2. Tang, X. *et al.* On the origin and continuing evolution of SARS-CoV-2. *National Science Review* (2020).
3. Duffy, S., Shackelton, L.A. & Holmes, E.C. Rates of evolutionary change in viruses: patterns and determinants. *Nature Reviews Genetics* **9**, 267-276 (2008).
4. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270-273 (2020).
5. Wong, M.C., Cregeen, S.J.J., Ajami, N.J. & Petrosino, J.F. Evidence of recombination in coronaviruses implicating pangolin origins of nCoV-2019. *Biorxiv* (2020).
6. Tsai, H.J., Chi, L.-A. & Alice, L.Y. Monoclonal antibodies targeting the synthetic peptide corresponding to the polybasic cleavage site on H5N1 influenza hemagglutinin. *Journal of biomedical science* **19**, 37 (2012).

7. Perrella, A. *et al.* Editorial–Novel Coronavirus 2019 (Sars-CoV2): a global emergency that needs new approaches. *Eur Rev Med Pharmaco* **24**, 2162-2164 (2020).
8. Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121-4123 (2018).
9. Yao, X. *et al.* A pathological report of three COVID-19 cases by minimally invasive autopsies. *Zhonghua bing li xue za zhi= Chinese journal of pathology* **49**, E009-E009 (2020).
10. Perico, L., Benigni, A. & Remuzzi, G. Should COVID-19 concern nephrologists? Why and to what extent? The emerging impasse of angiotensin blockade. *Nephron*, 1-9 (2020).
11. Li H, L.L., Zhang D, Xu J, Dai H, Tang N, Su X, Cao B. SARS-CoV-2 and viral sepsis: observations and hypotheses. *The Lancet* (2020 Apr 17).
12. Fan, Y., Zhao, K., Shi, Z. & Zhou, P. Bat Coronaviruses in China. Viruses, 11 (3), 210. (2019).
13. Kearse, M. *et al.* Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647-1649 (2012).
14. Zhang, T., Wu, Q. & Zhang, Z. Probable pangolin origin of SARS-CoV-2 associated with the COVID-19 outbreak. *Current Biology* (2020).
15. Andersen, K.G., Rambaut, A., Lipkin, W.I., Holmes, E.C. & Garry, R.F. The proximal origin of SARS-CoV-2. *Nature medicine* **26**, 450-452 (2020).
16. Zhang, H., Penninger, J.M., Li, Y., Zhong, N. & Slutsky, A.S. Angiotensin-converting enzyme 2 (ACE2) as a SARS-CoV-2 receptor: molecular mechanisms and potential therapeutic target. *Intensive Care Medicine* **46**, 586-590 (2020).
17. Ye, M. *et al.* Glomerular Localization and Expression of Angiotensin-Converting Enzyme 2 and Angiotensin-Converting Enzyme: Implications for Albuminuria in Diabetes. *Journal of the American Society of Nephrology* **17**, 3067-3075 (2006).
18. Ge, X.-Y. *et al.* Coexistence of multiple coronaviruses in several bat colonies in an abandoned mineshaft. *Virologica Sinica* **31**, 31-40 (2016).
19. Hu, B. *et al.* Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus. *PLoS pathogens* **13**(2017).
20. Gallagher, P.E., Ferrario, C.M. & Tallant, E.A. Regulation of ACE2 in cardiac myocytes and fibroblasts. *American Journal of Physiology-Heart and Circulatory Physiology* **295**, H2373-H2379 (2008).
21. G, C. Bulk and single-cell transcriptomics identify tobacco-use disparity in lung gene expression of ACE2, the receptor of 2019-nCov. *MedRxiv* (2020 Jan 1).
22. Hatta, M., Gao, P., Halfmann, P. & Kawaoka, Y. Molecular basis for high virulence of Hong Kong H5N1 influenza A viruses. *Science* **293**, 1840-1842 (2001).
23. Yamada, Y. & Liu, D.X. Proteolytic activation of the spike protein at a novel RRRR/S motif is implicated in furin-dependent entry, syncytium formation, and infectivity of coronavirus infectious bronchitis virus in cultured cells. *Journal of virology* **83**, 8744-8758 (2009).

24. Petit, C.M. *et al.* Palmitoylation of the cysteine-rich endodomain of the SARS–coronavirus spike glycoprotein is important for spike-mediated cell fusion. *Virology* **360**, 264-274 (2007).
25. Ngounou Wetie, A.G. *et al.* Investigation of stable and transient protein-protein interactions: Past, present, and future. *Proteomics* **13**, 538-557 (2013).
26. Phan, T. Genetic diversity and evolution of SARS-CoV-2. *Infection, Genetics and Evolution* **81**, 104260 (2020).
27. Gordon, D.E. *et al.* A SARS-CoV-2-human protein-protein interaction map reveals drug targets and potential drug-repurposing. *BioRxiv* (2020).
28. Wang, S.M., Huang, K.J. & Wang, C.T. Severe acute respiratory syndrome coronavirus spike protein counteracts BST2-mediated restriction of virus-like particle release. *Journal of medical virology* **91**, 1743-1750 (2019).
29. Hoffmann, M. *et al.* SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* **181**, 271-280.e8 (2020).
30. Hamming, I. *et al.* Tissue distribution of ACE2 protein, the functional receptor for SARS coronavirus. A first step in understanding SARS pathogenesis. *The Journal of Pathology* **203**, 631-637 (2004).
31. Perico, L., Benigni, A. & Remuzzi, G. Should COVID-19 Concern Nephrologists? Why and to What Extent? The Emerging Impasse of Angiotensin Blockade. *Nephron* (2020).
32. Chu, K.H. *et al.* Acute renal impairment in coronavirus-associated severe acute respiratory syndrome. *Kidney International* **67**, 698-705 (2005).
33. Tisoncik, J.R. *et al.* Into the Eye of the Cytokine Storm. *Microbiology and Molecular Biology Reviews* **76**, 16-32 (2012).
34. Huang, K.-J. *et al.* An interferon-γ-related cytokine storm in SARS patients. *Journal of Medical Virology* **75**, 185-194 (2005).
35. Diao, B. *et al.* Human Kidney is a Target for Novel Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) Infection. *medRxiv*, 2020.03.04.20031120 (2020).
36. Wan, Y., Shang, J., Graham, R., Baric, R.S. & Li, F. Receptor Recognition by the Novel Coronavirus from Wuhan: an Analysis Based on Decade-Long Structural Studies of SARS Coronavirus. *Journal of Virology* **94**, e00127-20 (2020).
37. Wang T, H.M., Chen X, Fu Y, Lei C, Dong H, Zhou Y, Jia H, Chen X, Yan J. Caution on Kidney Dysfunctions of 2019-nCoV Patients. *medRxiv.* (February 7th 2020).
38. Cheng, Y. *et al.* Kidney impairment is associated with in-hospital death of COVID-19 patients. *medRxiv*, 2020.02.18.20023242 (2020).
39. Yao, X.H. *et al.* [A pathological report of three COVID-19 cases by minimally invasive autopsies]. *Zhonghua bing li xue za zhi = Chinese journal of pathology* **49**, E009 (2020).
40. Pan, X.-W. *et al.* Identification of a potential mechanism of acute kidney injury during the COVID-19 outbreak: a study based on single-cell transcriptome analysis. *Intensive care medicine*, 1-3 (2020).
41. Remuzzi, A. & Remuzzi, G. COVID-19 and Italy: what next? *The Lancet* **395**, 1225-1228 (2020).

42. Gallagher, P.E., Ferrario, C.M. & Tallant, E.A. MAP kinase/phosphatase pathway mediates the regulation of ACE2 by angiotensin peptides. *American Journal of Physiology-Cell Physiology* **295**, C1169-C1174 (2008).
43. Ishiyama, Y. *et al.* Upregulation of Angiotensin-Converting Enzyme 2 After Myocardial Infarction by Blockade of Angiotensin II Receptors. *Hypertension* **43**, 970-976 (2004).
44. Ferrario, C.M. *et al.* Effect of Angiotensin-Converting Enzyme Inhibition and Angiotensin II Receptor Blockers on Cardiac Angiotensin-Converting Enzyme 2. *Circulation* **111**, 2605-2610 (2005).
45. Jessup, J.A. *et al.* Effect of angiotensin II blockade on a new congenic model of hypertension derived from transgenic Ren-2 rats. *American Journal of Physiology-Heart and Circulatory Physiology* **291**, H2166-H2172 (2006).
46. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Molecular biology and evolution* **35**, 1547-1549 (2018).
47. Tamura, K., Nei, M. & Kumar, S. Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proceedings of the National Academy of Sciences* **101**, 11030-11035 (2004).
48. Altschul, S.F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* **25**, 3389-3402 (1997).
49. Pickett, B.E. *et al.* ViPR: an open bioinformatics database and analysis resource for virology research. *Nucleic acids research* **40**, D593-D598 (2012).
50. Pagel, K.A. *et al.* Pathogenicity and functional impact of non-frameshifting insertion/deletion variation in the human genome. *PLoS computational biology* **15**, e1007112 (2019).

Figures

Fig. 1a: Circular maps of the SARS-CoV-2 samples analysed by using Geneious Prime Software Version 2019.2.1[13] (Yellow color indicates gene, Green color is for CDS and orange color represents nascent peptide)

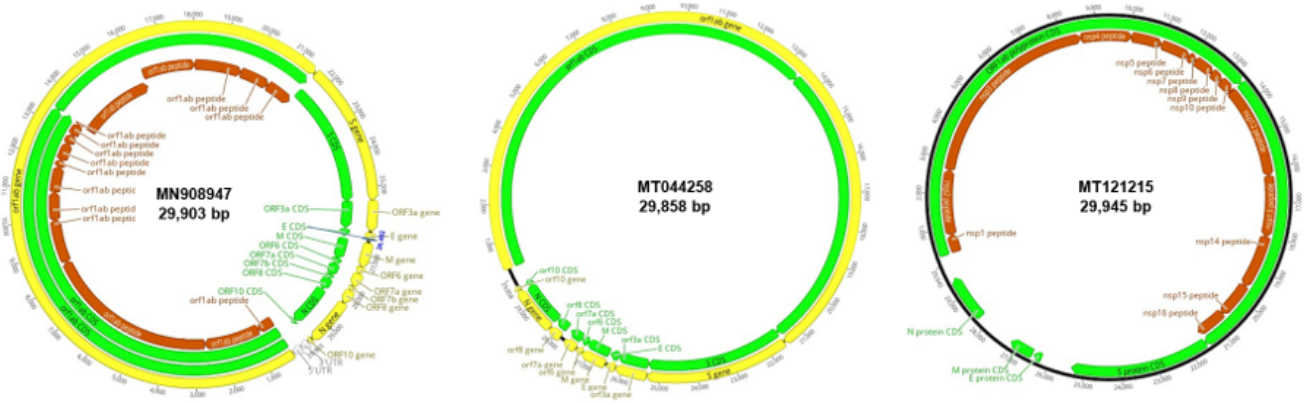

Fig 1b: Comparative genome analysis of SARS-CoV-2 isolates Wuhan-Hu-1 (NC045512), CA-6 (MT044258) and SH01 (MT121215)

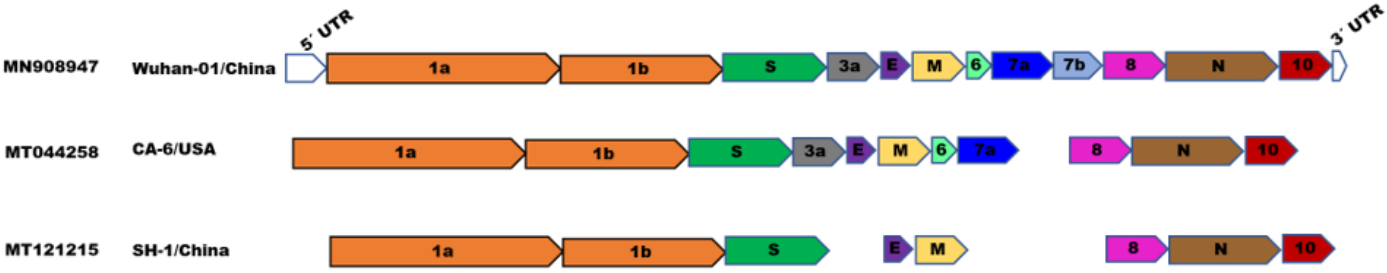

Figure 1

a: Circular maps of the SARS-CoV-2 samples analysed by using Geneious Prime Software Version 2019.2.113 (Yellow color indicates gene, Green color is for CDS and orange color represents nascent peptide). b: Comparative genome analysis of SARS-CoV-2 isolates Wuhan-Hu-1 (NC045512), CA-6 (MT044258) and SH01 (MT121215)

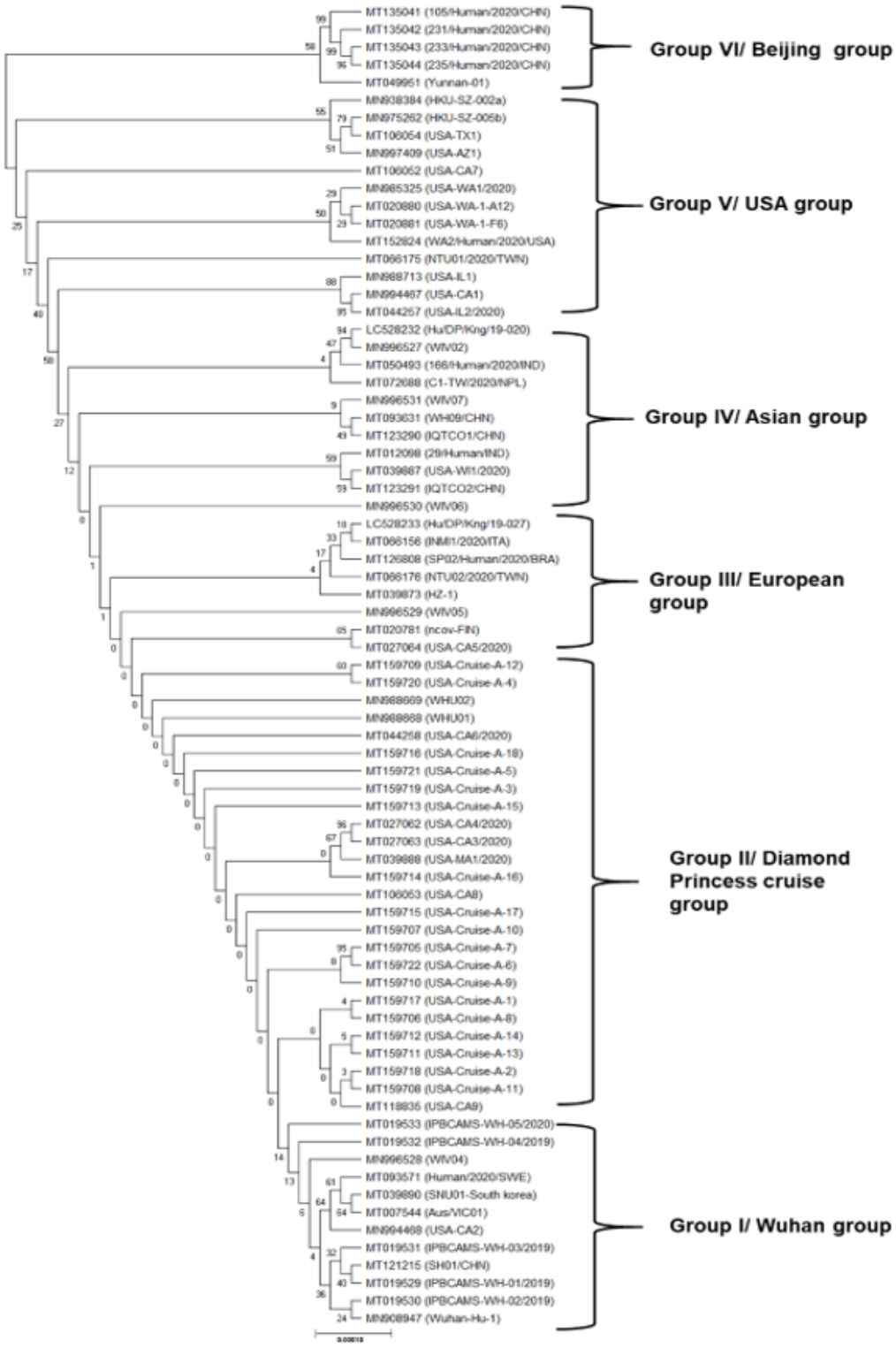

Figure 2

Phylogenetic emergence of COVID-19 among 75 prevalent samples globally In this study, each sample was given an ID, however, the ethnicity and geographical location of most of the patients details were not available.