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-110007, 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 -19 (COVID19) 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 14,500 lives and has impacted human lives worldwide. Development of universal vaccines against the novel SARS-CoV-2 holds utmost urgency to control COVID19 pandemic that appears to be more severe than any of the previous outbreaks of severe acute respiratory syndrome (SARS) and Middle-East respiratory syndrome (MERS). However, as the novel SARS-CoV-2 displays high transmission rates, the underlying severity hidden in the SARS-CoV2 genomes is required to be fully understood. We studied the complete genomes of 95 strains of SARS-CoV-2 reported from different geographical regions worldwide to uncover the pattern of spread of the novel SARS-CoV-2 across the globe. We show that there is no direct transmission pattern of the virus among neighbouring 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, nsp14, nsp15, nsp16 (present in ORF1b polyprotein region) and S-Protein within 10 viral isolates from USA. These viral proteins are involved in RNA replication and processing, indicating highly evolved strains of the novel SARS-CoV-2 circulating in the population of USA than in other countries. Furthermore, we found an isolate from USA (MT188341) to carry frameshift mutation between positions 2540 and 2570 of nsp16 which functions as mRNA cap-1 methyltransferase (2′-O-MTase). Thus, we reason that the replicative machinery of the novel SARS-CoV-2 is fast evolving to evade host challenges and survival. These mutations are needed to be considered, otherwise it will be difficult to develop effective treatment strategies. The two proteins also had dN/dS values approaching 1- ORF1ab polyprotein (dN/dS= 0.996, 0.575) and S protein (dN/dS= 0.88) and might confer selective advantage to the virus. Through the construction of SARS-CoV-2-human interactome, we further reveal multiple host proteins (PHB, PPP1CA, TGF-beta, JACK1, JACK2, SOCS3,STAT3, JAK1-2, SMAD3, BCL2, CAV1 & SPECC1) which are manipulated by the virus proteins (nsp2, PL-PRO, N-protein, ORF7a, M-S-ORF3a complex, nsp7-nsp8-nsp9-RdRp complex) for mediating host immune mechanism for its survival.

Background

Since the current outbreak of pandemic coronavirus disease 19 (COVID-19) caused by Severe Acute Respiratory Syndrome-related Coronavirus (SARS-CoV-2), the assessment of biogeographical pattern of SARS-CoV-2 isolates and the mutations at nucleotide and protein level is of high interest to many research groups (Wang *et al.*, 2020; Wall *et al.*, 2020; Wu *et al.*, 2020). Coronaviruses (CoVs), members of *Coronaviridae* family, order *Nidovirales*, have been known as human pathogens from last six decades (Tyrrell and Bynoe,1966). Their target is not just limited to humans, but also to other mammals and birds (Woo *et al.*, 2012). Coronaviruses have been classified under alpha, beta, gamma and delta-coronavirus groups (Li, 2016) in which former two are known to infect mammals while latter two primarily infect bird species (Tang *et al.*, 2015). Illness in humans varies 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) (Fehr and Perlman, 2015). 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 (Lau *et al.*, 2005; Mayer *et al.*, 2014). 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 (Zhang *et al.*, 2020).

The inherently high recombination frequency and mutation rates of coronavirus genomes allows for their easy transmission from different intermediate hosts. Structurally, they are positive-sense single stranded RNA (ssRNA) virions with characteristic spikes projecting from the surface of capsid coating (Newman *et al.*, 2006, Barcena *et al.*, 2009). Their genome is nearly 27 to 31 Kb long, largest among the RNA viruses, with 5'cap and 3' polyA tail, for translation (Brian and Baric, 2005). Their spherical capsid and spikes give them crown-like appearance due to which they were named as 'corona', meaning 'crown' or 'halo' in *Latin*. Coronavirus consists four main proteins, spike (S), membrane (M), envelope (E) and nucleocapsid (N). The spike (~150 kDa) mediates its attachment to host receptor proteins (Colins *et al.*, 1982). Membrane protein (~25-30 kDa) attaches with nucleocapsid and maintains curvature of virus membrane (Neumann *et al.*, 2011). 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 (Ruch and Machamer, 2012).

N-protein, the fourth protein, helps in packaging of virus particles into capsids and promotes replicase-transcriptase complex (RTC) (McBride *et al.,* 2014).

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/emergencies/diseases/novel-coronavirus-2019/situation-reports/).World Health Organisation (WHO) has declared COVID19, the disease caused by the novel SARS-CoV-2 a pandemic, affecting more than 186 countries and territories where China has most reported cases 81,601 and Italy has highest mortality rate 9.26% (59,138 infected individuals, 5476 deaths) (WHO situation report-63). As on date (March 24, 2020), more than 0.3 million individuals have been infected by SARS-CoV-2 and nearly 14,510 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 is sufficiently diverged from SARS-CoV (Lu *et al.,* 2020). SARS-CoV-2 is assumed to have originated from bats, which serve as reservoir host of the virus (Lu *et al.,* 2020). Other studies have also reported the genome composition and divergence patterns of SARS-CoV-2 (Sah *et al.,* 2020; Wu *et al.,* 2020). 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 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 analysed the ORFs of the isolates to reveal unique point and frameshift mutations in the isolates from the USA. In addition, we analysed 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 functional implications of the viral infection host cells. The results obtained from the analyses indicate high severity of SARS-CoV-2 isolates with 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 and based on quality assessment two genomes with multiple Ns were removed from the study. Further the genomes were annotated using Prokka (Seemann, 2014). 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) (Seemann, 2014). Further the GC content information were generated using QUAST standalone tool (Gurevich *et al.*, 2013).

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 (Edgar *et al.*, 2004) and further processed for removing stop codons using HyPhy v2.2.4 (Pond *et al.*, 2005). Single-Likelihood Ancestor Counting (SLAC) method in Datamonkey v2.0 (Weaver *et al.*, 2018) (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

In order to infer the phylogeny, the core gene alignment was generated using MAFFT (Nakamura *et al*., 2018) present within the Roary Package (Page *et al*., 2015). Further, the phylogeny was inferred using the Maximum Likelihood method based and Tamura-Nei model (Tamura and Nei, 1993) in MEGAX (Kumar *et al*., 2016) and visualized in interactive Tree of Life (iTOL) (Letunic and Bork, 2016) and GrapeTree (Zhou *et al*., 2018).

In order to determine the single nucleotide polymorphism (SNP), whole-genome alignments were made using libMUSCLE aligner. For this, we used SARS-CoV-2/SH01/human/2020/CHN (Accession no. MT121215) as the reference. As only genomes within a specified MUMI distance threshold are recruited, we used option -c to force include all the strains and for output, it produces 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 (Treangen *et al*., 2014). Further, the tree was visualized in interactive Tree of Life (iTOL) (Letunic and Bork, 2016).

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 (Altschul *et al.,* 1990). 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 virSITE (Stano *et al.,* 2016) and interaction databases such as STRING Virus v10.5 (Cook *et al.,* 2018) and IntRact (Kerrien *et al.,* 2013) 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) signalling 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 (Szklarczyk *et al*., 2018; Cook *et al*., 2018).

Filtered proteins (confidence value: 0.7) were mapped to their Entrez ID (Maglott *et al.,* 2005) based on the NCBI database used for interactome analysis. HPI were stimulated using Cytoscape v.3.7.2 (Shannon *et al.,* 2003).

Functional enrichment analysis

Next, functional studies were performed using the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000; Kanehisa *et al.,* 2016) and Gene Ontology (GO) enrichment analyses using UniPROT database (UniProt Consortium, 2007) to evaluate the biological relevance and functional pathways of the HCoV-associated proteins. All functional analyses were performed using STRINGenrichment and STRINGify, plugin of Cytoscape v.3.7.2 (Shannon *et al.,* 2003). Network analysis was performed by tool NetworkAnalyzer, plugin of Cytoscape with orthogonal layout.

Results and Discussion

General genomic attributes of SARS-CoV-2

In this study, we analysed a total of 95 SARS-CoV-2 strains 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 (Figure 1). 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 was found to be 29879 ± 26.6 bp and 37.99 ± 0.018%, respectively. All these isolates were found to harbour 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 is in consensus with a recently published study (Ren *et al*., 2020). The ORF1a harbours 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) (Ren *et al.*, 2020). Similarly, ORF1b contains four putative nsp's namely nsp13 (helicase or Hel), nsp14 (3′-to-5′ exoribonuclease or ExoN), nsp15 and nsp16.

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 neighbour of SARS-CoV-2 as SARSr-CoV-RaTG13, a bat coronavirus (Gorbaleneya *et al.*, 2020). 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 2A & C) and single nucleotide polymorphisms (SNPs) (Figure 2B). Phylogenies obtained were annotated with country of isolation of each strain (Figure 2A & B). The phylogenetic clustering was found majorly concordant by both core-genome (Figure 1A) and SNP based methods (Figure 2B). 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 neighbouring 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 exception to these two strains, we did not find any connection between strains of neighbouring 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 2A & 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 2A & B). Similarly, Australian strain (MT007544) showed close clustering with USA strain (Figure 2A & B) and one strain from Taiwan (MT066175) clustered nearly with China strain (Figure 2B). Isolates from Italy (MT012098) and Brazil (MT126808) clustered with different USA strains (Figure 2A & 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 neighbouring 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 Brazil strains with USA strains 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 biasness is evident in analysed 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 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 analysed 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) (Snijder *et al.*, 2016). 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 (Jang *et al.*, 2020) whereas nsp14 and nsp15 are exoribonuclease and endoribonuclease, respectively (Becares *et al.*, 2016; Athmer *et al.*, 2017). nsp16 functions as a mRNA cap-1 methyltransferase (von Grotthuss *et al.*, 2003). 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 analysed 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 frameshift mutation between positions 2540 and 2570 (Figure 3), which might affect the functioning of nsp16 (2′-O-MTase). 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

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 4D). Values of dN/dS >>1 are indicative of strong divergent lineage (Kryazhimskiy *et. al.*, 2008). 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 (Pond *et al*., 2005). The other gene clusters namely M-protein and orf1a polyprotein did not possess at least three unique sequences necessary for the analysis, hence, 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 immunopathogenicity 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 SAR-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 (Table 3). We identified a total 28 proteins from host proteome forming close association with 25 viral proteins present in 9 ORFs of SARS-CoV-2 (Figure 4C). 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 4A). Detailed analysis of interactome highlighted 9 host proteins in direct association with 6 viral proteins. Further, the network was analysed 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 (Huang *et al.*, 2020) and showed that MAPK activation certainly plays a major defence 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 Prohbitin (PHB), a known regulator of cell proliferation and maintains functional integrity of mitochondria (Tatsuta *et al.*, 2005). SARS-CoV nsp2 is also known for its interaction with host PHB1 and PHB2 (Cromwell T *et al.*, 2009). Nsp2 is a methyltransferase like domain which 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 (Chen *et al.*, 2011). This 5'-capping of viral RNA plays a crucial role in escape of virus from innate immunity recognition (Chen *et al.*, 2011). Hence, nsp2 -is responsible for modulating host cell survival strategies by altering host cell environment (Cromwell *et al.*, 2009). Based on network predicted we propose nsp16/nsp10 interface as a better drug target for anti-coronavirus drugs corresponding to the prediction made by Cormwell and group (2011).

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. 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 (Zhang *et al.*, 2007). Reports have shown the PL-PRO induced upregulation of TGFβ in human promonocytes via MAPK pathway result in pro-fibrotic responses (Li *et al*., 2016). This reflect that viral PL-PRO antagonises innate immune system and is directly involved in the pathogenicity of SARS-CoV 2 induced pulmonary fibrosis (tu sing fung, 2019; Li *et al*., 2016). Many COVID-19 patients develop acute respiratory distress syndrome (ADRS) which leads to pulmonary edema and lung failure (Xu Z *et al*., 2020, Huang *et al*., 2019). These symptoms are because of cytokine storm manifesting elevated levels of pro-inflammatory cytokines like IL6, IFNγ, IL17, IL1β etc (Huang *et al*., 2019). 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 signalling. It is well known that TGFβ along with IL6 and STAT3 promotes Th17 differentiation by inhibiting SOCS3. Th17 is a source of IL17, which is commonly found in serum samples of COVID19 patients. Hence, our interactome is supported from these findings where we found SOCS3, STAT3, JAK1/2 as an interacting partner. The results suggested that proinflammatory cytokine storm is one of the reasons for SARS-CoV -2 mediated immune-pathogenicity.

In 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 (Zhao *et al*, 2008), which is a successful strategy of immune evasion by 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 a responsible for lymphopenia, an abnormality found in SARS-CoV infected patients (Tan *et al.*, 2007). Another target of viral ORF7a protein is SGTA (Small glutamine-rich tetratricopeptide repeat) which is an ATPase regulator and promotes viral encapsulation (Fielding 2006). 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 (de Haan *et al*., 1999, Klumperman *et al*., 1994). 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 (Yuan *et al*., 2005). ORF3a protein also targets the functioning of CAV1 (Caveolin 1), caveolae protein, act as a scaffolding protein within caveolar membranes. CAV1 have been reported to be involved in viral replication, persistence, and the potential role in pathogenesis in HIV infection also (Mergia, 2017). Thus, ORF3a interaction will upregulate viral replication thus playing 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 own survival while, their interaction is also a reason for immunopathogenicity.

Conclusions

As COVID19 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 viral pathogenesis. With the limited data so far available we observed no direct transmission pattern of the novel SARS-CoV-2 in the neighbouring countries through our analyses of the phylogenomic relatedness of geographical isolates. The isolates from same locations were distant phylogenetically, for instance, isolates from 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 COVID19 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 USA, were found to harbour unique amino acid SNPs and showed amino acid substitutions in ORF1b protein and S-protein, while one of them also harboured a frameshift mutation. This is suggestive of the severity of the mutating viral genomes within the population of USA. These proteins are directly involved in 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 in order to develop strategies for COVID19 treatment. The ORF1ab polyprotein protein and S-protein were also found to have dN/dS values approaching to 1 and thus might confer selective advantage to evade host responsive mechanisms. The construction of SARS-CoV-2-human interactome revealed that its pathogenicity is mediated by 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 host cell is still unclear but it might be similar to SARS CoV-1 like viruses. Lastly, we believe that COVID-19 is being transmitted from human to human, but as more data accumulate the picture will be more clear, as these virus spread beyond the imagination of scientific community.

Authors Contribution

RL, RK, VG conceived the study and designed the study. RK, HV, NS, US, VG, MS, SN, PH executed the analysis and prepared figures. RK, HV, RS, 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, New Delhi (CSIR) for doctoral fellowships. HV would like to thank Ramjas College, University of Delhi, Delhi for providing support. VG and MS acknowledge Phixgen Pvt. Ltd. and NMCG for research fellowship. PH would like to thank Maitrey College, University of Delhi, Delhi for providing support.

References:

1. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic Local Alignment Search Tool. J Mol Biol 215, 403-410.

2. Athmer, J., Fehr, A.R., Grunewald, M., Smith, E.C., Denison, M.R., and Perlman, S. (2017). In situ tagged nsp15 reveals interactions with coronavirus replication/transcription complex-associated proteins. mBio 8, e02320-16.

3. Barcena, M., Oostergetel, G.T., Bartelink, W., Faas, F.G., Verkleij, A., Rottier, P.J., Koster, A.J., and Bosch, B.J. (2009). Cryo-electron tomography of mouse hepatitis virus: Insights into the structure of the coronavirion. Proc Natl Acad Sci U S A 106, 582–587.

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

5. Brian, D.A., and Baric, R.S. (2005). Coronavirus genome structure and replication. Curr Top Microbiol Immunol 287, 1–30.

6. Chen, Y., Su, C., Ke, M., Jin, X., Xu, L., Zhang, Z., Wu, A., Sun, Y., Yang, Z., Tien, P., Ahola, T., Liang, Y., Liu, X., and Guo, D. (2011). Biochemical and structural insights into the mechanisms of SARS coronavirus RNA ribose 2'-O-methylation by nsp16/nsp10 protein complex. PLoS Pathog 7, e1002294.

7. Collins, A.R., Knobler, R.L., Powell, H., and Buchmeier, M.J. (1982). Monoclonal antibodies to murine hepatitis virus-4 (strain JHM) define the viral glycoprotein responsible for attachment and cell--cell fusion. Virology 119, 358–371.

8. Cook, H.V., Doncheva, N.T., Szklarczyk, D., von Mering, C., and Jensen, L.J. (2018). Viruses.STRING: A Virus-Host Protein-Protein Interaction Database. Viruses 10, 519.

9. Cornillez-Ty, C.T., Liao, L., Yates, J.R.3rd., Kuhn, P., and Buchmeier, M.J. (2009). Severe acute respiratory syndrome coronavirus nonstructural protein 2 interacts with a host protein complex involved in mitochondrial biogenesis and intracellular signaling. J Virol 83, 10314-10318.

10. de Haan, C.A., Smeets, M., Vernooij, F., Vennema, H., and Rottier, P.J. (1999). Mapping of the coronavirus membrane protein domains involved in interaction with the spike protein. J Virol 73, 7441–7452.

11. Edgar, E. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792-7.

12. Fehr, A.R., and Perlman, S. (2015). Coronaviruses: an overview of their replication and pathogenesis. Methods Mol Biol 1282, 1–23.

13. Fielding, B.C., Gunalan, V., Tan, T.H., Chou, C.F., Shen, S., Khan, S., Lim, S.G., Hong, W., and Tan, Y.J. (2006). Severe acute respiratory syndrome coronavirus protein 7a interacts with hSGT. Biochem Biophys Res Commun 343,:1201-8.

14. Fung, T.S., and Liu, D.X. (2019). Human Coronavirus: Host-Pathogen Interaction. Annu Rev Microbiol 73, 529–557.

15. Gorbalenya, A.E., Baker, S.C., Baric, R.S., de Groot, R.J., Drosten, C., Gulyaeva, A.A., Haagmans, B.L., Lauber, C., Leontovich, A.M., Neuman, B.W., Penzar, D., Perlman, S., et al. (2020). The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. Nat Microbiol 5, 536-544.

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

17. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports/

18. Huang, L., Shi, Y., Gong, B., Jiang, L., Liu, X., Yang, J., Tang, J., You, C., Jiang, Q., Long, B., Zeng, T., Luo, M., Zeng, F., Zeng, F., Wang, S., Yang, X., and Yang, Z. (2020). Blood single cell immune profiling reveals the interferon-MAPK pathway mediated adaptive immune response for COVID-19. BMJ, doi: https://doi.org/10.1101/2020.03.15.20033472.

19. Huang, Y., Wang, X., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X., Cheng, Z., Yu, T., Xia, J., Wei, Y., Wu, W., Xie, X., Yin, W., Li, H., Liu, M., Xiao, Y., Gao, H., Guo, L., Xie, J., Wang, G., Jiang, R., Gao, Z., Jin, Q., Wang, J., and Cao, B. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 395, 497-506.

20. Jang, K.J., Jeong, S., Kang, D.Y., Sp, N., Yang, Y.M., and Kim, D.E. (2020). A high ATP concentration enhances the cooperative translocation of the SARS coronavirus helicase nsP13 in the unwinding of duplex RNA. Sci Rep 10, 1-13.

21. Josset, L., Menachery, V.D., Gralinski, L.E., Agnihothram, S., Sova, P., Carter, V.S., Yount, B.L., Graham, R.L., Baric, R.S., and Katze, M.G. (2013). Cell host response to infection with novel human coronavirus EMC predicts potential antivirals and important differences with SARS coronavirus. mBio 4, e00165-13.

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

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

24. Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., Duesbury, M., Dumousseau, M., Feuermann, M., Hinz, U., et al. (2013). The IntAct molecular interaction database in 2012. Nucleic Acids Res 41, D43-D47.

25. Klumperman, J., Locker, J.K., Meijer, A., Horzinek, M.C., Geuze, H.J., and Rottier, P.J. (1994). Coronavirus M proteins accumulate in the Golgi complex beyond the site of virion budding. J Virol 68, 6523–6534.

26. Kosakovsky Pond, S.L. and Frost, S.D. (2005). Not So Different After All: A Comparison of methods for detecting amino acid sites under selection. Mol Biol Evol 22, 1208–1222.

27. Kryazhimskiy, S., and Plotkin, J.B. (2008). The Population Genetics of dN/dS. PLoS Genet 4, e1000304.

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

29. Lau, S.K., Woo, P.C., Li, K.S., Huang, Y., Tsoi, H.W., Wong, B.H., Wong, S.S., Leung, S.Y., Chan, K.H., and Yuen, K.Y. (2005). Severe acute respiratory syndrome coronavirus-like virus in Chinese horseshoe bats. Proc Natl Acad Sci U S A 102, 14040–14045.

30. Li, F. (2016). Structure, function, and evolution of coronavirus spike proteins. Annu Rev Virol 3, 237-261.

31. Li, S.W., Wang, C.Y., Jou, Y.J., Jou, Y.J., Tang, T.C., Huang, S.H., Wan, L., Lin, Y.J., and Lin, C.W. (2016). SARS coronavirus papain-like protease induces Egr-1-dependent up-regulation of TGF-β1 via ROS/p38 MAPK/STAT3 pathway. Sci Rep 6, 25754.

32. Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., Wang, W., Ma, X., Zhan, F., Wang, L., Hu, T., Zhou, H., Hu, Z.. et al. (2020). Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancelet 395, 565-574.

33. McBride, R., van Zyl, M., and Fielding, B.C. (2014). The coronavirus nucleocapsid is a multifunctional protein. Viruses 6, 2991–3018.

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

35. Maglott, D., Ostell, J., Pruitt, K.D., and Tatusova, T. (2005). Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res. 33, D54–D58.

36. Mergia, A. (2017). The Role of Caveolin 1 in HIV Infection and Pathogenesis. Viruses 9, 129.

37. Meyer, B., Muller, M.A., Corman, V.M., Reusken, C.B., Ritz, D., Godeke, G. J., Lattwein, E., Kallies, S., Siemens, A., van Beek, J., Drexler, J.F., Muth, D., Bosch, B.J., Wernery, U., Koopmans, M.P., Wernery, R., and Drosten, C. (2014). Antibodies against MERS coronavirus in dromedary camels, United Arab Emirates, 2003 and 2013. Emerg Infect Dis 20, 552–559.

38. Nakamura, Y., and Tomii, K. (2018). Parallelization of MAFFT for large-scale multiple sequence alignments. Bioinformatics 34, 2490–2492.

39. Neuman, B.W., Adair, B.D., Yoshioka, C., Quispe, J.D., Orca, G., Kuhn, P., Milligan, R.A., Yeager, M., and Buchmeier, M.J. (2006). Supramolecular architecture of severe acute respiratory syndrome coronavirus revealed by electron cryomicroscopy. J virol 80,7918–7928.

40. Neuman, B.W., Kiss, G., Kunding, A.H., Bhella, D., Baksh, M.F., Connelly, S., Droese, B., Klaus, J.P., Makino, S., Sawicki, S.G., Siddell, S.G., Stamou, D.G., Wilson, I.A., Kuhn, P., and Buchmeier, M.J. (2011). A structural analysis of M protein in coronavirus assembly and morphology. J Struct Biol 174, 11–22.

41. Page, A.J., Cummins, C.A., Hunt, M., Wong, V.K., Reuter, S., Holden, M.T., Fookes, M., Falush, D., Keane, J.A., and Parkhill, J. (2015). Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 31, 3691-3693.

42. Pond, SL., Frost, S.D., and Muse, S.V. (2005). HyPhy: hypothesis testing using phylogenies. Bioinformatics. 21: 676-9.

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

44. Qin, H., Wang, L., Feng, T., Elson, C.O., Niyongere, S.A., Lee, A.J., Reynolds, S.L., Weaver, C.T., Roarty, K., Serra, R., Benveniste, E.N., and Cong, Y. (2009). TGF-beta promotes Th17 cell development through inhibition of SOCS3. J Immunol 183, 97–105.

45. Ren. L.L., Wang, Y.M., Wu, Z.Q., Xiang, Z.C., Guo, L., Xu, T., Jiang, Y.Z., Xiong, Y., Li, Y.J., Li, X.W., Li, H., Fan, G.H., Gu, X.Y., Xiao, Y., Gao, H., Xu, J.Y., Yang, F., Wang, X.M., Wu, C., Chen, L., Liu, Y.W., Liu, B., Yang, J., Wang, X.R., Dong, J., Li, L., Huang, C.L., Zhao, J.P., Hu, Y., Cheng, Z.S., Liu, L.L., Qian, Z.H., Qin, C., Jin, Q., Cao, B., and Wang, J.W. (2020). Identification of a novel coronavirus causing severe pneumonia in human: a descriptive study. Chin Med J (Engl). doi: 10.1097/CM9.0000000000000722.

46. Ruch, T.R., and Machamer, C.E. (2012). The coronavirus E protein: assembly and beyond. Viruses 4, 363-382.

47. Sah, R., Alfonso, J., Rodriguez-Morales., Jha, R., Daniel, K.W., Chu, H.G., Peiris, M., Bastola, A., Lal, B.K., Ojha, H.C., Rabaan, A.A., Zambrano, L.I., Costello, A., Morita, K., Pandey, B.D., and Poon, L.L.M. (2020). Complete genome sequence of a 2019 novel coronavirus (SARS-CoV-2) strain isolated in Nepal. Microbiol Res Announce 9, e00169-20.

48. Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics 30, 2068-2069.

49. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13, 2498-2504.

50. Snijder, E.J., Decroly, E., and Ziebuhr, J. (2016). The non-structural proteins directing coronavirus RNA synthesis and processing. Adv Virus Res 96, 59-126.

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

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

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

54. Tang, Q., Song, Y., Shi, M., Cheng, Y., Zhang, W., Xia, X.Q. (2015). Inferring the hosts of coronavirus using dual statistical models based on nucleotide composition. Sci Rep 5, 17155.

55. Tan,Y.X.,Timothy, H.P. Tan, Marvin, J.-R., Lee, Tham, P.Y., Gunalan, V., Druce, J., Birch, C., Catton, M., Fu, N.Y., Yu, V.C., and Tan, Y.J. (2007). 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 81, 6346-6355.

56. Tatsuta, T., Model, K., and Langer, T. (2005). Formation of membrane-bound ring complexes by prohibitins in mitochondria. Mol Biol Cell 16, 248-259.

57. Treangen, T.J., Ondov, B.D., Koren, S. and Phillippy, A.M. (2014). The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. Genome Biol 15, 524.

58. Tyrrell, D.A., and Bynoe, M.L. (1966). Cultivation of viruses from a high proportion of patients with colds. Lancet 1, 76–77.

59. UniProt Consortium. (2007). The universal protein resource (UniProt). Nucleic Acids Res **36,** D190-195.

60. von Grotthuss, M., Wyrwicz, L.S., and Rychlewski, L. (2003). mRNA cap-1 methyltransferase in the SARS genome. Cell 113, 701-702.

61. Wall, A.C., Park, Y.J., Tortorici, M.A., Wall, A., McGuire, A.T., and Vessler, D. (2020). Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. Cell 180, 1-12.

62. Wang, Q., Zhang, Y., Wu, L., Niu, S., Song, C.,Zhang, Z., Lu, G., Qiao,C., Hu, Y., Yuen, K.Y., Wang, Q., Zhou, H., Yan, J., and Qi, J. (2020). Structural and functional basis of SARS-CoV-2 entry by using human ACE2. Cell doi: 10.1016/j.cell.2020.03.045.

63. Weaver, S., Shank, S.D., Spielman, S.J., Li, M., Muse, S.V., Kosakovsky, Pond. S.L. (2018). Datamonkey 2.0: A Modern Web Application for Characterizing Selective and Other Evolutionary Processes. Mol Biol Evol. 35: 773-777.

64. Woo, P.C., Lau, S.K., Lam, C.S., Lau, C.C., Tsang, A.K., Lau, J.H., Bai, R., Teng, J.L., Tsang, C.C., Wang, M., Zheng, B.J., Chan, K.H., and Yuen, K.Y. (2012). 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 86, 3995-4008.

65. Wu, A., Peng, Y., Huang, B., Huang, B., Ding, X., Wang, X., Niu, P., Meng, J., Zhu, Z.Z., Wang, J., Sheng, J., Quan, L., Xia, Z., Tan, W., Cheng, G., and Jiang, T. (2020). Genome composition and divergence of the novel coronavirus (2019-nCoV) originating in China. Cell Host Microbe 27, 325-328.

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

67. Xu, Z., Shi, L., Wang, Y., Zhang, J., Huang, L., Zhang, C., Liu, S., Zhao, P., Liu, H., Zhu, L., Tai, Y., Bai, C., Gao, T., Song, J., Xia, P., Dong, J., Zhao, J., and Wang, F.S. (2020). Pathological findings of COVID-19 associated with acute respiratory distress syndrome. Lancet Respir Med doi: 10.1016/S2213-2600(20)30076-X.

68. Yuan, X., Li, J., Shan, Y., Yang, Z., Zhao, Z., Chen, B., Yao, Z., Dong, B., Wang, S., Chen, J., and Chong, Y. (2005). subcellular localization and membrane association of SARS-CoV 3a protein. Virus Res 109, 191-202.

69. Zhang, S. (2018). The role of transforming growth factor *β* in T helper 17 differentiation. Immunology. 155: 24–35.

70. Zhang. C., Zheng, W., Huang, X., Bell, E.W., Zhou, X., and Zhang, Y. (2020). 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. arXiv:2002.03173[q-bio.GN].

71. Zhao, X., Nicholls, J.M., and Chen, Y.G. (2008). Sars-cov nucleocapsid protein interacts with smad3 and modulates TGF-β signaling. J Biol Chem 8, 3272-3280.

72. Zhou, Z., Alikhan, N.F., Sergeant, M.J., Luhmann, N., Vaz, C., Francisco, A.P., Carrico, J.A., and Achtman, M. (2018). GrapeTree: visualization of core genomic relationships among 100,000 bacterial pathogens. Genome Res 28, 1395-1404.

Figure legends

Figure 1: The distribution pattern of COVID-19 cases across the globe and the number of isolates (in circle) included in the present study.

Figure 2: 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 3: 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 the frameshift mutation between positions 2540 and 2570.

Figure 4: (A) SARS-CoV-2 -Host interactome analysis. Sub-setnetwork 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

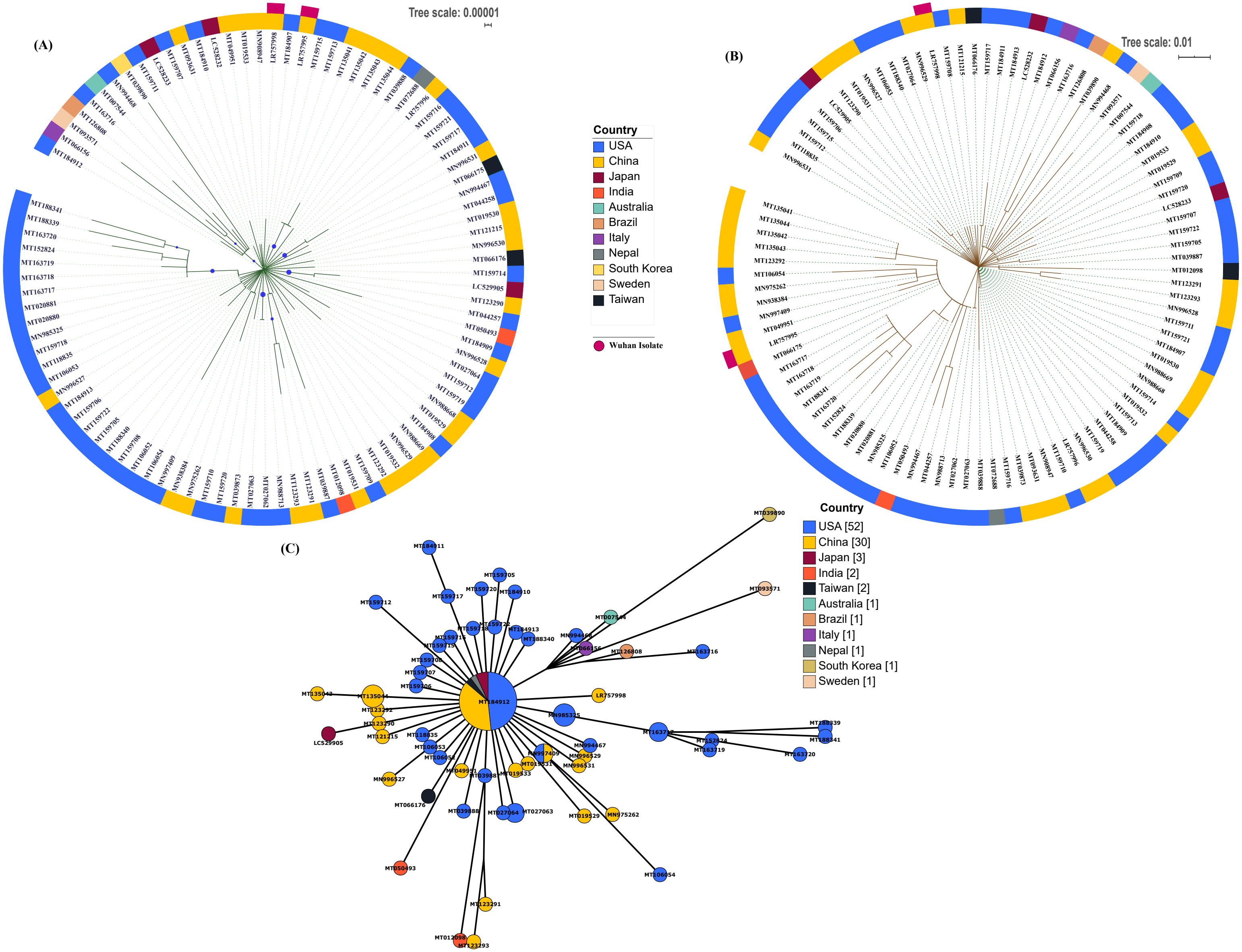
Kb compared with that of SARS-CoV generated using CGView. The ruler for genome size is shown as innermost ring where Kbp stands for kilo base pairs. Concentric circles from inside to outside denote: SARS-CoV genome (used as reference), G + C content, G + C skew, predicted ORFs in SARS-CoV-2 genome and 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 STRING Virus 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.

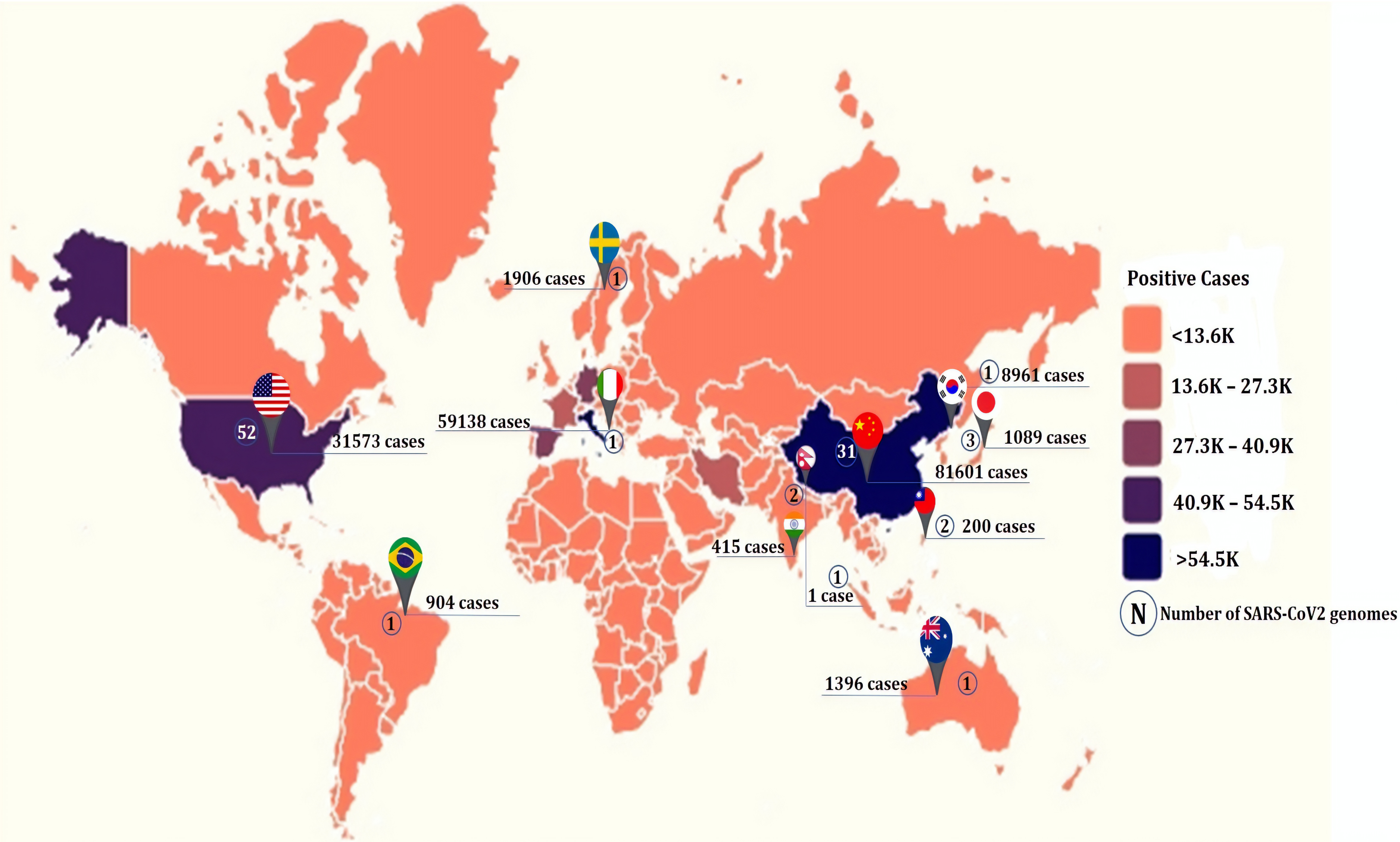
Tables Legends

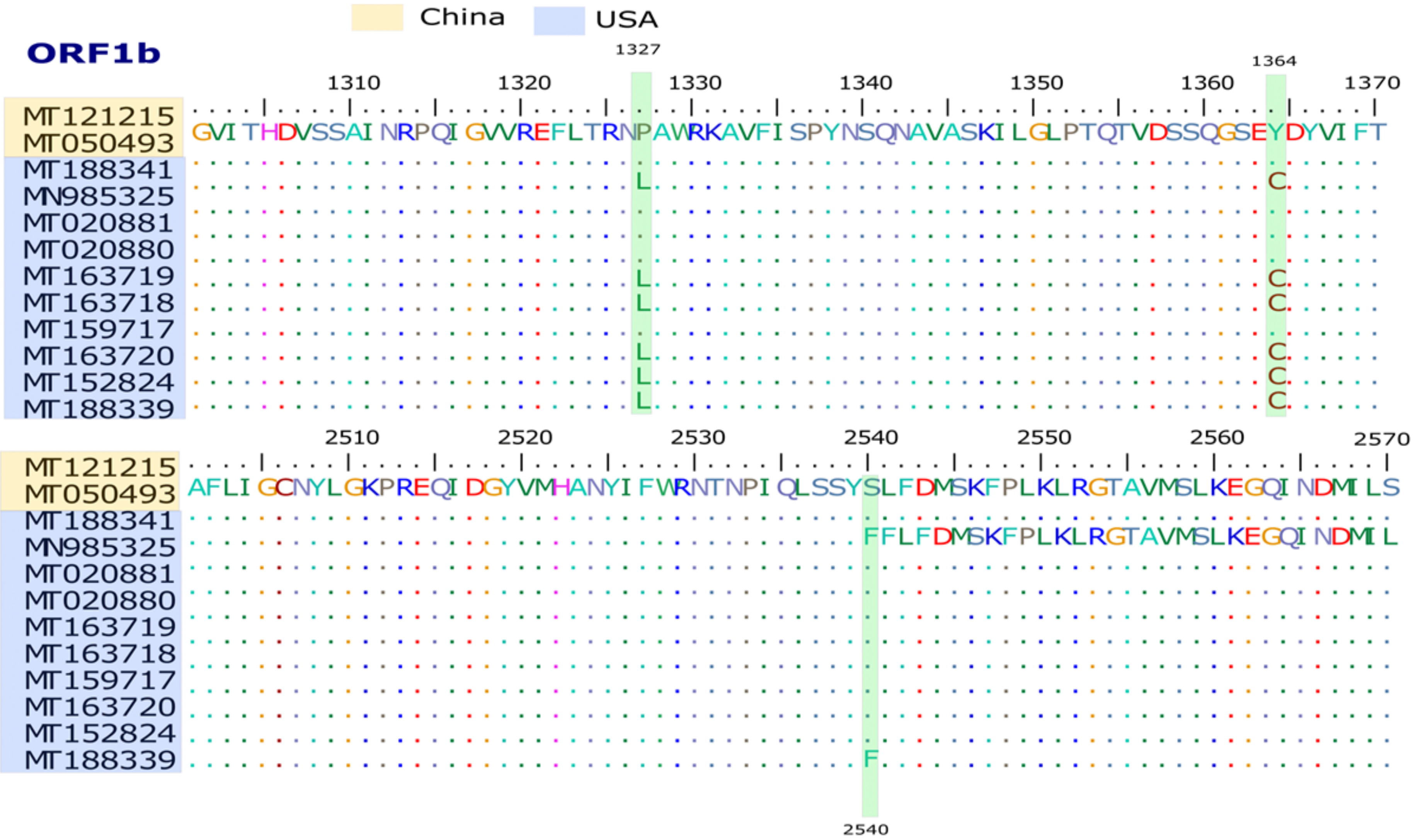
Table 1: General genomic attributes of SARS-CoV-2 strains.

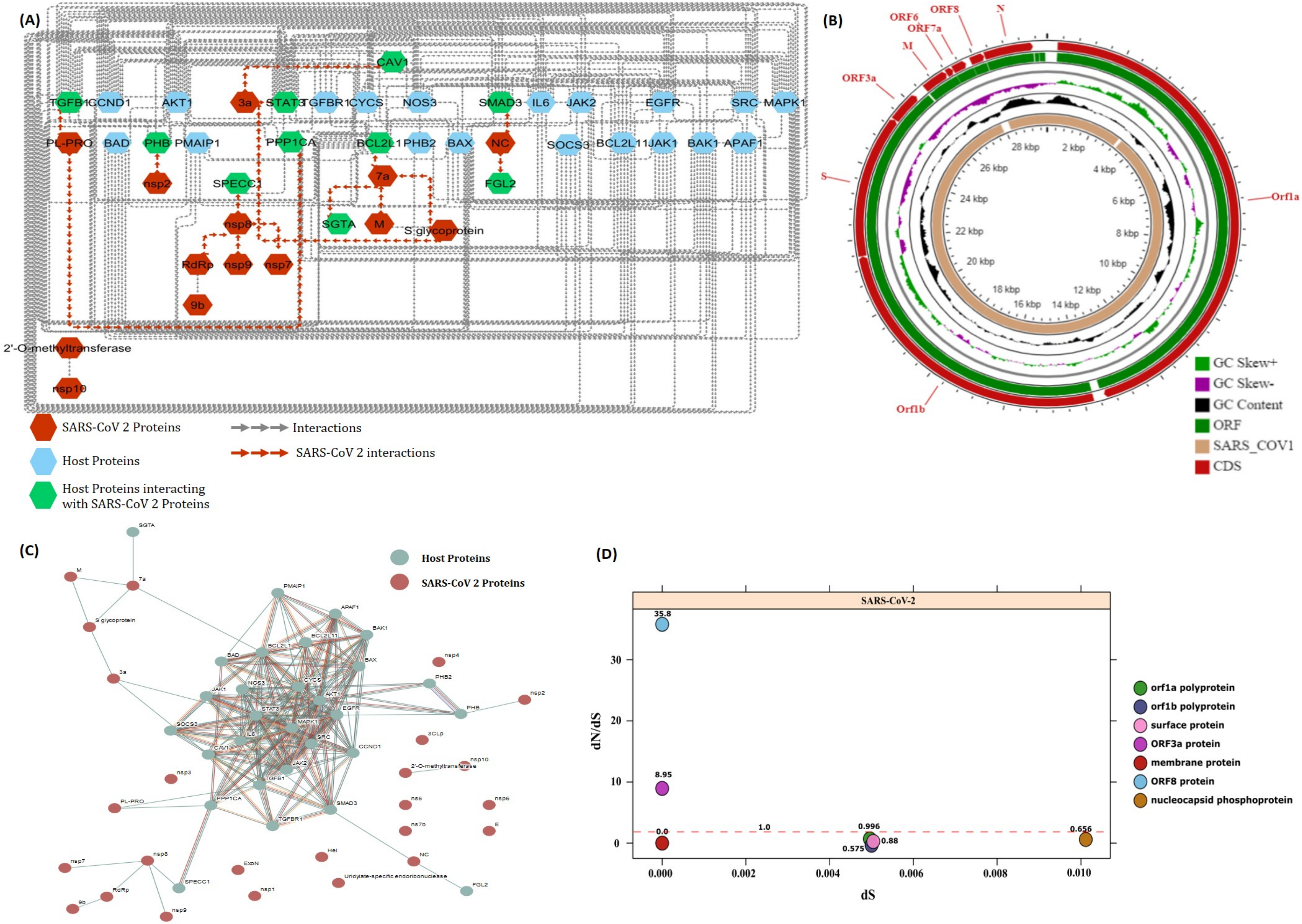
Table 2: Major mutations present in different isolates of SARS-CoV-2 at different locations.

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


(A)
Tree scale: 0.00001
(B)
Tree scale: 0.01
(C)
Country
🟦 USA
🟨 China
Japan
🟥 India
Australia
Brazil
🟪 Italy
Nepal
South Korea
Sweden
⬛ Taiwan
● Wuhan Isolate
Country
🟦 USA [52]
🟨 China [30]
Japan [3]
🟥 India [2]
⬛ Taiwan [2]
Australia [1]
Brazil [1]
🟪 Italy [1]
Nepal [1]
South Korea [1]
Sweden [1]
MT184912
MT039890
MT093571
MT007544
MT994468
MT066156
MT126808
MT163716
LR757998
MN985325
MT163717
MT152824
MT163719
MT188339
MT188341
MT163720
MT106054
MT184911
MT159705
MT159720
MT184910
MT159712
MT159717
MT159722
MT159718
MT184913
MT188340
MT159716
MT159715
MT159708
MT159707
MT159706
MT135042
MT135044
MT123292
MT123290
MT121215
LC529905
MT118835
MT106053
MT106052
MN996527
MT049951
MT066176
MT039888
MT039887
MT027064
MT027063
MT019533
MT019531
MN997409
MN994467
MN996529
MN996531
MT019529
MN975262
MT050493
MT123291
MT012098
MT123293


1906 cases
1
52
31573 cases
59138 cases
1
904 cases
1
1
8961 cases
3
1089 cases
31
81601 cases
2
2
200 cases
415 cases
1
1 case
1396 cases
1
Positive Cases
<13.6K
13.6K – 27.3K
27.3K – 40.9K
40.9K – 54.5K
>54.5K
N Number of SARS-CoV2 genomes


China
USA
ORF1b
1327
1364
1310
1320
1330
1340
1350
1360
1370
MT121215
MT050493
GVITHDVSSAINRPQIGVVREFLTRNPAWRKAVFISPYNSQNAVASKILGLPTQTVDSSQGSEYDYVIFT
MT188341
MN985325
MT020881
MT020880
MT163719
MT163718
MT159717
MT163720
MT152824
MT188339
L
C
2510
2520
2530
2540
2550
2560
2570
MT121215
MT050493
AFLIGCNYLGKPREQIDGYVMHANYIFWRNTNPIQLSSYSLFDMSKFPLKLRGTAVMSLKEGQINDMILS
MT188341
MN985325
FFLFDMSKFPLKLRGTAVMSLKEGQINDMIL
MT020881
MT020880
MT163719
MT163718
MT159717
MT163720
MT152824
MT188339
F
2540


(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
nsp8
SGTA
M
S glycoprotein
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
2 kbp
4 kbp
6 kbp
8 kbp
10 kbp
12 kbp
14 kbp
16 kbp
18 kbp
20 kbp
22 kbp
24 kbp
26 kbp
28 kbp
GC Skew+
GC Skew-
GC Content
ORF
SARS_COV1
CDS
(C)
Host Proteins
SARS-CoV 2 Proteins
SGTA
M
7a
S glycoprotein
3a
PMAIP1
APAF1
BAK1
BCL2L11
BCL2L1
BAD
BAX
nsp4
PHB2
PHB
nsp2
JAK1
NOS3
CYCS
AKT1
STAT3
EGFR
SOCS3
MAPK1
IL6
SRC
3CLp
CAV1
CCND1
JAK2
nsp10
2'-O-methyltransferase
nsp3
TGFB1
PL-PRO
PPP1CA
SMAD3
ns6
nsp6
TGFBR1
ns7b
E
nsp7
nsp8
Hel
NC
ExoN
SPECC1
RdRp
9b
nsp9
nsp1
Uridylate-specific endoribonuclease
FGL2
(D)
SARS-CoV-2
dN/dS
dS
0
10
20
30
0.000
0.002
0.004
0.006
0.008
0.010
35.8
8.95
0.0
1.0
0.996
0.575
0.88
0.656
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	LR757996.1	Wuhan seafood market pneumonia virus	China: Wuhan	29857	38	NA	01/01/2020
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_HK	China	29891	37.98	Oronasopharynx	11/01/2020

		U-SZ-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-	Kerala, India	29851	38.01	Oronasopharynx	31/01/2020

		2/166/human/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-	USA	29882	38	Oronasopharynx	17/02/2020

		CruiseA-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-	USA	29882	38	Oronasopharynx	24/02/2020

		CruiseA-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-	USA	29880	38	Oronasopharynx	17/02/2020

		CruiseA-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			