Comparative Genomic Analysis of Rapidly Evolving SARS-CoV-2 Viruses Reveal Mosaic Pattern of Phylogeographical Distribution

Roshan Kumar[1], Helianthous Verma[2], Nirjara Singhvi[3], Utkarsh Sood[4], Vipin Gupta[5], Mona Singh[5], Rashmi Kumari[6], Princy Hira[7], Shekhar Nagar[3], Chandni Talwar[3], Namita Nayyar[8], Shailly Anand[9], Charu Dogra Rawat[2], Mansi Verma[8], Ram Krishan Negi[3], Yogendra Singh[3] and Rup Lal[4*]

Authors Affiliations

[1]P.G. Department of Zoology, Magadh University, Bodh Gaya, Bihar-824234, India

[2]Department of Zoology, Ramjas College, University of Delhi, New Delhi-110007, India

[3]Department of Zoology, University of Delhi, New Delhi-110007, India

[4]The Energy and Resources Institute, Darbari Seth Block, IHC Complex, Lodhi Road, New Delhi-110003, India

[5]PhiXGen Private Limited, Gurugram, Haryana 122001, India

[6]Department of Zoology, College of Commerce, Arts & Science, Patliputra University, Patna, Bihar-800020, India

[7]Department of Zoology, Maitreyi College, University of Delhi, New Delhi-110021, India

[8]Department of Zoology, Sri Venkateswara College, University of Delhi, New Delhi-110021, India

[9]Department of Zoology, Deen Dayal Upadhyaya College, University of Delhi, New Delhi-110078, India

***Corresponding Author**

Email: ruplal@gmail.com

Abstract

The Coronavirus Disease-2019 (COVID-19) that started in Wuhan, China in December 2019 has spread worldwide emerging as a global pandemic. The severe respiratory pneumonia caused by the novel SARS-CoV-2 has so far claimed more than 60,000 lives and has impacted human lives worldwide. However, as the novel SARS-CoV-2 displays high transmission rates, their underlying genomic severity is required to be fully understood. We studied the complete genomes of 95 SARS-CoV-2 strains from different geographical regions worldwide to uncover the pattern of the spread of the virus. We show that there is no direct transmission pattern of the virus among neighboring countries suggesting that the outbreak is a result of travel of infected humans to different countries. We revealed unique single nucleotide polymorphisms (SNPs) in nsp13-16 (ORF1b polyprotein) and S-Protein within 10 viral isolates from the USA. These viral proteins are involved in RNA replication, indicating highly evolved viral strains circulating in the population of USA than other countries. Furthermore, we found an amino acid addition in nsp16 (mRNA cap-1 methyltransferase) of the USA isolate (MT188341) leading to shift in amino acid frame from position 2540 onwards. Through the construction of SARS-CoV-2-human interactome, we further revealed that multiple host proteins (PHB, PPP1CA, TGF-β, SOCS3, STAT3, JAK1/2, SMAD3, BCL2, CAV1 & SPECC1) are manipulated by the viral proteins (nsp2, PL-PRO, N-protein, ORF7a, M-S-ORF3a complex, nsp7-nsp8-nsp9-RdRp complex) for mediating host immune evasion. Thus, the replicative machinery of SARS-CoV-2 is fast evolving to evade host challenges which need to be considered for developing effective treatment strategies.

Background

Since the current outbreak of pandemic coronavirus disease 2019 (COVID-19) caused by Severe Acute Respiratory Syndrome-related Coronavirus-2 (SARS-CoV-2), the assessment of the biogeographical pattern of SARS-CoV-2 isolates and the mutations at nucleotide and protein level is of high interest to many research groups [1, 2, 3]. Coronaviruses (CoVs), members of *Coronaviridae* family, order *Nidovirales*, have been known as human pathogens from the last six decades [4]. Their target is not just limited to humans, but also other mammals and birds [5]. Coronaviruses have been classified under alpha, beta, gamma and delta-coronavirus groups [6] in which former two are known to infect mammals while the latter two primarily infect bird species [7]. Symptoms in humans vary from common cold to respiratory and gastrointestinal distress of varying intensities. In the past, more severe forms caused major outbreaks that include Severe Acute Respiratory Syndrome (SARS-CoV) (outbreak in 2003, China) and Middle East Respiratory Syndrome (MERS-CoV) (outbreak in 2012, Middle East) [8]. Bats are known to host coronaviruses acting as their natural reservoirs which may be transmitted to humans through an intermediate host. SARS-CoV and MERS-CoV were transmitted from intermediate hosts, palm civets and camel, respectively [9, 10]. It is not, however, yet clear which animal served as the intermediate host for transmission of SARS-CoV-2 transmission from bats to humans which is most likely suggested to be a warm-blooded vertebrate [11].

The inherently high recombination frequency and mutation rates of coronavirus genomes allow for their easy transmission among different hosts. Structurally, they are positive-sense single stranded RNA (ssRNA) virions with characteristic spikes projecting from the surface of capsid coating [12, 13]. The spherical capsid and spikes give them crown-like appearance due to which they were named as ‘corona’, meaning ‘crown’ or ‘halo’ in *Latin*. Their genome is nearly 30 Kb long, largest among the RNA viruses, with 5’cap and 3’ polyA tail, for translation [14]. Coronavirus consists of four main proteins, spike (S), membrane (M), envelope (E) and nucleocapsid (N). The spike (~150 kDa) mediates its attachment to host receptor proteins [15]. Membrane protein (~25-30 kDa) attaches with nucleocapsid and maintains curvature of virus membrane [16]. E protein (8-12 kDa) is responsible for the pathogenesis of the virus as it eases assembly and release of virion particles and also has ion channel activity as integral membrane protein [17]. N-protein, the fourth protein, helps in the packaging of virus particles into capsids and promotes replicase-transcriptase complex (RTC) [18].

Recently, in December 2019, the outbreak of novel beta-coronavirus (2019-nCoV) or SARS-CoV-2 in Wuhan, China has shown devastating effects worldwide (https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200403-sitrep-74-covid-19-mp.pdf?sfvrsn=4e043d03_4)).World Health Organization (WHO) has declared COVID-19, the disease caused by the novel SARS-CoV-2 a pandemic, affecting more than 186 countries and territories where USA has most reported cases 2,13,600 and Italy has highest mortality rate 12.08% (1,15,242 infected individuals, 13,917 deaths) (WHO situation report-74). As on date (April 4, 2020), more than 1 million individuals have been infected by SARS-CoV-2 and nearly 60,000 have died worldwide. Virtually, all human lives have been impacted with no foreseeable end of the pandemic. A recent study on ten novel coronavirus strains by Lu *et al.*, suggested that SARS-CoV-2 has sufficiently diverged from SARS-CoV [19]. SARS-CoV-2 is assumed to have originated from bats, which serve as a reservoir host of the virus [19]. A recent study has shown similar mutation patterns in Bat-SARS-CoV RaTG13 and SARS CoV-2, but the dataset was limited to 21 strains including few SARS-CoV-2 strains and other neighbors [20]. Other studies have also reported the genome composition and divergence patterns of SARS-CoV-2 [3, 21]. However, no study has yet explained the biogeographical pattern of this emerging pathogen. In this study, we selected 95 strains of SARS-CoV-2, isolated and sequenced from 11 different countries to understand the transmission patterns, evolution and pathogenesis of the virus. Using core genome and Single Nucleotide Polymorphism (SNP) based phylogeny, we attempted to uncover any existence of a transmission pattern of the virus across the affected countries, which was not known earlier. We analyzed the ORFs of the isolates to reveal unique point mutations and amino-acid substitutions/additions in the isolates from the USA. In addition, we analyzed the gene/protein mutations in these novel strains and estimated the direction of selection to decipher their evolutionary divergence rate. Further, we also established the interactome of SARS-CoV-2 with the human host proteins to predict the functional implications of the viral infection host cells. The results obtained from the analyses indicate the high severity of SARS-CoV-2 isolates with the inherent capability of unique mutations and the evolving viral replication strategies to adapt to human hosts.

Materials and Methods

Selection of genomes and annotation

Sequences of different strains were downloaded from NCBI database https://www.ncbi.nlm.nih.gov/genbank/sars-cov-2-seqs/ (Table 1). A total of 97 genomes were downloaded on March 19, 2020 from NCBI database and based on quality assessment two genomes with multiple Ns were removed from the study. Further the genomes were annotated using Prokka [22]. A manually annotated reference database was generated using the Genbank file of Severe acute respiratory syndrome coronavirus 2 isolate- SARS-CoV-2/SH01/human/2020/CHN (Accession number: MT121215) and open reading frames (ORFs) were predicted against the formatted database using prokka (-gcode 1) [22]. Further the GC content information was generated using QUAST standalone tool [23].

Analysis of natural selection

To determine the evolutionary pressure on viral proteins, dN/dS values were calculated for 9 ORFs of all strains. The orthologous gene clusters were aligned using MUSCLE v3.8 [24] and further processed for removing stop codons using HyPhy v2.2.4 [25]. Single-Likelihood Ancestor Counting (SLAC) method in Datamonkey v2.0 [26] (http://www.datamonkey.org/slac) was used to calculate dN/dS value for each orthologous gene cluster. The dN/dS values were plotted in R (R Development Core Team, 2015).

Phylogenetic analysis

To infer the phylogeny, the core gene alignment was generated using MAFFT [27] present within the Roary Package [28]. Further, the phylogeny was inferred using the Maximum Likelihood method based and Tamura-Nei model [29] in MEGAX [30] and visualized in interactive Tree of Life (iTOL) [31] and GrapeTree [32].

To determine the single nucleotide polymorphism (SNP), whole-genome alignments were made using libMUSCLE aligner. For this, we used annotated genbank of SARS-CoV-2/SH01/human/2020/CHN (Accession no. MT121215) as the reference in the parsnp tool of Harvest suite [33]. As only genomes within a specified MUMI distance threshold are recruited, we used option -c to force include all the strains. For output, it produced a core-genome alignment, variant calls and a phylogeny based on Single nucleotide polymorphisms. The SNPs were further visualized in Gingr, a dynamic visual platform [33]. Further, the tree was visualized in interactive Tree of Life (iTOL) [31].

SARS-CoV-2 protein annotation and host-pathogenic interactions

SARS-CoV-2/SH01/human/2020/CHN virus genome having accession no. MT121215.1 was used for protein-protein network analysis. Since, none of the SARS-CoV-2 genomes are updated in any protein database, we first annotated the genes using BLASTp tool [34]. The similarity searches were performed against SARS-CoV isolate Tor2 having accession no. AY274119 selected from NCBI at default parameters. The annotated SARS-CoV-2 proteins were mapped against viruSITE [35] and interaction databases such as Virus.STRING v10.5 [36] and IntAct [37] for predicting their interaction against host proteins. These proteins were either the direct targets of HCoV proteins or were involved in critical pathways of HCoV infection identified by multiple experimental sources. To build a comprehensive list of human PPIs, we assembled data from a total of 18 bioinformatics and systems biology databases with five types of experimental evidence: (i) binary PPIs tested by high-throughput yeast two-hybrid (Y2H) systems; (ii) binary, physical PPIs from protein 3D structures; (iii) kinase-substrate interactions by literature-derived low-throughput or high-throughput experiments; (iv) signaling network by literature-derived low-throughput experiments; and (v) literature-curated PPIs identified by affinity purification followed by mass spectrometry (AP-MS), Y2H, or by literature-derived low [36, 38].

Filtered proteins (confidence value: 0.7) were mapped to their Entrez ID [39] based on the NCBI database used for interactome analysis. HPI were stimulated using Cytoscape v.3.7.2 [40].

Functional enrichment analysis

Next, functional studies were performed using the Kyoto Encyclopedia of Genes and Genomes (KEGG) [41, 42] and Gene Ontology (GO) enrichment analyses using UniProt database [43] to evaluate the biological relevance and functional pathways of the HCoV-associated proteins. All functional analyses were performed using STRING enrichment and STRINGify, plugin of Cytoscape v.3.7.2 [40]. Network analysis was performed by tool NetworkAnalyzer, plugin of Cytoscape with the orthogonal layout.

Results and Discussion

General genomic attributes of SARS-CoV-2

In this study, we analyzed a total of 95 SARS-CoV-2 strains (available on March 19, 2020) isolated between December 2019-March 2020 from 11 different countries namely USA (n=52), China (n=30), Japan (n=3), India (n=2), Taiwan (n=2) and one each from Australia, Brazil, Italy, Nepal, South Korea and Sweden. A total of 68 strains were isolated from either oronasopharynges or lungs, while two of them were isolated from faeces suggesting both respiratory and gastrointestinal connection of SARS-CoV-2 (Table 1). No information of the source of isolation of the remaining isolates is available. The average genome size and GC content were found to be 29879 ± 26.6 bp and 37.99 ± 0.018%, respectively. All these isolates were found to harbor 9 open reading frames coding for ORF1a (13218 bp) and ORF1b (7788 bp) polyproteins, surface glycoprotein or S-protein (3822 bp), ORF3a protein (828 bp), membrane glycoprotein or M-protein (669 bp), ORF6 protein (186 bp), ORF7a protein (366 bp), ORF8 protein (366 bp), and nucleocapsid phosphoprotein or N-protein (1260 bp) which agrees with a recently published study [44]. The ORF1a harbors 12 non-structural protein (nsp) namely nsp1, nsp2, nsp3 (papain-like protease or PLpro domain), nsp4, nsp5 (3C-like protease or 3CLpro), nsp6, nsp7, nsp8, nsp9, nsp10, nsp11and nsp12 (RNA-dependent RNA polymerase or RdRp) [44]. Similarly, ORF1b contains four putative nsp's namely nsp13 (helicase or Hel), nsp14 (3′-to-5′ exoribonuclease or ExoN), nsp15 and nsp16 (mRNA cap-1 methyltransferase).

Phylogenomic analysis: defining evolutionary relatedness

Our analysis revealed that strains of human infecting SARS-CoV-2 are novel and highly identical (>99.9%). A recent study established the closest neighbor of SARS-CoV-2 as SARSr-CoV-RaTG13, a bat coronavirus [45]. As COVID19 transits from epidemic to pandemic due to extremely contagious nature of the SARS-CoV-2, it was interesting to draw the relation between strains and their geographical locations. In this study, we employed two methods to delineate phylogenomic relatedness of the isolates: core genome (Figure 1A & C) and single nucleotide polymorphisms (SNPs) (Figure 1B). Phylogenies obtained were annotated with country of isolation of each strain (Figure 1A & B). The phylogenetic clustering was found majorly concordant by both core-genome (Figure 1A) and SNP based methods (Figure 1B). The strains formed a monophyletic clade, in which MT093571.1 (South Korea) and MT039890.1 (Sweden) were most diverged. Focusing on the edge-connection between the neighboring countries from where the transmission is more likely to occur, we noted a strain from Taiwan (MT066176) closely clustered with another from China (MT121215.1). With the exception of these two strains, we did not find any connection between strains of neighboring countries. Thus, most strains belonging to the same country clustered distantly from each other and showed relatedness with strains isolated from distant geographical locations (Figure 1A & B). For instance, a SARS-CoV-2 strain isolated from Nepal (MT072688) clustered with a strain from USA (MT039888). Also, strains from Wuhan (LR757998 and LR757995), where the virus was originated, showed highest identity with USA as well as China strains; strains from India, MT012098 and MT050493 clustered closely with China and USA strains, respectively (Figure 1A & B). Similarly, Australian strain (MT007544) showed close clustering with USA strain (Figure 1A & B) and one strain from Taiwan (MT066175) clustered nearly with Chinese isolates (Figure 1B). Isolates from Italy (MT012098) and Brazil (MT126808) clustered with different USA strains (Figure 1A & B). Notably, isolates from same country or geographical location formed a mosaic pattern of phylogenetic placements of countries' isolates. For viral transmission, contact between the individuals is also an important factor, supposedly due to which the spread of identical strains across the border of neighboring countries is more likely. But we obtained a pattern where Indian strains showed highest similarity with USA and China strains, Australian strains with USA strains, Italy and Brazilian strains with strains isolated from USA among others. This depicts the viral spread across different communities. However, as genomes of SARS-CoV-2 were available mostly from USA and China, sampling biases is evident in analyzed dataset as available on NCBI. Thus, it is plausible for strains from other countries to show most similarity with strains from these two countries. In the near future as more and more genome sequences will become available from different geographical locations; more accurate patterns of their relatedness across the globe will become available

SNPs in the SARS-CoV-2 genomes

SNPs in all predicted ORFs in each genome were analyzed using SARS-CoV-2/SH01/human/2020/CHN as a reference. SNPs were determined using maximum unique matches between the genomes of coronavirus, we observed that the strains isolated from USA (MT188341; MN985325; MT020881; MT020880; MT163719; MT163718; MT163717; MT152824; MT163720; MT188339) are the most evolved and they carry set of unique point mutations (Table2) in nsp13, nsp14, nsp15, nsp16 (present in orf1b polyprotein region) and S-Protein. All the mutated proteins are non-structural proteins (NSP) functionally involved in forming viral replication-transcription complexes (RTC) [46]. For instance, non-structural protein 13 (nsp13), belongs to helicase superfamily 1 and is putatively involved in viral RNA replication through RNA-DNA duplex unwinding [47] whereas nsp14 and nsp15 are exoribonuclease and endoribonuclease, respectively [48, 49]. nsp16 functions as a mRNA cap-1 methyltransferase [50]. All these proteins containing SNPs at several positions (Table 2) indicate that viral machinery for its RNA replication and processing is utmost evolved in strains from USA as compared to the other countries. Further, we analyzed the SNPs at protein level and interestingly in ORF1b protein, there were amino acid substitutions at P1327L, Y1364C and S2540F in USA isolates. One isolate namely USA0/MN1-MDH1/2020 (MT188341) carried amino-acid addition at 2540 position leading to shift in amino acid frame their onwards (Figure 2), which might affect the functioning of nsp16 (2′-O-MTase). But no changes were observed in Indian isolates, thus found similar to Chinese isolate. As the proteins involved in viral replication are evolving rapidly, this highlights the need to consider these mutants in order to develop the treatment strategies.

Direction of selection of SARS-CoV-2 genes

Our analysis revealed that ORF8 (121 a.a.) (dN/dS= 35.8) along with ORF3a (275 bp) (dN/dS= 8.95) showed highest dN/dS values among the nine ORFs thus, have much greater number of non-synonymous substitutions than the synonymous substitution (Figure 3D). Values of dN/dS >>1 are indicative of strong divergent lineage [51]. Thus, both of these proteins are evolving under high selection pressure and are highly divergent ORFs across strains. Two other proteins, ORF1ab polyprotein (dN/dS= 0.996, 0.575) and S protein (dN/dS= 0.88) might confer selective advantage with host challenges and survival. The dN/dS rates nearly 1 and greater than 1 suggests that the strains are coping up with the challenges *i.e.*, immune responses and inhibitory environment of host cells [52]. The other gene clusters namely M-protein and orf1a polyprotein did not possess at least three unique sequences necessary for the analysis, hence, they should be similar across the strains. The two genes ORF1ab polyprotein encodes for protein translation and post translation modification found to be evolved which actively translates, enhance the multiplication and facilitates growth of virus inside the host. Similarly, the S protein which helps in the entry of virus to the host cells by surpassing the cell membrane found to be accelerated towards positive selection confirming the successful ability of enzyme to initiate the infection. Another positive diversifying gene N protein encodes for nucleocapsid formation which protects the genetic material of virus form host immune responses such as cellular proteases. Overall, the data represent that the growth and multiplication related genes are highly evolving. The other proteins with dN/dS values equal to zero suggesting a conserved repertoire of genes.

SARS-CoV-2-Host interactome unveils immunopathogenesis of COVID-19

Although the primary mode of infection is human to human transmission through close contact, which occurs via spraying of nasal droplets from the infected person, yet the primary site of infection and pathogenesis of SARS-CoV-2 is still not clear and under investigation. To explore the role of SARS-CoV-2 proteins in host immune evasion, the SARSCoV-2 proteins were mapped over host proteome database (Figure 3B & Table 3). We identified a total of 28 proteins from host proteome forming close association with 25 viral proteins present in 9 ORFs of SARS-CoV-2 (Figure 3C). The network was trimmed in Cytoscape v3.7.2 where only interacting proteins were selected. Only 12 viral proteins were found to interact with host proteins (Figure 3A). Detailed analysis of interactome highlighted 9 host proteins in direct association with 6 viral proteins. Further, the network was analyzed for identification of regulatory hubs based on degree analysis. We identified mitogen activated protein kinase 1 (MAPK1) and AKT proteins as major hubs forming 24 and 21 interactions in the network respectively, highlighting their crucial role in pathogenesis. Recently, Huang *et al*, demonstrated the role of Mitogen activated protein kinase (MAPK) in COVID-19 mediated blood immune responses in infected patients [53] and showed that MAPK activation certainly plays a major defense mechanism.

Gene Ontology based functional annotation studies predicted the role of direct interactions of several viral proteins with host proteins. One such protein is non-structural protein2 (nsp2) which directly interacts with host Prohibitin (PHB), a known regulator of cell proliferation and maintains functional integrity of mitochondria [54]. SARS-CoV nsp2 is also known for its interaction with host PHB1 and PHB2 [55]. Nsp2 is a methyltransferase like domain that is known to mediate mRNA cap 2'-O-ribose methylation to the 5'-cap structure of viral mRNAs. This N7-methylguanosine cap is required for the action of nsp16 (2'-O-methyltransferase) and nsp10 complex [56]. This 5'-capping of viral RNA plays a crucial role in escape of virus from innate immunity recognition [56]. Hence, nsp2 -is responsible for modulating host cell survival strategies by altering host cell environment [55]. Based on network predicted we propose nsp16/nsp10 interface as a better drug target for anti-coronavirus drugs corresponding to the prediction made by Chen and group (2011) [56].

Similarly, the viral protein Papain-like proteinase (PL-PRO) which has deubiquitinase and deISGylating activity is responsible for cleaving viral polyprotein into 3 mature proteins which are essential for viral replication [57]. Our study showed that PL-PRO directly interacts with PPP1CA which is a protein phosphatase that associates with over 200 regulatory host proteins to form highly specific holoenzymes. PL-PRO is also found to interact with TGFβ which is a beta transforming growth factor and promotes T- helper 17 cells (Th17) and regulatory T-cells (T_{reg}) differentiation [58]. Reports have shown the PL-PRO induced upregulation of TGFβ in human promonocytes via MAPK pathway result in pro-fibrotic responses [59]. This reflects that viral PL-PRO antagonises innate immune system and is directly involved in the pathogenicity of SARS-CoV-2 induced pulmonary fibrosis [56, 58]. Many COVID-19 patients develop acute respiratory distress syndrome (ADRS) which leads to pulmonary edema and lung failure [60, 61]. These symptoms are because of cytokine storm manifesting elevated levels of pro-inflammatory cytokines like IL6, IFNγ, IL17, IL1β etc [61]. These results are in agreement with our prediction where we found IL6 as an interacting partner. Our study also showed JAK1/2 as an interacting partner which is known for IFNγ signaling. It is well known that TGFβ along with IL6 and STAT3 promotes Th17 differentiation by inhibiting SOCS3 [62]. Th17 is a source of IL17, which is commonly found in serum samples of COVID-19 patients [61, 63]. Hence, our interactome is supported from these findings where we found SOCS3, STAT3, JAK1/2 as an interacting partner [64]. The results suggested that proinflammatory cytokine storm is one of the reasons for SARS-CoV-2 mediated immunopathogenesis.

In the next cycle of physical events the viral protein NC (nucleoprotein), which is a major structural part of SARV family associates with the genomic RNA to form a flexible, helical nucleocapsid. Interaction of this protein with SMAD3 leads to inhibition of apoptosis of SARS-CoV infected lung cells [65], which is a successful strategy of immune evasion by the virus. More complex and multiple associations of ORF7a viral protein which is a non-structural protein and known as growth factor for SARS family viruses, directly captures BCL2L1 which is a potent regulator of apoptosis. Tan *et al.* (2007) have shown that SARS-CoV ORF7a protein induces apoptosis by interacting with Bcl X_L protein which is responsible for lymphopenia, an abnormality found in SARS-CoV infected patients [66]. Another target of viral ORF7a protein is SGTA (Small glutamine-rich tetratricopeptide repeat) which is an ATPase regulator and promotes viral encapsulation [67]. Subordinate viral proteins M (Membrane), S (Glycoprotein) and ORF3a (viroporin) were found to interact with each other. This interaction is important for viral cell formation and budding [68, 69]. Studies have shown the localization of ORF3a protein in Golgi apparatus of SARS-CoV infected patients along with M protein and responsible for viral budding and cell injury [70]. ORF3a protein also targets the functioning of CAV1 (Caveolin 1), caveolae protein, acts as a scaffolding protein within caveolar membranes. CAV1 has been reported to be involved in viral replication, persistence, and the potential role in pathogenesis in HIV infection also [71]. Thus, ORF3a interactions will upregulate viral replication thus playing a very crucial role in pathogenesis. Multiple methyltransferase assembly viral proteins (nsp7, nsp8, nsp9, RdRp) which are nuclear structural proteins were observed to target the SPECC1 proteins and linked with cytokinesis and spindle formations during division. Thus, major viral assembly also targets the proteins linked with immunity and cell division. Taken together, we estimated that SARS-CoV-2 manipulate multiple host proteins for its survival while, their interaction is also a reason for immunopathogenesis.

Conclusions

As COVID-19 continues to impact virtually all human lives worldwide due to its extremely contagious nature, it has spiked the interest of scientific community all over the world to understand better the pathogenesis of the novel SARS-CoV-2. In this study, the analysis was performed on the genomes of the novel SARS-CoV-2 isolates recently reported from different countries to understand viral pathogenesis. With the limited data available so far, we observed no direct transmission pattern of the novel SARS-CoV-2 in the neighboring countries through our analyses of the phylogenomic relatedness of geographical isolates. The isolates from same locations were phylogenetically distant, for instance, isolates from the USA and China. Thus, there appears to be a mosaic pattern of transmission indicative of the result of infected human travel across different countries. As COVID-19 transited from epidemic to pandemic within a short time, it does not look surprising from the genome structures of the viral isolates. The genomes of six isolates, specifically from the USA, were found to harbor unique amino acid SNPs and showed amino acid substitutions in ORF1b protein and S-protein, while one of them also harbored an amino-acid addition. This is suggestive of the severity of the mutating viral genomes within the population of the USA. These proteins are directly involved in the formation of viral replication-transcription complexes (RTC). Therefore, we argue that the novel SARS-CoV-2 has fast evolving replicative machinery and that it is urgent to consider these mutants to develop strategies for COVID-19 treatment. The ORF1ab polyprotein protein and S-protein were also found to have dN/dS values approaching 1 and thus might confer a selective advantage to evade host responsive mechanisms. The construction of SARS-CoV-2-human interactome revealed that its pathogenicity is mediated by a surge in pro-inflammatory cytokine. It is predicted that major immune-pathogenicity mechanism by SARS-CoV-2 includes the host cell environment alteration by disintegration by signal transduction pathways and immunity evasion by several protection mechanisms. The mode of entry of this virus by S-proteins inside the host cell is still unclear but it might be similar to SARS CoV-1 like viruses. Lastly, we believe as more data accumulate for COVID-19 the evolutionary pattern will become much clear.

Authors Contribution

RL, RK, HV, VG, US conceived and designed the study. RK, HV, NS, US, VG, MS, SN, PH executed the analysis and prepared figures. RK, HV, RK, NS, US, VG, MS, SN, PH, CT, NN, SA, CDR, MV wrote the manuscript with contributions from all authors. YS and RKN provided time to time guidance.

Conflict of Interest

Authors declare no conflict of Interest

Acknowledgements

RK acknowledges Magadh University, Bodh Gaya for providing support. RL and US also acknowledge The National Academy of Sciences, India, for support under the NASI-Senior Scientist Platinum Jubilee Fellowship Scheme. NS, SN, CT acknowledge Council of Scientific and Industrial Research (CSIR), New Delhi for doctoral fellowships. HV would like to thank Ramjas College, University of Delhi, Delhi for providing support. VG and MS acknowledge Phixgen Pvt. Ltd. for research fellowship. PH would like to thank Maitreyi College, University of Delhi, Delhi for providing support.

References:

1. Wang Q, Zhang Y, Wu L, Niu S, Song C, Zhang Z, *et al.* Structural and functional basis of SARS-CoV-2 entry by using human ACE2. Cell. 2020; doi: 10.1016/j.cell.2020.03.045.
2. Wall AC, Park YJ, Tortorici MA, Wall A, McGuire AT, Vessler D. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. Cell. 2020; 180: 1-12.
3. Wu A, Peng Y, Huang B, Huang B, Ding X, Wang X, *et al.* Genome composition and divergence of the novel coronavirus (2019-nCoV) originating in China. Cell Host Microbe. 2020; 27: 325-328.

4. Tyrrell DA, Bynoe ML. Cultivation of viruses from a high proportion of patients with colds. Lancet. 1966; 1: 76–77.

5. Woo PC, Lau SK, Lam CS, Lau CC, Tsang AK, Lau JH, *et al.* Discovery of seven novel Mammalian and avian coronaviruses in the genus deltacoronavirus supports bat coronaviruses as the gene source of alphacoronavirus and betacoronavirus and avian coronaviruses as the gene source of gammacoronavirus and deltacoronavirus. J Virol. 2012; 86: 3995-4008.

6. Li F. Structure, function, and evolution of coronavirus spike proteins. Annu Rev Virol. 2016; 3: 237-261.

7. Tang Q, Song Y, Shi M, Cheng Y, Zhang W, Xia XQ. Inferring the hosts of coronavirus using dual statistical models based on nucleotide composition. Sci Rep. 2015; 5: 17155.

8. Fehr AR, Perlman S. Coronaviruses: an overview of their replication and pathogenesis. Methods Mol Biol. 2015; 1282: 1–23.

9. Lau SK, Woo PC, Li KS, Huang Y, Tsoi HW, Wong BH, *et al.* Severe acute respiratory syndrome coronavirus-like virus in Chinese horseshoe bats. Proc Natl Acad Sci U S A. 2005; 102: 14040–14045.

10. Meyer B, Muller MA, Corman VM, Reusken CB, Ritz D, Godeke GJ, *et al.* Antibodies against MERS coronavirus in dromedary camels, United Arab Emirates, 2003 and 2013. Emerg Infect Dis. 2014; 20: 552–559.

11. Zhang C, Zheng W, Huang X, Bell EW, Zhou X, Zhang Y. Protein structure and sequence re-analysis of 2019-nCoV genome does not indicate snakes as its intermediate host or the unique similarity between its spike protein insertions and HIV-1. 2020; arXiv:2002.03173[q-bio.GN].

12. Neuman BW, Adair BD, Yoshioka C, Quispe JD, Orca G, Kuhn P, *et al.* Supramolecular architecture of severe acute respiratory syndrome coronavirus revealed by electron cryomicroscopy. J virol. 2006; 80:7918–7928.

13. Barcena M, Oostergetel GT, Bartelink W, Faas FG, Verkleij A, Rottier PJ, *et al.* Cryo-electron tomography of mouse hepatitis virus: Insights into the structure of the coronavirion. Proc Natl Acad Sci U S A. 2009; 106: 582–587.

14. Chen Y, Liu Q, Guo D. Emerging coronaviruses: Genome structure, replication, and pathogenesis. J. Med. Virol. 2020; 92: 418-423.

15. Collins AR, Knobler RL. Powell H, Buchmeier MJ. Monoclonal antibodies to murine hepatitis virus-4 (strain JHM) define the viral glycoprotein responsible for attachment and cell--cell fusion. Virology. 1982; 119: 358–371.

16. Neuman BW, Kiss G, Kunding AH, Bhella D, Baksh MF, Connelly S, *et al.* A structural analysis of M protein in coronavirus assembly and morphology. J Struct Biol. 2011; 174: 11–22.

17. Ruch TR, Machamer CE. The coronavirus E protein: assembly and beyond. Viruses. 2012; 4: 363-382.

18. McBride R, van Zyl M, Fielding BC. The coronavirus nucleocapsid is a multifunctional protein. Viruses. 2014; 6: 2991–3018.

19. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, *et al.* Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancelet. 2020; 395: 565-574.

20. Lv L, Li G, Chen J, Liang X, Li Y. Comparative genomic analysis revealed specific mutation pattern between human coronavirus SARS-CoV-2 and Bat-SARSr-CoV RaTG13. bioRxiv, 2020; doi: https://doi.org/10.1101/2020.02.27.969006.

21. Sah R, Alfonso J, Rodriguez-Morales, Jha R, Daniel KW, Chu HG, *et al.* Complete genome sequence of a 2019 novel coronavirus (SARS-CoV-2) strain isolated in Nepal. Microbiol Res Announce. 2020; 9: e00169-20.

22. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014; 30: 2068-2069.

23. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013; 29: 1072-1075.

24. Edgar E. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004; 32:1792-1797.

25. Pond SL, Frost SD, Muse, SV. HyPhy: hypothesis testing using phylogenies. Bioinformatics. 2005; 21: 676-679.

26. Weaver S, Shank SD, Spielman SJ, Li M, Muse SV, Kosakovsky Pond SL. Datamonkey 2.0: A Modern Web Application for Characterizing Selective and Other Evolutionary Processes. Mol Biol Evol. 2018; 35: 773-777.

27. Nakamura Y, Tomii K. Parallelization of MAFFT for large-scale multiple sequence alignments. Bioinformatics. 2018; 34: 2490–2492.

28. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, *et al.* Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015; 31: 3691-3693.

29. Tamura K, Nei M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol. 1993; 10: 512-526.

30. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. Mol Biol Evol. 2016; 33: 1870-1874.

31. Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. 2016; 44: W242-245.

32. Zhou Z, Alikhan NF, Sergeant MJ, Luhmann N, Vaz C, Francisco AP, *et al.* GrapeTree: visualization of core genomic relationships among 100,000 bacterial pathogens. Genome Res. 2018; 28: 1395-1404.

33. Treangen TJ, Ondov BD, Koren S, Phillippy AM. The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. Genome Biol. 2014; 15: 524.

34. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic Local Alignment Search Tool. J Mol Biol. 1990; 215: 403-410.

35. Stano M, Beke G, Klucar L. viruSITE—integrated database for viral genomics. Database 2, 2016; article ID baw162; doi:10.1093/database/baw162.

36. Cook HV, Doncheva NT, Szklarczyk D, von Mering C, Jensen LJ. Viruses.STRING: A Virus-Host Protein-Protein Interaction Database. Viruses. 2018; 10: 519.

37. Kerrien S, Aranda B, Breuza, L, Bridge A, Broackes-Carter, F, Chen C, et al. The IntAct molecular interaction database in 2012. Nucleic Acids Res. 2013; 41: D43-D47.

38. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, *et al.* STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 2015; **43:** D447-452.

39. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res. 2005; 33: D54–D58.

40. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003; 13: 2498-2504.

41. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000; 28: 27–30.

42. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 2016; 44: D457–D462.

43. UniProt Consortium. The universal protein resource (UniProt). Nucleic Acids Res. 2007; **36:** D190-195.

44. Ren LL, Wang YM, Wu ZQ, Xiang ZC, Guo L, Xu T, *et al.* Identification of a novel coronavirus causing severe pneumonia in human: a descriptive study. Chin Med J (Engl). 2020; doi: 10.1097/CM9.0000000000000722.

45. Gorbalenya AE, Baker SC, Baric RS, de Groot RJ, Drosten C, Gulyaeva AA, *et al*. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. Nat Microbiol. 2020; 5: 536-544.

46. Snijder EJ, Decroly E, Ziebuhr J. The non-structural proteins directing coronavirus RNA synthesis and processing. Adv Virus Res. 2016; 96: 59-126.

47. Jang KJ, Jeong S, Kang DY, Sp N, Yang YM, Kim DE. A high ATP concentration enhances the cooperative translocation of the SARS coronavirus helicase nsP13 in the unwinding of duplex RNA. Sci Rep. 2020; 10: 1-13.

48. Becares M, Pascual-Iglesias A, Nogales A, Sola I, Enjuanes L, Zuñiga S. Mutagenesis of coronavirus nsp14 reveals its potential role in modulation of the innate immune response. J Virol. 2016; 90: 5399-5414.

49. Athmer J, Fehr AR, Grunewald M, Smith EC, Denison MR, Perlman S. In situ tagged nsp15 reveals interactions with coronavirus replication/transcription complex-associated proteins. mBio. 2017; 8: e02320-16.

50. Von Grotthuss M, Wyrwicz LS, Rychlewski L. mRNA cap-1 methyltransferase in the SARS genome. Cell. 2003; 113: 701-702.

51. Kryazhimskiy S, Plotkin JB. The Population Genetics of dN/dS. PLoS Genet. 2008; 4: e1000304.

52. Kosakovsky Pond SL. Frost SD. (2005). Not So Different After All: A Comparison of methods for detecting amino acid sites under selection. Mol Biol Evol. 2005; 22:1208–1222.

53. Huang L, Shi Y, Gong B, Jiang L, Liu X, Yang J, *et al.* Blood single cell immune profiling reveals the interferon-MAPK pathway mediated adaptive immune response for COVID-19. BMJ. 2020; ,doi: https://doi.org/10.1101/2020.03.15.20033472.

54. Tatsuta T, Model K, Langer T. Formation of membrane-bound ring complexes by prohibitins in mitochondria. Mol Biol Cell. 2005; 16: 248-259.

55. Cornillez-Ty CT, Liao L, Yates JR3rd, Kuhn P, Buchmeier MJ. Severe acute respiratory syndrome coronavirus nonstructural protein 2 interacts with a host protein complex involved in mitochondrial biogenesis and intracellular signaling. J Virol. 2009; 83: 10314-10318.

56. Chen Y, Su C, Ke M, Jin X, Xu L, Zhang Z*, et al.* Biochemical and structural insights into the mechanisms of SARS coronavirus RNA ribose 2'-O-methylation by nsp16/nsp10 protein complex. PLoS Pathog. 2011; 7: e1002294.

57. Fung TS, Liu DX. Human Coronavirus: Host-Pathogen Interaction. Annu Rev Microbiol. 2019; 73: 529–557.

58. Wan YY, Flavell RA. ‘Yin-Yang’ functions of transforming growth factor-beta and T regulatory cells in immune regulation. Immunol Rev. 2007; 220: 199-213.

59. Li SW, Wang CY, Jou YJ, Jou YJ, Tang TC, Huang SH, *et al.* SARS coronavirus papain-like protease induces Egr-1-dependent up-regulation of TGF-β1 via ROS/p38 MAPK/STAT3 pathway. Sci Rep. 2016; 6: 25754.

60. Xu Z, Shi L, Wang Y, Zhang J, Huang L, Zhang C, *et al.* Pathological findings of COVID-19 associated with acute respiratory distress syndrome. Lancet Respir Med. 2020; doi: 10.1016/S2213-2600(20)30076-X.

61. Huang Y, Wang X, Li X, Ren L, Zhao J, Hu Y, *et al.* Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet. 2020; 395: 497-506.

62. Qin H, Wang L, Feng T, Elson CO, Niyongere SA, Lee AJ, *et al.* TGF-beta promotes Th17 cell development through inhibition of SOCS3. J Immunol. 2009; 183: 97–105.

63. Josset L, Menachery VD, Gralinski LE, Agnihothram S, Sova P, Carter VS, *et al.* Cell host response to infection with novel human coronavirus EMC predicts potential antivirals and important differences with SARS coronavirus. mBio. 2013; 4: e00165-13.

64. Prompetchara E, Ketloy C, Palaga T. Immune responses in COVID-19 and potential vaccines: Lessons learned from SARS and MERS epidemic. Asian Pac J Allergy Immunol. 2020; 38: 1-9.

65. Zhao X, Nicholls JM, Chen Y. Sars-cov nucleocapsid protein interacts with smad3 and modulates TGF-β signaling. J Biol Chem. 2008; 283: 3272-80.

66. Tan YX, Tan THP, Lee MJ-R, Tham PY, Gunalan V, Druce J, *et al*. Induction of apoptosis by the severe acute respiratory syndrome coronavirus 7a protein is dependent on its interaction with the Bcl-X_L protein. J Virol. 2007; 81: 6346-6355.

67. Fielding BC, Gunalan V, Tan TH, Chou CF, Shen S, Khan S, *et al.* Severe acute respiratory syndrome coronavirus protein 7a interacts with hSGT. Biochem Biophys Res Commun. 2006; 343: 1201-8.

68. de Haan CA, Smeets M, Vernooij F, Vennema H, Rottier PJ. Mapping of the coronavirus membrane protein domains involved in interaction with the spike protein. J Virol. 1999; 73: 7441–7452.

69. Klumperman J, Locker JK, Meijer A, Horzinek MC, Geuze HJ, Rottier PJ. Coronavirus M proteins accumulate in the Golgi complex beyond the site of virion budding. J Virol. 1994; 68: 6523–6534.

70. Yuan X, Li J, Shan Y, Yang Z, Zhao Z, Chen B, *et al.* Subcellular localization and membrane association of SARS-CoV 3a protein. Virus Res. 2005; 109: 191-202.

71. Mergia A. The Role of Caveolin 1 in HIV Infection and Pathogenesis. Viruses. 2017; 9: 129.

Figure legends

Figure 1: A) Core genome based phylogenetic analysis of SARS-CoV-2 isolates using the Maximum Likelihood method based on the Tamura-Nei model. The analysis involved 95 SARS-CoV-2 sequences with a total of 28451 nucleotide positions. Bootstrap values more than 70% are shown on branches as blue dots with sizes corresponding to the bootstrap values. The

coloured circle represents the country of origin of each isolate. The two isolates from Wuhan are marked separately on the outside of the ring. B) SNP based phylogeny of SARS-CoV-2 isolates. Highly similar genomes of coronaviruses were taken as input by Parsnp. Whole-genome alignments were made using libMUSCLE aligner using the annotated genome of MT121215 strain as reference. Parsnp identifies the maximal unique matches (MUMs) among the query genomes provided in a single directory. As only genomes within a specified MUMI distance threshold are recruited, option -c to force include all the strains was used. The output phylogeny based on Single nucleotide polymorphisms was obtained following variant calling on core-genome alignment. C) The minimum spanning tree generated using Maximum Likelihood method and Tamura-Nei model showing the genetic relationships of SARS-CoV-2 isolates with their geographical distribution.

Figure 2: Multiple sequence alignment of ORF1b protein showing amino acid substitutions at three positions: P1327L, Y1364C and S2540F. The isolate USA/MN1-MDH1/2020 (MT188341) showed an amino-acid addition leading to change in amino acid frame from position 2540 onwards.

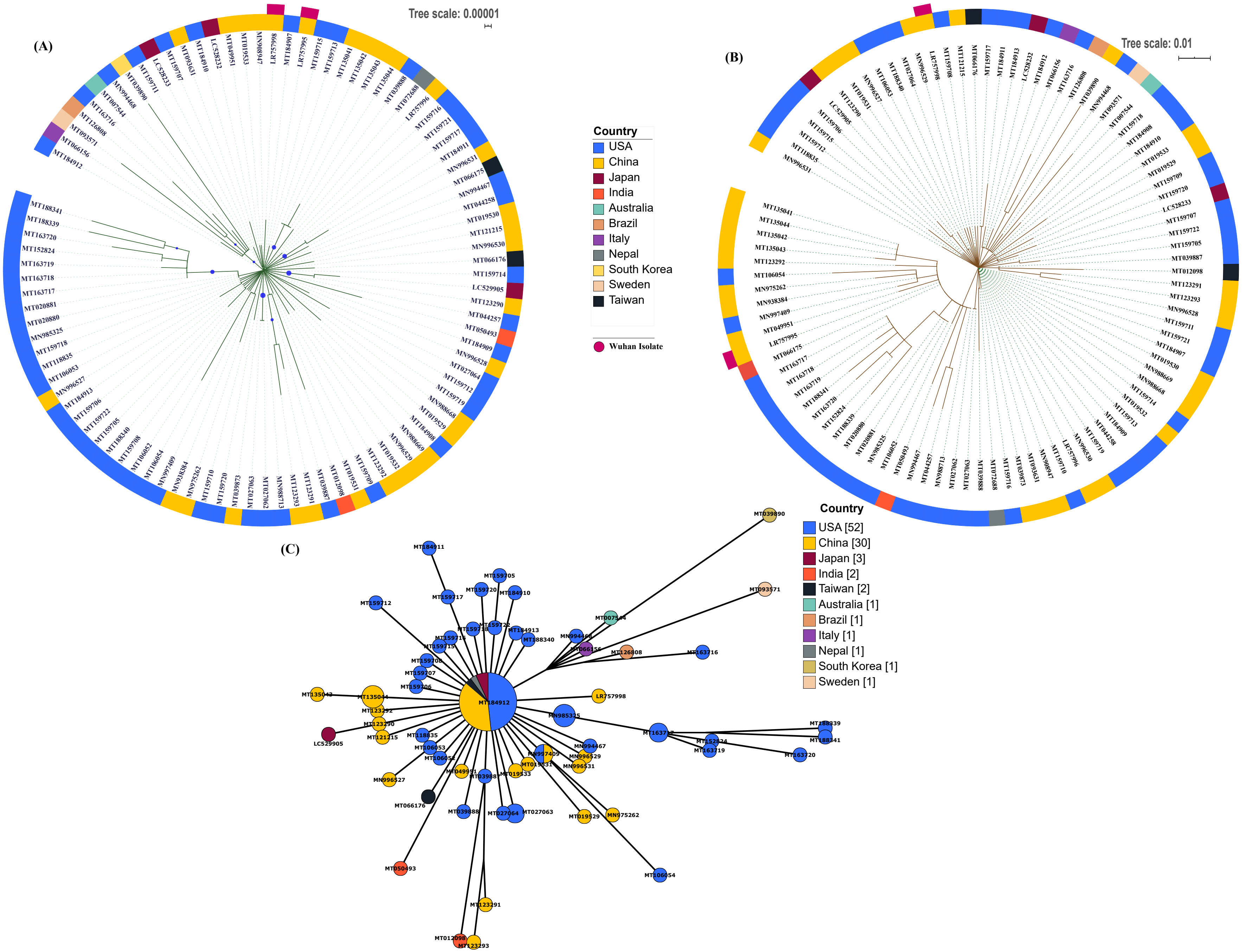
Figure 3: (A) SARS-CoV-2 -Host interactome analysis. Sub-set network highlighting SARS-CoV-2 and host nodes targeting each other. In total, nine direct interactions were observed (shown with red arrows). (B) Circular genome map of SARS-CoV-2 with genome size of 29.8 Kb generated using CGView. The genome of SARS-CoV 2 is also compared with that of SARS-CoV genome. The ruler for genome size is shown as innermost ring where Kbp stands for kilo base pairs. Concentric circles from inside to outside denote: SARS-CoV genome (used as reference), G + C content, G + C skew, predicted ORFs in SARS-CoV-2 genome and annotated CDS in SARS-CoV-2 genome. Gaps in alignment are shown in white. The positive and negative deviation from mean G + C content and G + C skew are represented with outward and inward peaks respectively. (C) SARS-CoV 2 and Host interactome generated using Virus.STRING interaction database v10.5. Both interacting and non-interacting viral proteins are shown. (D) Estimation of purifying natural selection pressure in nine coding sequences of SARS-CoV-2. dN/dS values are plotted as a function of dS.

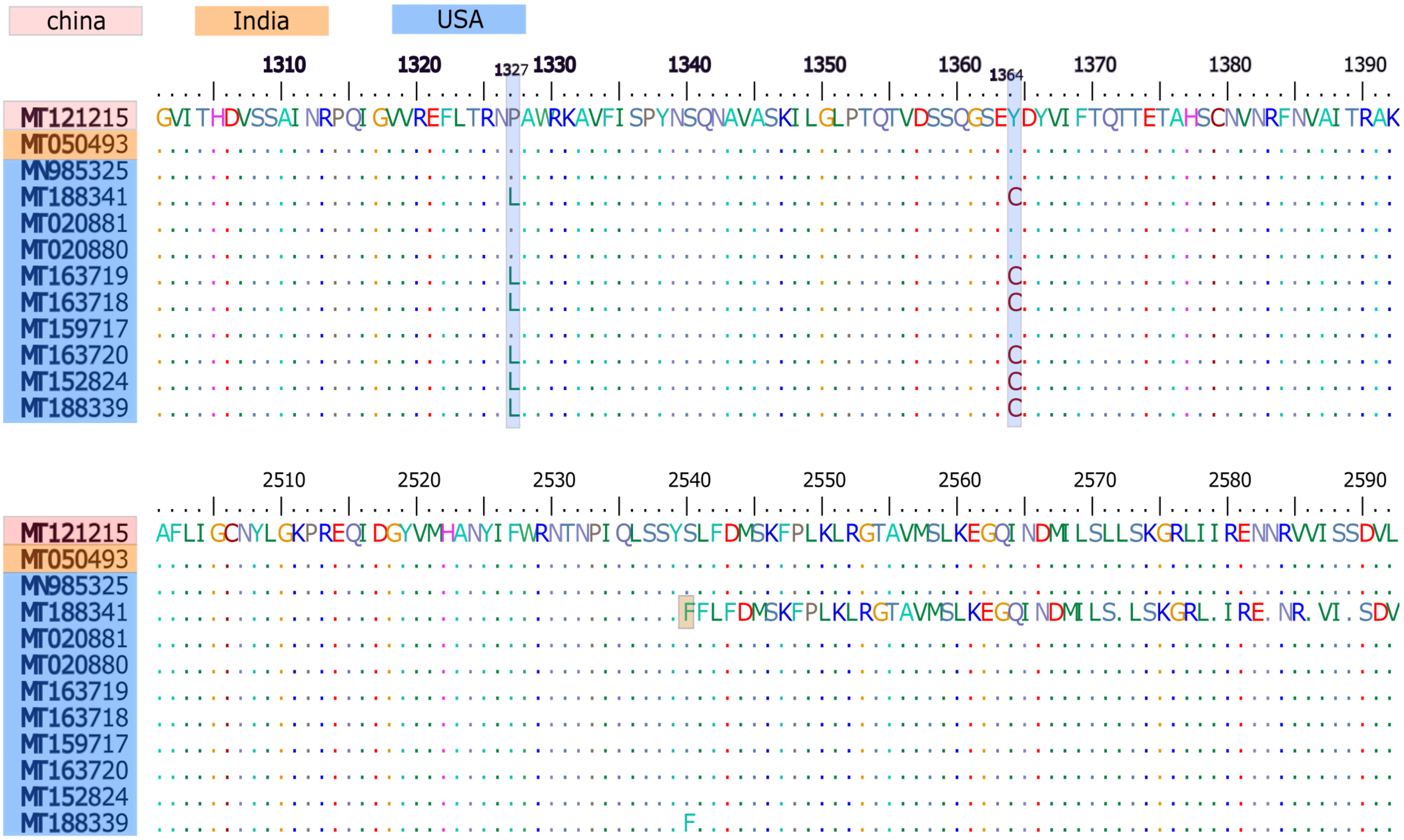
Tables Legends

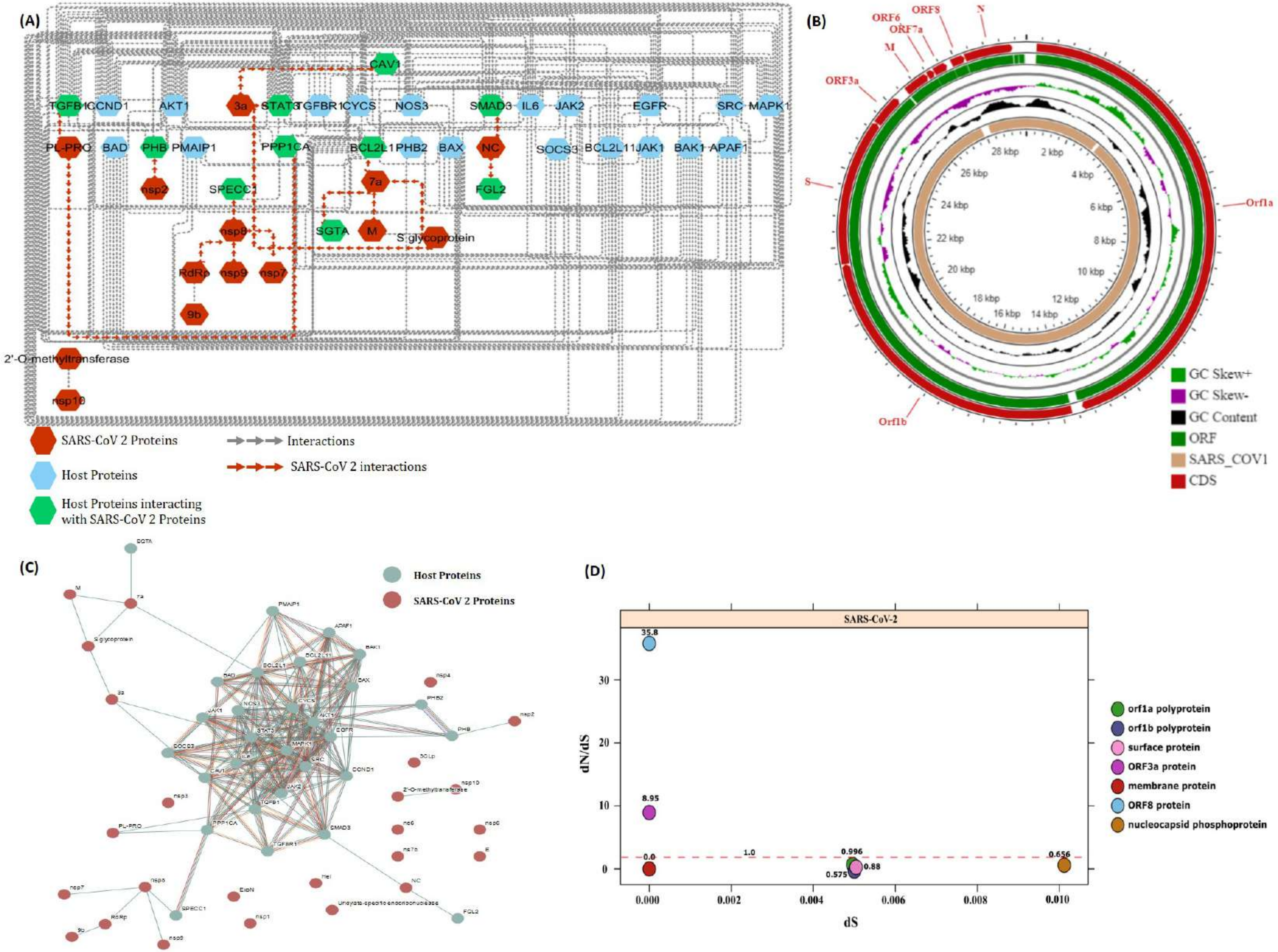
Table 1: General genomic attributes of SARS-CoV-2 strains.

Table 2: Major mutations present in different isolates of SARS-CoV-2 at different locations.

Table 3: Description of SARS-CoV2 proteins and its similarity in comparison to SARS-CoV used for PPI prediction.


(A)
Tree scale: 0.00001
MT184912
MT066156
MT093571
MT126808
MT163716
MT007544
MN994468
MT039890
MT159711
LC528233
MT159707
MT093631
MT184910
LC528232
MT049951
MT019533
MN908947
LR757998
MT184907
LR757995
MT159715
MT159713
MT135041
MT135042
MT135043
MT135044
MT039888
MT072688
LR757996
MT159716
MT159721
MT159717
MT184911
MN996531
MT066175
MN994467
MT044258
MT019530
MT121215
MN996530
MT066176
MT159714
LC529905
MT123290
MT044257
MT050493
MT184909
MN996528
MT027064
MT159712
MT159719
MN988668
MT019529
MT184908
MN988669
MN996529
MT019532
MT123292
MT159709
MT019531
MT012098
MT039887
MT123291
MT123293
MN988713
MT027062
MT027063
MT039873
MT159720
MT159710
MN975262
MN938384
MN997409
MT106054
MT106052
MT159708
MT188340
MT159705
MT159722
MT159706
MT184913
MN996527
MT106053
MT118835
MT159718
MN985325
MT020880
MT020881
MT163717
MT163718
MT163719
MT152824
MT163720
MT188339
MT188341
Country
USA
China
Japan
India
Australia
Brazil
Italy
Nepal
South Korea
Sweden
Taiwan
Wuhan Isolate
(B)
Tree scale: 0.01
MN996531
MT118835
MT159712
MT159715
MT159706
LC529905
MT123290
MT019531
MN996527
MT106053
MT188340
MT027064
MN996529
LR757998
MT159708
MT121215
MT066176
MT159717
MT184911
MT184913
LC528232
MT184912
MT066156
MT163716
MT126808
MT039890
MN994468
MT093571
MT007544
MT159718
MT184908
MT184910
MT019533
MT019529
MT159709
MT159720
LC528233
MT159707
MT159722
MT159705
MT039887
MT012098
MT123291
MT123293
MN996528
MT159711
MT159721
MT184907
MT019530
MN988669
MN988668
MT159714
MT019532
MT159713
MT184909
MT044258
MT159719
MN996530
LR757996
MT159710
MN908947
MT093631
MT039873
MT159716
MT072688
MT039888
MT027063
MT027062
MN988713
MT044257
MN994467
MT050493
MT106052
MN985325
MT020881
MT020880
MT188339
MT152824
MT163720
MT188341
MT163719
MT163718
MT163717
MT066175
LR757995
MT049951
MN997409
MN938384
MN975262
MT106054
MT123292
MT135043
MT135042
MT135044
MT135041
(C)
Country
USA [52]
China [30]
Japan [3]
India [2]
Taiwan [2]
Australia [1]
Brazil [1]
Italy [1]
Nepal [1]
South Korea [1]
Sweden [1]
MT184911
MT159705
MT159720
MT184910
MT159712
MT159717
MT007544
MT039890
MT093571
MT159719
MT159722
MT184913
MT188340
MN994468
MT066156
MT126808
MT163716
MT159716
MT159715
MT159708
MT159707
MT159706
MT135043
MT135044
MT123292
MT123290
MT121215
LC529905
MT184912
LR757998
MN985325
MT163717
MT152824
MT163719
MT188339
MT188341
MT163720
MT118835
MT106053
MT106052
MN994467
MN997409
MN996529
MT019531
MN996531
MN996527
MT049951
MT039887
MT019533
MT066176
MT039888
MT027064
MT027063
MT019529
MN975262
MT106054
MT050493
MT123291
MT012098
MT123293


china
India
USA
1310
1320
1327
1330
1340
1350
1360
1364
1370
1380
1390
MT121215
MT050493
MN985325
MT188341
MT020881
MT020880
MT163719
MT163718
MT159717
MT163720
MT152824
MT188339
GVITHDVSSAINRPQIGVVREFLTRNPAWRKAVFISPYNSQNAVASKILGLPTQTVDSSQGSEYDYVIFTQTTETAHSCNVNRFNVAITRAK
L
C
2510
2520
2530
2540
2550
2560
2570
2580
2590
AFLIGCNYLGKPREQIDGYVMHANYIFWRNTNPIQLSSYSLFDMSKFPLKLRGTAVMSLKEGQINDMILSLLSKGRLIIRENNRVVISSDVL
FFLFDMSKFPLKLRGTAVMSLKEGQINDMILS.LSKGRL.IRE.NR.VI.SDV
F


(A)
CAV1
TGFB1
CCND1
AKT1
3a
STAT3
TGFBR1
CYCS
NOS3
SMAD3
IL6
JAK2
EGFR
SRC
MAPK1
PL-PRO
BAD
PHB
PMAIP1
PPP1CA
BCL2L1
PHB2
BAX
NC
SOCS3
BCL2L11
JAK1
BAK1
APAF1
nsp2
SPECC1
7a
FGL2
SGTA
M
S glycoprotein
nsp8
RdRp
nsp9
nsp7
9b
2'-O-methyltransferase
nsp10
SARS-CoV 2 Proteins
Host Proteins
Host Proteins interacting with SARS-CoV 2 Proteins
Interactions
SARS-CoV 2 interactions
(B)
ORF6
ORF8
N
ORF7a
M
ORF3a
S
Orf1a
Orf1b
28 kbp
2 kbp
26 kbp
4 kbp
24 kbp
6 kbp
22 kbp
8 kbp
20 kbp
10 kbp
18 kbp
12 kbp
16 kbp
14 kbp
GC Skew+
GC Skew-
GC Content
ORF
SARS_COV1
CDS
(C)
Host Proteins
SARS-CoV 2 Proteins
SGTA
M
7a
S glycoprotein
PMAIP1
APAF1
BAK1
BCL2L11
BCL2L1
BAD
BAX
nsp4
PHB2
3a
NOS3
CYCS
JAK1
AKT1
nsp2
STAT3
EGFR
PHB
SOCS3
MAPK1
IL6
SRC
3CLp
CAV1
CCND1
nsp10
2'-O-methyltransferase
nsp3
JAK2
TGFB1
PL-PRO
PPP1CA
SMAD3
ns6
nsp6
TGFBR1
ns7b
E
nsp7
nsp8
Hel
NC
ExoN
SPECC1
Uridylate-specific endoribonuclease
RdRp
FGL2
9b
nsp1
nsp9
(D)
SARS-CoV-2
35.8
8.95
0.0
1.0
0.996
0.575
0.88
0.656
30
20
10
0
dN/dS
0.000
0.002
0.004
0.006
0.008
0.010
dS
orf1a polyprotein
orf1b polyprotein
surface protein
ORF3a protein
membrane protein
ORF8 protein
nucleocapsid phosphoprotein

Sr. No.	Accession No.	Virus (SARS-CoV-2)	Country of origin	Genome Size (bp)	GC %	Isolation source	Date of Isolation
1	LC528232.1	Hu/DP/Kng/19-020	Japan	29902	37.98	Oronasopharynx	10/02/2020
2	LC528233.1	Hu/DP/Kng/19-027	Japan	29902	38.02	Oronasopharynx	10/02/2020
3	LC529905.1	TKYE6182_2020	Japan	29903	37.97	NA	01/2020
4	LR757995.1	Wuhan seafood market pneumonia virus	China: Wuhan	29872	38	NA	05/01/2020
5	MT163720.1	WA8-UW5/human/2020/USA	USA	29732	37.97	NA	01/03/2020
6	LR757998.1	Wuhan seafood market pneumonia virus	China: Wuhan	29866	37.99	NA	26/12/2020
7	MN908947.3	Wuhan-Hu-1	China	29903	37.97	NA	12/2019
8	MN938384.1	2019-nCoV_HKU-SZ-002a_2020	China:Shenzhen	29838	38.02	Oronasopharynx	10/01/2020
9	MN975262.1	2019-nCoV_HKU-SZ-	China	29891	37.98	Oronasopharynx	11/01/2020

		005b_2020					
10	MN985325.1	2019-nCoV/USA-WA1/2020	USA	29882	38	Oronasopharynx	19/01/2020
11	MN988668.1	2019-nCoV WHU01	China	29881	38	NA	02/01/2020
12	MN988669.1	2019-nCoV WHU02	China	29881	38	NA	02/01/2020
13	MN988713.1	2019-nCoV/USA-IL1/2020	USA	29882	37.99	Lung, Oronasopharynx	21/01/2020
14	MN994467.1	2019-nCoV/USA-CA1/2020	USA	29882	38	Oronasopharynx	23/12/2020
15	MN994468.1	2019-nCoV/USA-CA2/2020	USA	29883	37.99	Oronasopharynx	22/01/2020
16	MN996527.1	WIV02	China	29825	38.02	Lung	30/12/2019
17	MN996528.1	WIV04	China	29891	37.99	Lung	30/12/2019
18	MN996529.1	WIV05	China	29852	38.02	Lung	30/12/2019
19	MN996530.1	WIV06	China	29854	38.03	Lung	30/12/2019
20	MN996531.1	WIV07	China	29857	38.02	Lung	30/12/2019
21	MN997409.1	2019-nCoV/USA-AZ1/2020	USA	29882	37.99	Feces	22/01/2020
22	MT007544.1	Australia/VIC01/2020	Australia	29893	37.97	NA	25/01/2020

23	MT012098.1	SARS-CoV-2/29/human/2020/IND	Kerala, India	29854	38.02	Oronasopharynx	27/01/2020
24	MT019529.1	BetaCoV/Wuhan/IPBCAMS-WH-01/2019	China	29899	37.98	Lung	23/12/2020
25	MT019530.1	BetaCoV/Wuhan/IPBCAMS-WH-02/2019	China	29889	38	Lung	30/12/2019
26	MT019531.1	BetaCoV/Wuhan/IPBCAMS-WH-03/2019	China	29899	37.98	Lung	30/12/2019
27	MT019532.1	BetaCoV/Wuhan/IPBCAMS-WH-04/2019	China	29890	37.99	Lung	30/12/2019
28	MT019533.1	BetaCoV/Wuhan/IPBCAMS-WH-05/2020	China	29883	37.99	Lung	01/01/2020
29	MT020880.1	2019-nCoV/USA-WA1-A12/2020	USA	29882	38	Oronasopharynx	25/01/2020
30	MT020881.1	2019-nCoV/USA-WA1-F6/2020	USA	29882	38	Oronasopharynx	25/01/2020

31	MT027062.1	2019-nCoV/USA-CA3/2020	USA	29882	38	Oronasopharynx	29/01/2020
32	MT027063.1	2019-nCoV/USA-CA4/2020	USA	29882	38	Oronasopharynx	29/01/2020
33	MT027064.1	2019-nCoV/USA-CA5/2020	USA	29882	37.99	Oronasopharynx	29/01/2020
34	MT039873.1	HZ-1	China	29833	38.02	Lung, Oronasopharynx	20/01/2020
35	MT039887.1	2019-nCoV/USA-WI1/2020	USA	29879	38	Oronasopharynx	31/01/2020
36	MT039888.1	2019-nCoV/USA-MA1/2020	USA	29882	37.99	Oronasopharynx	29/01/2020
37	MT039890.1	SNU01	South Korea	29903	37.96	NA	01/2020
38	MT044257.1	2019-nCoV/USA-IL2/2020	USA	29882	38	Lung, Oronasopharynx	28/01/2020
39	MT044258.1	2019-nCoV/USA-CA6/2020	USA	29858	38	Oronasopharynx	27/01/2020
40	MT049951.1	SARS-CoV-2/Yunnan-01/human/2020/CHN	China	29903	37.97	Lung, Oronasopharynx	17/01/2020
41	MT050493.1	SARS-CoV-2/166/huma	Kerala, India	29851	38.01	Oronasopharynx	31/01/2020

		n/2020/IND					
42	MT066156.1	SARS-CoV-2/NM	Italy	29867	38.01	Lung, Oronasopharynx	30/01/2020
43	MT066175.1	SARS-CoV-2/NTU01/2020/TWN	Taiwan	29870	38.01	NA	31/01/2020
44	MT066176.1	SARS-CoV-2/NTU02/2020/TWN	Taiwan	29870	38.01	NA	05/02/2020
45	MT072688.1	SARS0CoV-2/61-TW/human/2020/ NPL	Nepal	29811	38.02	Oronasopharynx	13/02/2020
46	MT093571.1	SARS-CoV-2/01/human/2020/SWE	Sweden	29886	38	NA	07/02/2020
47	MT093631.2	SARS-CoV-2/WH-09/human/2020/CHN	China	29860	38.02	Oronasopharynx	08/01/2020
48	MT106052.1	2019-nCoV/USA-CA7/2020	USA	29882	37.99	Oronasopharynx	06/02/2020
49	MT106053.1	2019-nCoV/USA-CA8/2020	USA: CA	29882	38	Oronasopharynx	10/02/2020
50	MT106054.1	2019-nCoV/USA-TX1/2020	USA:TX	29882	38	Lung, Oronasopharynx	11/02/2020

51	MT118835.1	2019-nCoV/USA-CA9/2020	USA: CA	29882	38	Lung	23/02/2020
52	MT121215.1	SARS-CoV-2/SH01/human/2020/CHN	China	29945	37.91	Oronasopharynx	02/02/2020
53	MT123290.1	SARS-CoV-2/IQTC01/human/2020/CHN	China	29891	38	Oronasopharynx	05/02/2020
54	MT123291.2	SARS-CoV-2/IQTC02/human/2020/CHN	China	29882	37.99	Lung	29/01/2020
55	MT123292.2	SARS-CoV-2/QT	China	29923	38.02	Lung, Oronasopharynx	27/01/2020
56	MT123293.2	SARS-CoV-2/IQTC03/human/2020/CHN	China	29871	38	Feces	29/01/2020
57	MT126808.1	SARS-CoV-2/SP02/human/2020/BRA	Brazil	29876	38	Oronasopharynx	28/02/2020
58	MT135041.1	SARS-CoV-2/105/huma	China:Beijing	29903	37.97	NA	26/01/2020

		n/2020/CHN					
59	MT135042.1	SARS-CoV-2/231/human/2020/CHN	China:Beijing	29903	37.97	NA	28/01/2020
60	MT135043.1	SARS-CoV-2/233/human/2020/CHN	China:Beijing	29903	37.97	NA	28/01/2020
61	MT135044.1	SARS-CoV-2/235/human/2020/CHN	China:Beijing	29903	37.97	NA	28/01/2020
62	MT152824.1	SARS-CoV-2/WA2/human/2020/USA	USA:WA	29878	38	Mid nasal swab	24/02/2020
63	MT159705.1	2019-nCoV/USA-CruiseA-7/2020	USA	29882	37.99	Oronasopharynx	17/02/2020
64	MT159706.1	2019-nCoV/USA-CruiseA-8/2020	USA	29882	38	Oronasopharynx	17/02/2020
65	MT159707.1	2019-nCoV/USA-CruiseA-	USA	29882	38	Oronasopharynx	17/02/2020

		10/2020					
66	MT159708.1	2019-nCoV/USA-CruiseA-11/2020	USA	29882	38	Oronasopharynx	17/02/2020
67	MT159709.1	2019-nCoV/USA-CruiseA-12/2020	USA	29882	38	Oronasopharynx	20/02/2020
68	MT159710.1	2019-nCoV/USA-CruiseA-9/2020	USA	29882	38	Oronasopharynx	17/02/2020
69	MT159711.1	2019-nCoV/USA-CruiseA-13/2020	USA	29882	38	Oronasopharynx	20/02/2020
70	MT159712.1	2019-nCoV/USA-CruiseA-14/2020	USA	29882	37.99	Oronasopharynx	25/02/2020
71	MT159713.1	2019-nCoV/USA-CruiseA-15/2020	USA	29882	38	Oronasopharynx	18/02/2020
72	MT159714.1	2019-nCoV/USA-CruiseA-16/2020	USA	29882	38	Oronasopharynx	18/02/2020
73	MT159715.1	2019-nCoV/USA-CruiseA-	USA	29882	38	Oronasopharynx	24/02/2020

		17/2020					
74	MT159716.1	2019-nCoV/USA-CruiseA-18/2020	USA	29867	38	Oronasopharynx	24/02/2020
75	MT159717.1	2019-nCoV/USA-CruiseA-1/2020	USA	29882	37.99	Oronasopharynx	17/02/2020
76	MT159718.1	2019-nCoV/USA-CruiseA-2/2020	USA	29882	37.99	Oronasopharynx	18/02/2020
77	MT159719.1	2019-nCoV/USA-CruiseA-3/2020	USA	29882	38	Oronasopharynx	18/02/2020
78	MT159720.1	2019-nCoV/USA-CruiseA-4/2020	USA	29882	37.99	Oronasopharynx	21/02/2020
79	MT159721.1	2019-nCoV/USA-CruiseA-5/2020	USA	29882	38	Oronasopharynx	21/02/2020
80	MT159722.1	2019-nCoV/USA-CruiseA-6/2020	USA	29882	37.99	Oronasopharynx	21/02/2020
81	MT163716.1	SARS-CoV-2/WA3-	USA:WA	29903	37.95	NA	27/02/2020

		UW1/human/2020/USA					
82	MT163717.1	SARS-CoV-2/WA4-UW2/human/2020/USA	USA:WA	29897	37.97	NA	28/02/2020
83	MT163718.1	SARS-CoV-2/WA6-UW3/human/2020/USA	USA:WA	29903	37.97	NA	29/02/2020
84	MT163719.1	SARS-CoV-2/WA7-UW4/human/2020/USA	USA:WA	29903	37.97	NA	01/03/2020
85	LR757996.1	Wuhan seafood market pneumonia virus	China: Wuhan	29732	37.96	NA	01/01/2020
86	MT184907.1	2019-nCoV/USA-CruiseA-19/2020	USA	29882	38	Oronasopharynx	18/02/2020
87	MT184908.1	2019-nCoV/USA-CruiseA-	USA	29880	38	Oronasopharynx	17/02/2020

		21/2020					
88	MT184909.1	2019-nCoV/USA-CruiseA-22/2020	USA	29882	38	Oronasopharynx	21/02/2020
89	MT184910.1	2019-nCoV/USA-CruiseA-23/2020	USA	29882	37.99	Oronasopharynx	18/02/2020
90	MT184911.1	2019-nCoV/USA-CruiseA-24/2020	USA	29882	37.97	Oronasopharynx	17/02/2020
91	MT184912.1	2019-nCoV/USA-CruiseA-25/2020	USA	29882	38	Oronasopharynx	17/02/2020
92	MT184913.1	2019-nCoV/USA-CruiseA-26/2020	USA	29882	37.99	Oronasopharynx	24/02/2020
93	MT188339.1	USA/MN3-MDH3/2020	USA:MN	29783	38.01	Oronasopharynx	07/03/2020
94	MT188340.1	USA/MN2-MDH2/2020	USA:MN	29845	37.98	Oronasopharynx	09/03/2020
95	MT188341.1	USA/MN1-MDH1/2020	USA:MN	29835	37.99	Oronasopharynx	05/03/2020

Strains having major mutations	**Protein**	**Position in reference genome**	**Variant Nucleotide different from reference**	**Nucleotide in Reference Genome**
MT188341; MN985325; MT020881; MT020880; MT163719; MT163718; MT163717; MT152824; MT163720; MT188339	NSP14	18060	T	C
MT188341; MT163719; MT163718; MT163717; MT152824; MT163720; MT188339;	NSP13	17747	T	C
MT188341; MT163719; MT163718; MT163717; MT152824; MT163720; MT188339;	NSP13	17858	G	A
MT188341	NSP13	16467	G	A
Several Strains under study	NSP3	6026	C	T
MT039888	NSP3	3518	T	G
MT039888	NSP3	17423	G	A
MT163719	NSP15	20281	G	T
MT188339	NSP16	21147	C	T
MT188341	S-Protein	23185	T	C
MT163720	S-Protein	23525	T	C
MT188339	S-Protein	22432	T	C
MT159716	S-Protein	22033	A	C
MT050493 (INDIAN)	S-Protein	24351	T	C

CDS	SARS-CoV (NC_004718.3)		SARS-CoV 2 (MT121215.1)		Similarity %
	Positions	Protein ID	Positions	Protein ID	
Orf1a polyprotein	265-21482	NP_828849.2	266-13468, 13468-21555	QII57165.1	86
Nsp1	265-804	NP_828860.2	266-805		84.44
Nsp2	805-2718	NP_828861.2	806-2719		68.34
Nsp3/PL-PRO	2719-8484	NP_828862.2	2720-8554		75.77
Nsp4	8485-9984	NP_904322.1	8555-10054		<80
Nsp5/3CLp	9985-10902	NP_828863.1	10055-10972		<90
Nsp6	10903-11772	NP_828864.1	10973-11842		88.15
Nsp7	11773-12021	NP_828865.1	11843-12091		98.80
Nsp8	12022-12615	NP_828866.1	12092-12685		97.47
Nsp9	12616-12954	NP_828867.1	12686-13024		97.35
Nsp10	12955-13371	NP_828868.1	13025-13441		97.12
Nsp12 (RdRp)	13372-13398, 13398-16166	NP_828869.1	13442-13468, 13468-16236		
Orf1b					

polyprotein					
Nsp13 (Hel)	16167-17969	NP_828870.1	16237-18039		99.83
Nsp14 (ExoN)	17970-19550	NP_828871.1	18040-19620		95.07
Nsp15	19551-20588	NP_828872.1	19621-20658		88.73
Nsp16(O-methyl)	20589-21482	NP_828873.2	20659-21552		93.29
S	21492-25259	NP_828851.1	21563-25384	QII57161.1	75.96
Sars3a/Orf3a	25268-26092	NP_828852.2	25393-26220		
Sars3b/Orf3b	25689-26153	NP_828853.1	25814-26281		78.68
E	26117-26347	NP_828854.1	26245-26472	QII57162.1	94.74
M	26398-27063	NP_828855.1	26523-27191	QII57163.1	90.54
Sars6	26913-26918	NP_828856.1			
Sars7a/Orf7	27273-27641	NP_828857.1	27394-27759		82.21
Sars7b/Orf8	27638-27772	NP_849175.1	27756-27878		87.10
N/Sars9a	28120-29388	NP_828858.1	28274-29533	QII57164.1	90.52
Sars9b	28130-28426	NP_828859.1			