

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

3 Roshan Kumar¹, Helianthous Verma², Nirjara Singhvi³, Utkarsh Sood⁴, Vipin Gupta⁵, Mona
4 Singh⁵, Rashmi Kumari⁶, Princy Hira⁷, Shekhar Nagar³, Chandni Talwar³, Namita Nayyar⁸,
5 Shailly Anand⁹, Charu Dogra Rawat², Mansi Verma⁸, Ram Krishan Negi³, Yogendra Singh³
6 and Rup Lal^{4*}

7

8

9

10 **Authors Affiliations**

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

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

13 ³Department of Zoology, University of Delhi, New Delhi-110007, India

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

16 ⁵PhiXGen Private Limited, Gurugram, Haryana 122001, India

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

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

20 ⁸Department of Zoology, Sri Venkateswara College, University of Delhi, New Delhi-110021,
21 India

22 ⁹Department of Zoology, Deen Dayal Upadhyaya College, University of Delhi, New Delhi-
23 110078, India

24

25

26

27

28 ***Corresponding Author**

29 Email: ruplal@gmail.com

30

31 **Abstract**

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

52

53

54

55

56

57

58

59

60

61 **Background**

62 Since the current outbreak of pandemic coronavirus disease 2019 (COVID-19) caused by
63 Severe Acute Respiratory Syndrome-related Coronavirus-2 (SARS-CoV-2), the assessment of
64 the biogeographical pattern of SARS-CoV-2 isolates and the mutations at nucleotide and
65 protein level is of high interest to many research groups [1, 2, 3]. Coronaviruses (CoVs),
66 members of *Coronaviridae* family, order *Nidovirales*, have been known as human pathogens
67 from the last six decades [4]. Their target is not just limited to humans, but also other mammals
68 and birds [5]. Coronaviruses have been classified under alpha, beta, gamma and delta-
69 coronavirus groups [6] in which former two are known to infect mammals while the latter two
70 primarily infect bird species [7]. Symptoms in humans vary from common cold to respiratory
71 and gastrointestinal distress of varying intensities. In the past, more severe forms caused major
72 outbreaks that include Severe Acute Respiratory Syndrome (SARS-CoV) (outbreak in 2003,
73 China) and Middle East Respiratory Syndrome (MERS-CoV) (outbreak in 2012, Middle East)
74 [8]. Bats are known to host coronaviruses acting as their natural reservoirs which may be
75 transmitted to humans through an intermediate host. SARS-CoV and MERS-CoV were
76 transmitted from intermediate hosts, palm civets and camel, respectively [9, 10]. It is not,
77 however, yet clear which animal served as the intermediate host for transmission of SARS-
78 CoV-2 transmission from bats to humans which is most likely suggested to be a warm-blooded
79 vertebrate [11].

80 The inherently high recombination frequency and mutation rates of coronavirus genomes allow
81 for their easy transmission among different hosts. Structurally, they are positive-sense single
82 stranded RNA (ssRNA) virions with characteristic spikes projecting from the surface of capsid
83 coating [12, 13]. The spherical capsid and spikes give them crown-like appearance due to
84 which they were named as ‘corona’, meaning ‘crown’ or ‘halo’ in *Latin*. Their genome is
85 nearly 30 Kb long, largest among the RNA viruses, with 5’cap and 3’ polyA tail, for translation
86 [14]. Coronavirus consists of four main proteins, spike (S), membrane (M), envelope (E) and
87 nucleocapsid (N). The spike (~150 kDa) mediates its attachment to host receptor proteins [15].
88 Membrane protein (~25-30 kDa) attaches with nucleocapsid and maintains curvature of virus
89 membrane [16]. E protein (8-12 kDa) is responsible for the pathogenesis of the virus as it eases
90 assembly and release of virion particles and also has ion channel activity as integral membrane
91 protein [17]. N-protein, the fourth protein, helps in the packaging of virus particles into capsids
92 and promotes replicase-transcriptase complex (RTC) [18].

93 Recently, in December 2019, the outbreak of novel beta-coronavirus (2019-nCoV) or SARS-
94 CoV-2 in Wuhan, China has shown devastating effects worldwide
95 ([https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200403-sitrep-74-](https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200403-sitrep-74-covid-19-mp.pdf?sfvrsn=4e043d03_4)
96 [covid-19-mp.pdf?sfvrsn=4e043d03_4](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
97 COVID-19, the disease caused by the novel SARS-CoV-2 a pandemic, affecting more than
98 186 countries and territories where USA has most reported cases 2,13,600 and Italy has highest
99 mortality rate 12.08% (1,15,242 infected individuals, 13,917 deaths) (WHO situation report-
100 74). As on date (April 4, 2020), more than 1 million individuals have been infected by SARS-
101 CoV-2 and nearly 60,000 have died worldwide. Virtually, all human lives have been impacted
102 with no foreseeable end of the pandemic. A recent study on ten novel coronavirus strains by
103 Lu *et al.*, suggested that SARS-CoV-2 has sufficiently diverged from SARS-CoV [19]. SARS-
104 CoV-2 is assumed to have originated from bats, which serve as a reservoir host of the virus
105 [19]. A recent study has shown similar mutation patterns in Bat-SARS-CoV RaTG13 and
106 SARS CoV-2, but the dataset was limited to 21 strains including few SARS-CoV-2 strains and
107 other neighbors [20]. Other studies have also reported the genome composition and divergence
108 patterns of SARS-CoV-2 [3, 21]. However, no study has yet explained the biogeographical
109 pattern of this emerging pathogen. In this study, we selected 95 strains of SARS-CoV-2,
110 isolated and sequenced from 11 different countries to understand the transmission patterns,
111 evolution and pathogenesis of the virus. Using core genome and Single Nucleotide
112 Polymorphism (SNP) based phylogeny, we attempted to uncover any existence of a
113 transmission pattern of the virus across the affected countries, which was not known earlier.
114 We analyzed the ORFs of the isolates to reveal unique point mutations and amino-acid
115 substitutions/additions in the isolates from the USA. In addition, we analyzed the gene/protein
116 mutations in these novel strains and estimated the direction of selection to decipher their
117 evolutionary divergence rate. Further, we also established the interactome of SARS-CoV-2
118 with the human host proteins to predict the functional implications of the viral infection host
119 cells. The results obtained from the analyses indicate the high severity of SARS-CoV-2 isolates
120 with the inherent capability of unique mutations and the evolving viral replication strategies to
121 adapt to human hosts.

122 **Materials and Methods**

123 **Selection of genomes and annotation**

124 Sequences of different strains were downloaded from NCBI database
125 <https://www.ncbi.nlm.nih.gov/genbank/sars-cov-2-seqs/> (Table 1). A total of 97 genomes were
126 downloaded on March 19, 2020 from NCBI database and based on quality assessment two
127 genomes with multiple Ns were removed from the study. Further the genomes were annotated
128 using Prokka [22]. A manually annotated reference database was generated using the Genbank
129 file of Severe acute respiratory syndrome coronavirus 2 isolate- SARS-CoV-
130 2/SH01/human/2020/CHN (Accession number: MT121215) and open reading frames (ORFs)
131 were predicted against the formatted database using prokka (-gcode 1) [22]. Further the GC
132 content information was generated using QUAST standalone tool [23].

133 **Analysis of natural selection**

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

140 **Phylogenetic analysis**

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

145 To determine the single nucleotide polymorphism (SNP), whole-genome alignments were
146 made using libMUSCLE aligner. For this, we used annotated genbank of SARS-CoV-
147 2/SH01/human/2020/CHN (Accession no. MT121215) as the reference in the parsnp tool of
148 Harvest suite [33]. As only genomes within a specified MUMI distance threshold are recruited,
149 we used option -c to force include all the strains. For output, it produced a core-genome
150 alignment, variant calls and a phylogeny based on Single nucleotide polymorphisms. The SNPs
151 were further visualized in Gingr, a dynamic visual platform [33]. Further, the tree was
152 visualized in interactive Tree of Life (iTOL) [31].

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

154 SARS-CoV-2/SH01/human/2020/CHN virus genome having accession no. MT121215.1 was
155 used for protein-protein network analysis. Since, none of the SARS-CoV-2 genomes are
156 updated in any protein database, we first annotated the genes using BLASTp tool [34]. The
157 similarity searches were performed against SARS-CoV isolate Tor2 having accession no.
158 AY274119 selected from NCBI at default parameters. The annotated SARS-CoV-2 proteins
159 were mapped against viruSITE [35] and interaction databases such as Virus.STRING v10.5
160 [36] and IntAct [37] for predicting their interaction against host proteins. These proteins were
161 either the direct targets of HCoV proteins or were involved in critical pathways of HCoV
162 infection identified by multiple experimental sources. To build a comprehensive list of human
163 PPIs, we assembled data from a total of 18 bioinformatics and systems biology databases with
164 five types of experimental evidence: (i) binary PPIs tested by high-throughput yeast two-hybrid
165 (Y2H) systems; (ii) binary, physical PPIs from protein 3D structures; (iii) kinase-substrate
166 interactions by literature-derived low-throughput or high-throughput experiments; (iv)
167 signaling network by literature-derived low-throughput experiments; and (v) literature-curated
168 PPIs identified by affinity purification followed by mass spectrometry (AP-MS), Y2H, or by
169 literature-derived low [36, 38].

170 Filtered proteins (confidence value: 0.7) were mapped to their Entrez ID [39] based on the
171 NCBI database used for interactome analysis. HPI were stimulated using Cytoscape v.3.7.2
172 [40].

173 **Functional enrichment analysis**

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

180 **Results and Discussion**

181 **General genomic attributes of SARS-CoV-2**

182 In this study, we analyzed a total of 95 SARS-CoV-2 strains (available on March 19, 2020)
183 isolated between December 2019-March 2020 from 11 different countries namely USA (n=52),
184 China (n=30), Japan (n=3), India (n=2), Taiwan (n=2) and one each from Australia, Brazil,

185 Italy, Nepal, South Korea and Sweden. A total of 68 strains were isolated from either
186 oronasopharynges or lungs, while two of them were isolated from faeces suggesting both
187 respiratory and gastrointestinal connection of SARS-CoV-2 (Table 1). No information of the
188 source of isolation of the remaining isolates is available. The average genome size and GC
189 content were found to be 29879 ± 26.6 bp and $37.99 \pm 0.018\%$, respectively. All these isolates
190 were found to harbor 9 open reading frames coding for ORF1a (13218 bp) and ORF1b (7788
191 bp) polyproteins, surface glycoprotein or S-protein (3822 bp), ORF3a protein (828 bp),
192 membrane glycoprotein or M-protein (669 bp), ORF6 protein (186 bp), ORF7a protein (366
193 bp), ORF8 protein (366 bp), and nucleocapsid phosphoprotein or N-protein (1260 bp) which
194 agrees with a recently published study [44]. The ORF1a harbors 12 non-structural protein (nsp)
195 namely nsp1, nsp2, nsp3 (papain-like protease or PLpro domain), nsp4, nsp5 (3C-like protease
196 or 3CLpro), nsp6, nsp7, nsp8, nsp9, nsp10, nsp11 and nsp12 (RNA-dependent RNA
197 polymerase or RdRp) [44]. Similarly, ORF1b contains four putative nsp's namely nsp13
198 (helicase or Hel), nsp14 (3'-to-5' exoribonuclease or ExoN), nsp15 and nsp16 (mRNA cap-1
199 methyltransferase).

200 **Phylogenomic analysis: defining evolutionary relatedness**

201 Our analysis revealed that strains of human infecting SARS-CoV-2 are novel and highly
202 identical (>99.9%). A recent study established the closest neighbor of SARS-CoV-2 as SARSr-
203 CoV-RaTG13, a bat coronavirus [45]. As COVID19 transits from epidemic to pandemic due
204 to extremely contagious nature of the SARS-CoV-2, it was interesting to draw the relation
205 between strains and their geographical locations. In this study, we employed two methods to
206 delineate phylogenomic relatedness of the isolates: core genome (Figure 1A & C) and single
207 nucleotide polymorphisms (SNPs) (Figure 1B). Phylogenies obtained were annotated with
208 country of isolation of each strain (Figure 1A & B). The phylogenetic clustering was found
209 majorly concordant by both core-genome (Figure 1A) and SNP based methods (Figure 1B).
210 The strains formed a monophyletic clade, in which MT093571.1 (South Korea) and
211 MT039890.1 (Sweden) were most diverged. Focusing on the edge-connection between the
212 neighboring countries from where the transmission is more likely to occur, we noted a strain
213 from Taiwan (MT066176) closely clustered with another from China (MT121215.1). With the
214 exception of these two strains, we did not find any connection between strains of neighboring
215 countries. Thus, most strains belonging to the same country clustered distantly from each other
216 and showed relatedness with strains isolated from distant geographical locations (Figure 1A &
217 B). For instance, a SARS-CoV-2 strain isolated from Nepal (MT072688) clustered with a strain

218 from USA (MT039888). Also, strains from Wuhan (LR757998 and LR757995), where the
219 virus was originated, showed highest identity with USA as well as China strains; strains from
220 India, MT012098 and MT050493 clustered closely with China and USA strains, respectively
221 (Figure 1A & B). Similarly, Australian strain (MT007544) showed close clustering with USA
222 strain (Figure 1A & B) and one strain from Taiwan (MT066175) clustered nearly with Chinese
223 isolates (Figure 1B). Isolates from Italy (MT012098) and Brazil (MT126808) clustered with
224 different USA strains (Figure 1A & B). Notably, isolates from same country or geographical
225 location formed a mosaic pattern of phylogenetic placements of countries' isolates. For viral
226 transmission, contact between the individuals is also an important factor, supposedly due to
227 which the spread of identical strains across the border of neighboring countries is more likely.
228 But we obtained a pattern where Indian strains showed highest similarity with USA and China
229 strains, Australian strains with USA strains, Italy and Brazilian strains with strains isolated
230 from USA among others. This depicts the viral spread across different communities. However,
231 as genomes of SARS-CoV-2 were available mostly from USA and China, sampling biases is
232 evident in analyzed dataset as available on NCBI. Thus, it is plausible for strains from other
233 countries to show most similarity with strains from these two countries. In the near future as
234 more and more genome sequences will become available from different geographical locations;
235 more accurate patterns of their relatedness across the globe will become available

236 **SNPs in the SARS-CoV-2 genomes**

237 SNPs in all predicted ORFs in each genome were analyzed using SARS-CoV-
238 2/SH01/human/2020/CHN as a reference. SNPs were determined using maximum unique
239 matches between the genomes of coronavirus, we observed that the strains isolated from USA
240 (MT188341; MN985325; MT020881; MT020880; MT163719; MT163718; MT163717;
241 MT152824; MT163720; MT188339) are the most evolved and they carry set of unique point
242 mutations (Table2) in nsp13, nsp14, nsp15, nsp16 (present in orf1b polyprotein region) and S-
243 Protein. All the mutated proteins are non-structural proteins (NSP) functionally involved in
244 forming viral replication-transcription complexes (RTC) [46]. For instance, non-structural
245 protein 13 (nsp13), belongs to helicase superfamily 1 and is putatively involved in viral RNA
246 replication through RNA-DNA duplex unwinding [47] whereas nsp14 and nsp15 are
247 exoribonuclease and endoribonuclease, respectively [48, 49]. nsp16 functions as a mRNA cap-
248 1 methyltransferase [50]. All these proteins containing SNPs at several positions (Table 2)
249 indicate that viral machinery for its RNA replication and processing is utmost evolved in strains
250 from USA as compared to the other countries. Further, we analyzed the SNPs at protein level

251 and interestingly in ORF1b protein, there were amino acid substitutions at P1327L, Y1364C
252 and S2540F in USA isolates. One isolate namely USA0/MN1-MDH1/2020 (MT188341)
253 carried amino-acid addition at 2540 position leading to shift in amino acid frame their onwards
254 (Figure 2), which might affect the functioning of nsp16 (2'-O-MTase). But no changes were
255 observed in Indian isolates, thus found similar to Chinese isolate. As the proteins involved in
256 viral replication are evolving rapidly, this highlights the need to consider these mutants in order
257 to develop the treatment strategies.

258 **Direction of selection of SARS-CoV-2 genes**

259 Our analysis revealed that ORF8 (121 a.a.) (dN/dS= 35.8) along with ORF3a (275 bp) (dN/dS=
260 8.95) showed highest dN/dS values among the nine ORFs thus, have much greater number of
261 non-synonymous substitutions than the synonymous substitution (Figure 3D). Values of dN/dS
262 $\gg 1$ are indicative of strong divergent lineage [51]. Thus, both of these proteins are evolving
263 under high selection pressure and are highly divergent ORFs across strains. Two other proteins,
264 ORF1ab polyprotein (dN/dS= 0.996, 0.575) and S protein (dN/dS= 0.88) might confer selective
265 advantage with host challenges and survival. The dN/dS rates nearly 1 and greater than 1
266 suggests that the strains are coping up with the challenges *i.e.*, immune responses and inhibitory
267 environment of host cells [52]. The other gene clusters namely M-protein and orf1a polyprotein
268 did not possess at least three unique sequences necessary for the analysis, hence, they should
269 be similar across the strains. The two genes ORF1ab polyprotein encodes for protein translation
270 and post translation modification found to be evolved which actively translates, enhance the
271 multiplication and facilitates growth of virus inside the host. Similarly, the S protein which
272 helps in the entry of virus to the host cells by surpassing the cell membrane found to be
273 accelerated towards positive selection confirming the successful ability of enzyme to initiate
274 the infection. Another positive diversifying gene N protein encodes for nucleocapsid formation
275 which protects the genetic material of virus form host immune responses such as cellular
276 proteases. Overall, the data represent that the growth and multiplication related genes are
277 highly evolving. The other proteins with dN/dS values equal to zero suggesting a conserved
278 repertoire of genes.

279

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

281 Although the primary mode of infection is human to human transmission through close contact,
282 which occurs via spraying of nasal droplets from the infected person, yet the primary site of

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

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

309 Similarly, the viral protein Papain-like proteinase (PL-PRO) which has deubiquitinase and
310 deISGylating activity is responsible for cleaving viral polyprotein into 3 mature proteins which
311 are essential for viral replication [57]. Our study showed that PL-PRO directly interacts with
312 PPP1CA which is a protein phosphatase that associates with over 200 regulatory host proteins
313 to form highly specific holoenzymes. PL-PRO is also found to interact with TGF β which is a
314 beta transforming growth factor and promotes T- helper 17 cells (Th17) and regulatory T-cells
315 (T_{reg}) differentiation [58]. Reports have shown the PL-PRO induced upregulation of TGF β in

316 human promonocytes via MAPK pathway result in pro-fibrotic responses [59]. This reflects
317 that viral PL-PRO antagonises innate immune system and is directly involved in the
318 pathogenicity of SARS-CoV-2 induced pulmonary fibrosis [56, 58]. Many COVID-19 patients
319 develop acute respiratory distress syndrome (ARDS) which leads to pulmonary edema and
320 lung failure [60, 61]. These symptoms are because of cytokine storm manifesting elevated
321 levels of pro-inflammatory cytokines like IL6, IFN γ , IL17, IL1 β etc [61]. These results are in
322 agreement with our prediction where we found IL6 as an interacting partner. Our study also
323 showed JAK1/2 as an interacting partner which is known for IFN γ signaling. It is well known
324 that TGF β along with IL6 and STAT3 promotes Th17 differentiation by inhibiting SOCS3
325 [62]. Th17 is a source of IL17, which is commonly found in serum samples of COVID-19
326 patients [61, 63]. Hence, our interactome is supported from these findings where we found
327 SOCS3, STAT3, JAK1/2 as an interacting partner [64]. The results suggested that
328 proinflammatory cytokine storm is one of the reasons for SARS-CoV-2 mediated
329 immunopathogenesis.

330 In the next cycle of physical events the viral protein NC (nucleoprotein), which is a major
331 structural part of SARV family associates with the genomic RNA to form a flexible, helical
332 nucleocapsid. Interaction of this protein with SMAD3 leads to inhibition of apoptosis of SARS-
333 CoV infected lung cells [65], which is a successful strategy of immune evasion by the virus.
334 More complex and multiple associations of ORF7a viral protein which is a non-structural
335 protein and known as growth factor for SARS family viruses, directly captures BCL2L1 which
336 is a potent regulator of apoptosis. Tan *et al.* (2007) have shown that SARS-CoV ORF7a protein
337 induces apoptosis by interacting with Bcl XL protein which is responsible for lymphopenia, an
338 abnormality found in SARS-CoV infected patients [66]. Another target of viral ORF7a protein
339 is SGTA (Small glutamine-rich tetratricopeptide repeat) which is an ATPase regulator and
340 promotes viral encapsulation [67]. Subordinate viral proteins M (Membrane), S (Glycoprotein)
341 and ORF3a (viroporin) were found to interact with each other. This interaction is important for
342 viral cell formation and budding [68, 69]. Studies have shown the localization of ORF3a
343 protein in Golgi apparatus of SARS-CoV infected patients along with M protein and
344 responsible for viral budding and cell injury [70]. ORF3a protein also targets the functioning
345 of CAV1 (Caveolin 1), caveolae protein, acts as a scaffolding protein within caveolar
346 membranes. CAV1 has been reported to be involved in viral replication, persistence, and the
347 potential role in pathogenesis in HIV infection also [71]. Thus, ORF3a interactions will
348 upregulate viral replication thus playing a very crucial role in pathogenesis. Multiple

349 methyltransferase assembly viral proteins (nsp7, nsp8, nsp9, RdRp) which are nuclear
350 structural proteins were observed to target the SPECC1 proteins and linked with cytokinesis
351 and spindle formations during division. Thus, major viral assembly also targets the proteins
352 linked with immunity and cell division. Taken together, we estimated that SARS-CoV-2
353 manipulate multiple host proteins for its survival while, their interaction is also a reason for
354 immunopathogenesis.

355 **Conclusions**

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

381 viruses. Lastly, we believe as more data accumulate for COVID-19 the evolutionary pattern
382 will become much clear.

383 **Authors Contribution**

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

388

389 **Conflict of Interest**

390 Authors declare no conflict of Interest

391

392 **Acknowledgements**

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

400

401 **References:**

- 402 1. Wang Q, Zhang Y, Wu L, Niu S, Song C, Zhang Z, *et al.* Structural and functional basis of
403 SARS-CoV-2 entry by using human ACE2. *Cell.* 2020; doi: 10.1016/j.cell.2020.03.045.
- 404 2. Wall AC, Park YJ, Tortorici MA, Wall A, McGuire AT, Vessler D. Structure, function,
405 and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell.* 2020; 180: 1-12.
- 406 3. Wu A, Peng Y, Huang B, Huang B, Ding X, Wang X, *et al.* Genome composition and
407 divergence of the novel coronavirus (2019-nCoV) originating in China. *Cell Host Microbe.*
408 2020; 27: 325-328.

- 409 4. Tyrrell DA, Bynoe ML. Cultivation of viruses from a high proportion of patients with
410 colds. *Lancet*. 1966; 1: 76–77.
- 411 5. Woo PC, Lau SK, Lam CS, Lau CC, Tsang AK, Lau JH, *et al*. Discovery of seven novel
412 Mammalian and avian coronaviruses in the genus deltacoronavirus supports bat
413 coronaviruses as the gene source of alphacoronavirus and betacoronavirus and avian
414 coronaviruses as the gene source of gammacoronavirus and deltacoronavirus. *J Virol*.
415 2012; 86: 3995-4008.
- 416 6. Li F. Structure, function, and evolution of coronavirus spike proteins. *Annu Rev Virol*.
417 2016; 3: 237-261.
- 418 7. Tang Q, Song Y, Shi M, Cheng Y, Zhang W, Xia XQ. Inferring the hosts of coronavirus
419 using dual statistical models based on nucleotide composition. *Sci Rep*. 2015; 5: 17155.
- 420 8. Fehr AR, Perlman S. Coronaviruses: an overview of their replication and pathogenesis.
421 *Methods Mol Biol*. 2015; 1282: 1–23.
- 422 9. Lau SK, Woo PC, Li KS, Huang Y, Tsoi HW, Wong BH, *et al*. Severe acute respiratory
423 syndrome coronavirus-like virus in Chinese horseshoe bats. *Proc Natl Acad Sci U S A*.
424 2005; 102: 14040–14045.
- 425 10. Meyer B, Muller MA, Corman VM, Reusken CB, Ritz D, Godeke GJ, *et al*. Antibodies
426 against MERS coronavirus in dromedary camels, United Arab Emirates, 2003 and 2013.
427 *Emerg Infect Dis*. 2014; 20: 552–559.
- 428 11. Zhang C, Zheng W, Huang X, Bell EW, Zhou X, Zhang Y. Protein structure and sequence
429 re-analysis of 2019-nCoV genome does not indicate snakes as its intermediate host or the
430 unique similarity between its spike protein insertions and HIV-1. 2020;
431 arXiv:2002.03173[q-bio.GN].
- 432 12. Neuman BW, Adair BD, Yoshioka C, Quispe JD, Orca G, Kuhn P, *et al*. Supramolecular
433 architecture of severe acute respiratory syndrome coronavirus revealed by electron
434 cryomicroscopy. *J virol*. 2006; 80:7918–7928.
- 435 13. Barcena M, Oostergetel GT, Bartelink W, Faas FG, Verkleij A, Rottier PJ, *et al*. Cryo-
436 electron tomography of mouse hepatitis virus: Insights into the structure of the
437 coronavirion. *Proc Natl Acad Sci U S A*. 2009; 106: 582–587.
- 438 14. Chen Y, Liu Q, Guo D. Emerging coronaviruses: Genome structure, replication, and
439 pathogenesis. *J. Med. Virol*. 2020; 92: 418-423.

- 440 15. Collins AR, Knobler RL, Powell H, Buchmeier MJ. Monoclonal antibodies to murine
441 hepatitis virus-4 (strain JHM) define the viral glycoprotein responsible for attachment and
442 cell--cell fusion. *Virology*. 1982; 119: 358–371.
- 443 16. Neuman BW, Kiss G, Kunding AH, Bhella D, Baksh MF, Connelly S, *et al.* A structural
444 analysis of M protein in coronavirus assembly and morphology. *J Struct Biol*. 2011; 174:
445 11–22.
- 446 17. Ruch TR, Machamer CE. The coronavirus E protein: assembly and beyond. *Viruses*. 2012;
447 4: 363-382.
- 448 18. McBride R, van Zyl M, Fielding BC. The coronavirus nucleocapsid is a multifunctional
449 protein. *Viruses*. 2014; 6: 2991–3018.
- 450 19. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, *et al.* Genomic characterisation and
451 epidemiology of 2019 novel coronavirus: implications for virus origins and receptor
452 binding. *Lancet*. 2020; 395: 565-574.
- 453 20. Lv L, Li G, Chen J, Liang X, Li Y. Comparative genomic analysis revealed specific mutation
454 pattern between human coronavirus SARS-CoV-2 and Bat-SARSr-CoV RaTG13. *bioRxiv*, 2020;
455 doi: <https://doi.org/10.1101/2020.02.27.969006>.
- 456 21. Sah R, Alfonso J, Rodriguez-Morales, Jha R, Daniel KW, Chu HG, *et al.* Complete genome
457 sequence of a 2019 novel coronavirus (SARS-CoV-2) strain isolated in Nepal. *Microbiol*
458 *Res Announce*. 2020; 9: e00169-20.
- 459 22. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*. 2014; 30: 2068-
460 2069.
- 461 23. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome
462 assemblies. *Bioinformatics*. 2013; 29: 1072-1075.
- 463 24. Edgar E. MUSCLE: multiple sequence alignment with high accuracy and high throughput.
464 *Nucleic Acids Res*. 2004; 32:1792-1797.
- 465 25. Pond SL, Frost SD, Muse, SV. HyPhy: hypothesis testing using phylogenies.
466 *Bioinformatics*. 2005; 21: 676-679.
- 467 26. Weaver S, Shank SD, Spielman SJ, Li M, Muse SV, Kosakovsky Pond SL. Datamonkey
468 2.0: A Modern Web Application for Characterizing Selective and Other Evolutionary
469 Processes. *Mol Biol Evol*. 2018; 35: 773-777.

- 470 27. Nakamura Y, Tomii K. Parallelization of MAFFT for large-scale multiple sequence
471 alignments. *Bioinformatics*. 2018; 34: 2490–2492.
- 472 28. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, *et al.* Roary: rapid
473 large-scale prokaryote pan genome analysis. *Bioinformatics*. 2015; 31: 3691-3693.
- 474 29. Tamura K, Nei M. Estimation of the number of nucleotide substitutions in the control
475 region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol*. 1993; 10: 512-
476 526.
- 477 30. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis
478 version 7.0 for bigger datasets. *Mol Biol Evol*. 2016; 33: 1870-1874.
- 479 31. Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and
480 annotation of phylogenetic and other trees. *Nucleic Acids Res*. 2016; 44: W242-245.
- 481 32. Zhou Z, Alikhan NF, Sergeant MJ, Luhmann N, Vaz C, Francisco AP, *et al.* GrapeTree:
482 visualization of core genomic relationships among 100,000 bacterial pathogens. *Genome*
483 *Res*. 2018; 28: 1395-1404.
- 484 33. Treangen TJ, Ondov BD, Koren S, Phillippy AM. The Harvest suite for rapid core-genome
485 alignment and visualization of thousands of intraspecific microbial genomes. *Genome*
486 *Biol*. 2014; 15: 524.
- 487 34. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic Local Alignment Search
488 Tool. *J Mol Biol*. 1990; 215: 403-410.
- 489 35. Stano M, Beke G, Klucar L. viruSITE—integrated database for viral genomics. *Database*
490 2, 2016; article ID baw162; doi:10.1093/database/baw162.
- 491 36. Cook HV, Doncheva NT, Szklarczyk D, von Mering C, Jensen LJ. Viruses.STRING: A
492 Virus-Host Protein-Protein Interaction Database. *Viruses*. 2018; 10: 519.
- 493 37. Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter, F, Chen C, *et al.* The IntAct
494 molecular interaction database in 2012. *Nucleic Acids Res*. 2013; 41: D43-D47.
- 495 38. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, *et al.*
496 STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic*
497 *Acids Res*. 2015; **43**: D447-452.
- 498 39. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez Gene: gene-centered information at
499 NCBI. *Nucleic Acids Res*. 2005; 33: D54–D58.

- 500 40. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software
501 environment for integrated models of biomolecular interaction networks. *Genome Res.* 2003;
502 13: 2498-2504.
- 503 41. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids*
504 *Res.* 2000; 28: 27–30.
- 505 42. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference
506 resource for gene and protein annotation. *Nucleic Acids Res.* 2016; 44: D457–D462.
- 507 43. UniProt Consortium. The universal protein resource (UniProt). *Nucleic Acids Res.* 2007;
508 **36**: D190-195.
- 509 44. Ren LL, Wang YM, Wu ZQ, Xiang ZC, Guo L, Xu T, *et al.* Identification of a novel
510 coronavirus causing severe pneumonia in human: a descriptive study. *Chin Med J (Engl).*
511 2020; doi: 10.1097/CM9.0000000000000722.
- 512 45. Gorbalenya AE, Baker SC, Baric RS, de Groot RJ, Drosten C, Gulyaeva AA, *et al.* The
513 species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and
514 naming it SARS-CoV-2. *Nat Microbiol.* 2020; 5: 536-544.
- 515 46. Snijder EJ, Decroly E, Ziebuhr J. The non-structural proteins directing coronavirus RNA
516 synthesis and processing. *Adv Virus Res.* 2016; 96: 59-126.
- 517 47. Jang KJ, Jeong S, Kang DY, Sp N, Yang YM, Kim DE. A high ATP concentration
518 enhances the cooperative translocation of the SARS coronavirus helicase nsP13 in the
519 unwinding of duplex RNA. *Sci Rep.* 2020; 10: 1-13.
- 520 48. Becares M, Pascual-Iglesias A, Nogales A, Sola I, Enjuanes L, Zuñiga S. Mutagenesis of
521 coronavirus nsp14 reveals its potential role in modulation of the innate immune response. *J*
522 *Viro.* 2016; 90: 5399-5414.
- 523 49. Athmer J, Fehr AR, Grunewald M, Smith EC, Denison MR, Perlman S. In situ tagged
524 nsp15 reveals interactions with coronavirus replication/transcription complex-associated
525 proteins. *mBio.* 2017; 8: e02320-16.
- 526 50. Von Grotthuss M, Wyrwicz LS, Rychlewski L. mRNA cap-1 methyltransferase in the
527 SARS genome. *Cell.* 2003; 113: 701-702.
- 528 51. Kryazhimskiy S, Plotkin JB. The Population Genetics of dN/dS. *PLoS Genet.* 2008; 4:
529 e1000304.

- 530 52. Kosakovsky Pond SL, Frost SD. (2005). Not So Different After All: A Comparison of
531 methods for detecting amino acid sites under selection. *Mol Biol Evol.* 2005; 22:1208–
532 1222.
- 533 53. Huang L, Shi Y, Gong B, Jiang L, Liu X, Yang J, *et al.* Blood single cell immune profiling
534 reveals the interferon-MAPK pathway mediated adaptive immune response for COVID-
535 19. *BMJ.* 2020; ,doi: <https://doi.org/10.1101/2020.03.15.20033472>.
- 536 54. Tatsuta T, Model K, Langer T. Formation of membrane-bound ring complexes by
537 prohibitins in mitochondria. *Mol Biol Cell.* 2005; 16: 248-259.
- 538 55. Cornillez-Ty CT, Liao L, Yates JR3rd, Kuhn P, Buchmeier MJ. Severe acute respiratory
539 syndrome coronavirus nonstructural protein 2 interacts with a host protein complex
540 involved in mitochondrial biogenesis and intracellular signaling. *J Virol.* 2009; 83: 10314-
541 10318.
- 542 56. Chen Y, Su C, Ke M, Jin X, Xu L, Zhang Z, *et al.* Biochemical and structural insights into
543 the mechanisms of SARS coronavirus RNA ribose 2'-O-methylation by nsp16/nsp10
544 protein complex. *PLoS Pathog.* 2011; 7: e1002294.
- 545 57. Fung TS, Liu DX. Human Coronavirus: Host-Pathogen Interaction. *Annu Rev Microbiol.*
546 2019; 73: 529–557.
- 547 58. Wan YY, Flavell RA. ‘Yin-Yang’ functions of transforming growth factor-beta and T
548 regulatory cells in immune regulation. *Immunol Rev.* 2007; 220: 199-213.
- 549 59. Li SW, Wang CY, Jou YJ, Jou YJ, Tang TC, Huang SH, *et al.* SARS coronavirus papain-
550 like protease induces Egr-1-dependent up-regulation of TGF- β 1 via ROS/p38
551 MAPK/STAT3 pathway. *Sci Rep.* 2016; 6: 25754.
- 552 60. Xu Z, Shi L, Wang Y, Zhang J, Huang L, Zhang C, *et al.* Pathological findings of COVID-
553 19 associated with acute respiratory distress syndrome. *Lancet Respir Med.* 2020; doi:
554 10.1016/S2213-2600(20)30076-X.
- 555 61. Huang Y, Wang X, Li X, Ren L, Zhao J, Hu Y, *et al.* Clinical features of patients infected
556 with 2019 novel coronavirus in Wuhan, China. *Lancet.* 2020; 395: 497-506.
- 557 62. Qin H, Wang L, Feng T, Elson CO, Niyongere SA, Lee AJ, *et al.* TGF-beta promotes Th17
558 cell development through inhibition of SOCS3. *J Immunol.* 2009; 183: 97–105.

- 559 63. Josset L, Menachery VD, Gralinski LE, Agnihothram S, Sova P, Carter VS, *et al.* Cell
560 host response to infection with novel human coronavirus EMC predicts potential antivirals
561 and important differences with SARS coronavirus. *mBio*. 2013; 4: e00165-13.
- 562 64. Prompetchara E, Ketloy C, Palaga T. Immune responses in COVID-19 and potential
563 vaccines: Lessons learned from SARS and MERS epidemic. *Asian Pac J Allergy Immunol*.
564 2020; 38: 1-9.
- 565 65. Zhao X, Nicholls JM, Chen Y. Sars-cov nucleocapsid protein interacts with smad3 and
566 modulates TGF- β signaling. *J Biol Chem*. 2008; 283: 3272-80.
- 567 66. Tan YX, Tan THP, Lee MJ-R, Tham PY, Gunalan V, Druce J, *et al.* Induction of apoptosis
568 by the severe acute respiratory syndrome coronavirus 7a protein is dependent on its
569 interaction with the Bcl-X_L protein. *J Virol*. 2007; 81: 6346-6355.
- 570 67. Fielding BC, Gunalan V, Tan TH, Chou CF, Shen S, Khan S, *et al.* Severe acute respiratory
571 syndrome coronavirus protein 7a interacts with hSGT. *Biochem Biophys Res Commun*.
572 2006; 343: 1201-8.
- 573 68. de Haan CA, Smeets M, Vernooij F, Vennema H, Rottier PJ. Mapping of the coronavirus
574 membrane protein domains involved in interaction with the spike protein. *J Virol*. 1999;
575 73: 7441–7452.
- 576 69. Klumperman J, Locker JK, Meijer A, Horzinek MC, Geuze HJ, Rottier PJ. Coronavirus M
577 proteins accumulate in the Golgi complex beyond the site of virion budding. *J Virol*. 1994;
578 68: 6523–6534.
- 579 70. Yuan X, Li J, Shan Y, Yang Z, Zhao Z, Chen B, *et al.* Subcellular localization and
580 membrane association of SARS-CoV 3a protein. *Virus Res*. 2005; 109: 191-202.
- 581 71. Mergia A. The Role of Caveolin 1 in HIV Infection and Pathogenesis. *Viruses*. 2017; 9:
582 129.

583

584 **Figure legends**

585 **Figure 1:** A) Core genome based phylogenetic analysis of SARS-CoV-2 isolates using the
586 Maximum Likelihood method based on the Tamura-Nei model. The analysis involved 95
587 SARS-CoV-2 sequences with a total of 28451 nucleotide positions. Bootstrap values more than
588 70% are shown on branches as blue dots with sizes corresponding to the bootstrap values. The

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

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

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

617 **Tables Legends**

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

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

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

622

623

624

625

626

627

628

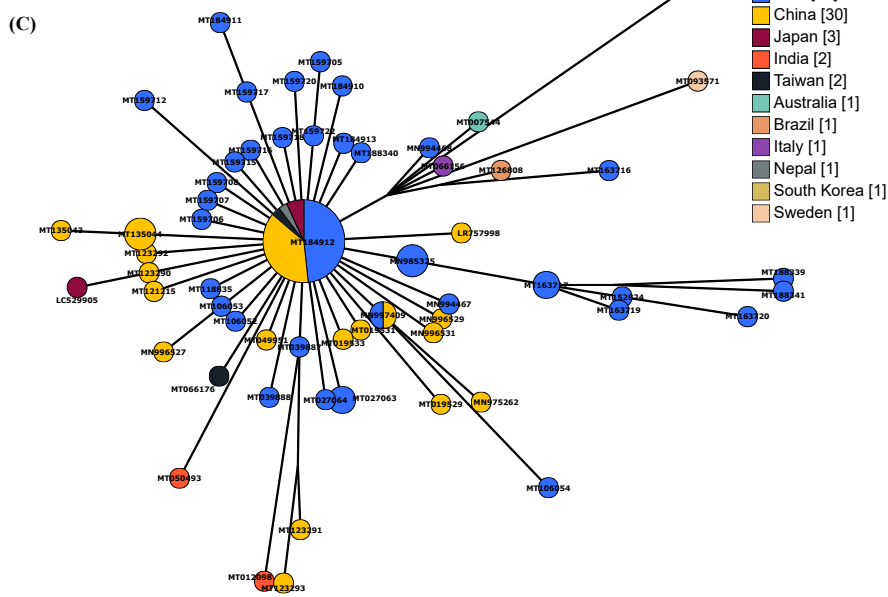
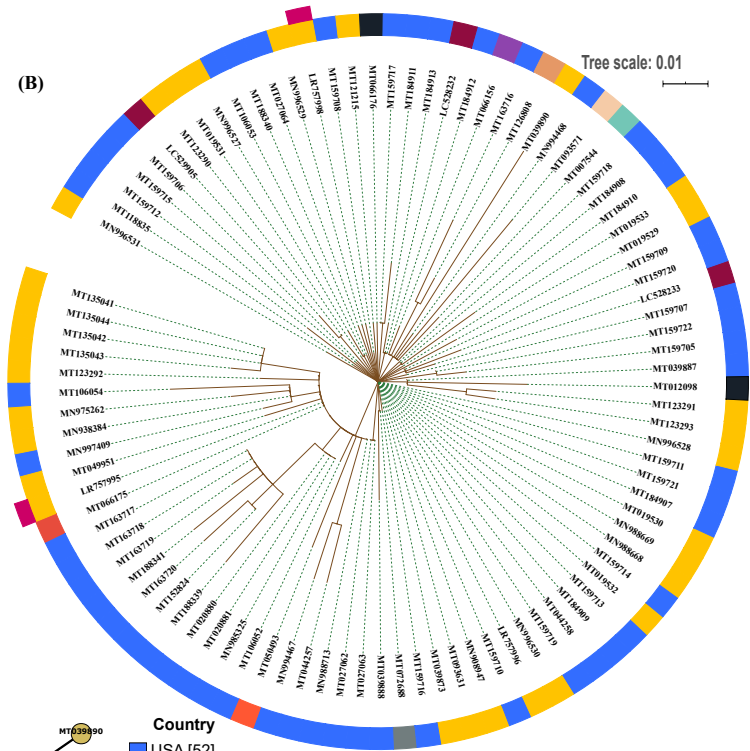
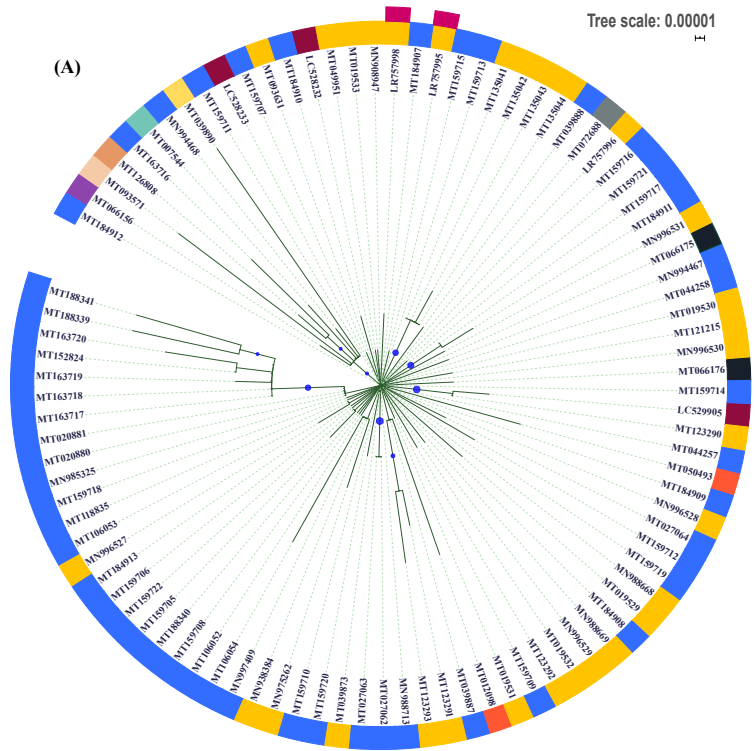
629

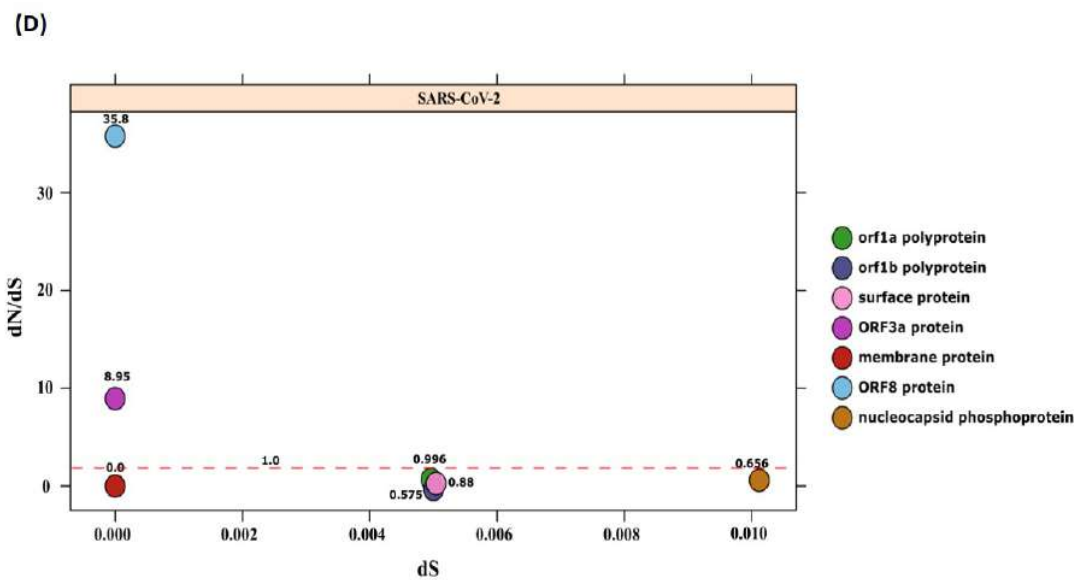
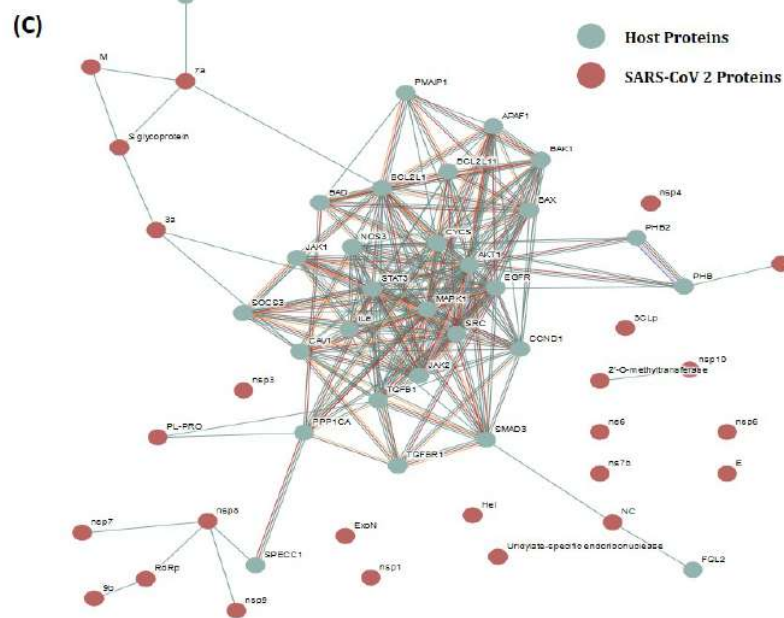
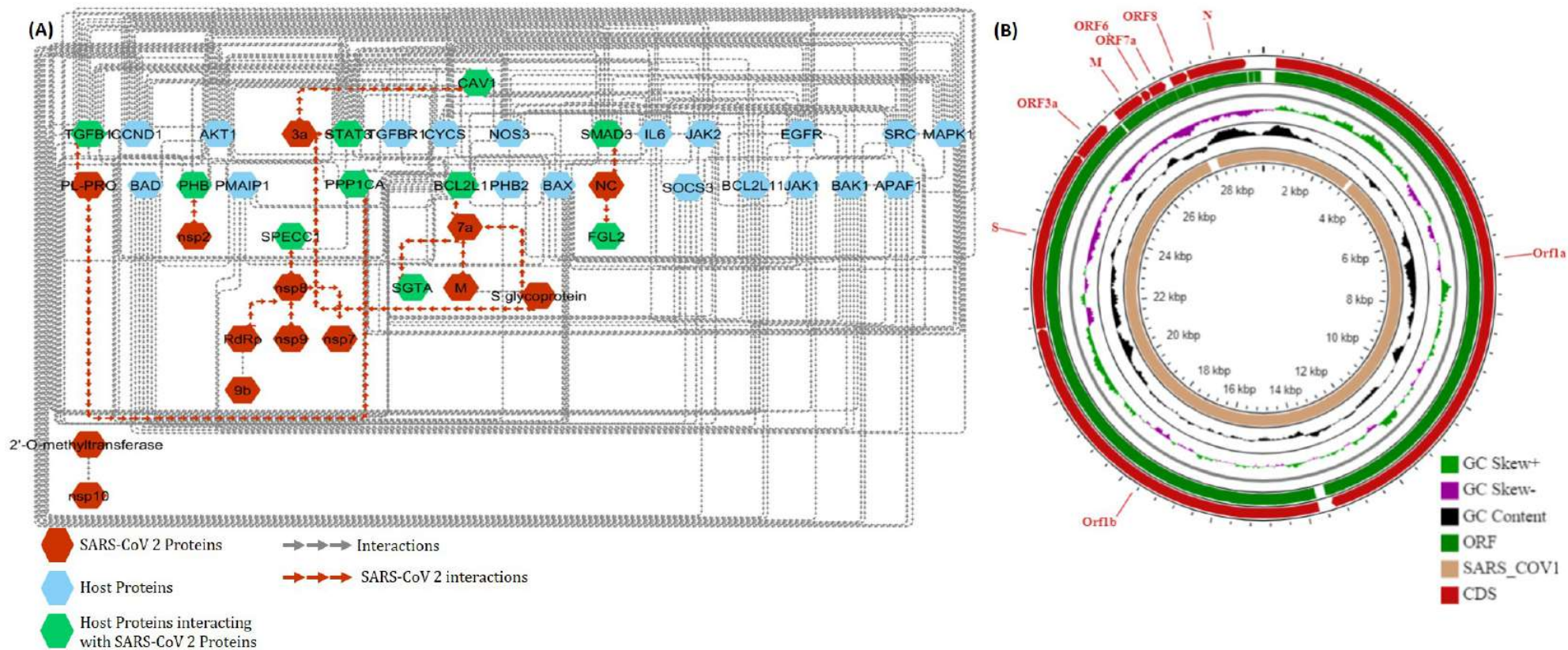
630

631

632

633





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_HK U-SZ-002a_2020	China: Shenzhen	29838	38.02	Oronasopharynx	10/01/2020
9	MN975262.1	2019-nCoV_HK U-SZ-	China	29891	37.98	Oronasopharynx	11/01/2020

		005b_2020					
10	MN985325.1	2019-nCoV/USA-WA1/2020	USA	2988 2	38	Oronasopharynx	19/01/20 20
11	MN988668.1	2019-nCoV WHU01	China	2988 1	38	NA	02/01/20 20
12	MN988669.1	2019-nCoV WHU02	China	2988 1	38	NA	02/01/20 20
13	MN988713.1	2019-nCoV/USA-IL1/2020	USA	2988 2	37.99	Lung, Oronasopharynx	21/01/20 20
14	MN994467.1	2019-nCoV/USA-CA1/2020	USA	2988 2	38	Oronasopharynx	23/12/20 20
15	MN994468.1	2019-nCoV/USA-CA2/2020	USA	2988 3	37.99	Oronasopharynx	22/01/20 20
16	MN996527.1	WIV02	China	2982 5	38.02	Lung	30/12/20 19
17	MN996528.1	WIV04	China	2989 1	37.99	Lung	30/12/20 19
18	MN996529.1	WIV05	China	2985 2	38.02	Lung	30/12/20 19
19	MN996530.1	WIV06	China	2985 4	38.03	Lung	30/12/20 19
20	MN996531.1	WIV07	China	2985 7	38.02	Lung	30/12/20 19
21	MN997409.1	2019-nCoV/USA-AZ1/2020	USA	2988 2	37.99	Feces	22/01/20 20
22	MT007544.1	Australia/VIC01/2020	Australia	2989 3	37.97	NA	25/01/20 20

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	2988 2	38	Oronasopharynx	29/01/20 20
32	MT027063.1	2019-nCoV/USA-CA4/2020	USA	2988 2	38	Oronasopharynx	29/01/20 20
33	MT027064.1	2019-nCoV/USA-CA5/2020	USA	2988 2	37.99	Oronasopharynx	29/01/20 20
34	MT039873.1	HZ-1	China	2983 3	38.02	Lung, Oronasopharynx	20/01/20 20
35	MT039887.1	2019-nCoV/USA-WI1/2020	USA	2987 9	38	Oronasopharynx	31/01/20 20
36	MT039888.1	2019-nCoV/USA-MA1/2020	USA	2988 2	37.99	Oronasopharynx	29/01/20 20
37	MT039890.1	SNU01	South Korea	2990 3	37.96	NA	01/2020
38	MT044257.1	2019-nCoV/USA-IL2/2020	USA	2988 2	38	Lung, Oronasopharynx	28/01/20 20
39	MT044258.1	2019-nCoV/USA-CA6/2020	USA	2985 8	38	Oronasopharynx	27/01/20 20
40	MT049951.1	SARS-CoV-2/Yunnan-01/human/2020/CHN	China	2990 3	37.97	Lung, Oronasopharynx	17/01/20 20
41	MT050493.1	SARS-CoV-2/166/huma	Kerala, India	2985 1	38.01	Oronasopharynx	31/01/20 20

		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	2988 2	38	Lung	23/02/20 20
52	MT121215.1	SARS-CoV-2/SH01/human/2020/CHN	China	2994 5	37.91	Oronasopharynx	02/02/20 20
53	MT123290.1	SARS-CoV-2/IQTC01/human/2020/CHN	China	2989 1	38	Oronasopharynx	05/02/20 20
54	MT123291.2	SARS-CoV-2/IQTC02/human/2020/CHN	China	2988 2	37.99	Lung	29/01/20 20
55	MT123292.2	SARS-CoV-2/QT	China	2992 3	38.02	Lung, Oronasopharynx	27/01/20 20
56	MT123293.2	SARS-CoV-2/IQTC03/human/2020/CHN	China	2987 1	38	Feces	29/01/20 20
57	MT126808.1	SARS-CoV-2/SP02/human/2020/BRA	Brazil	2987 6	38	Oronasopharynx	28/02/20 20
58	MT135041.1	SARS-CoV-2/105/human	China:Beijing	2990 3	37.97	NA	26/01/20 20

		n/2020/CH N					
59	MT135042.1	SARS- CoV- 2/231/huma n/2020/CH N	China:Be ijing	2990 3	37.97	NA	28/01/20 20
60	MT135043.1	SARS- CoV- 2/233/huma n/2020/CH N	China:Be ijing	2990 3	37.97	NA	28/01/20 20
61	MT135044.1	SARS- CoV- 2/235/huma n/2020/CH N	China:Be ijing	2990 3	37.97	NA	28/01/20 20
62	MT152824.1	SARS- CoV- 2/WA2/hum an/2020/US A	USA:W A	2987 8	38	Mid nasal swab	24/02/20 20
63	MT159705.1	2019- nCoV/USA- CruiseA- 7/2020	USA	2988 2	37.99	Oronasopharynx	17/02/20 20
64	MT159706.1	2019- nCoV/USA- CruiseA- 8/2020	USA	2988 2	38	Oronasopharynx	17/02/20 20
65	MT159707.1	2019- nCoV/USA- CruiseA-	USA	2988 2	38	Oronasopharynx	17/02/20 20

		10/2020					
66	MT159708.1	2019- nCoV/USA- CruiseA- 11/2020	USA	2988 2	38	Oronasopharynx	17/02/20 20
67	MT159709.1	2019- nCoV/USA- CruiseA- 12/2020	USA	2988 2	38	Oronasopharynx	20/02/20 20
68	MT159710.1	2019- nCoV/USA- CruiseA- 9/2020	USA	2988 2	38	Oronasopharynx	17/02/20 20
69	MT159711.1	2019- nCoV/USA- CruiseA- 13/2020	USA	2988 2	38	Oronasopharynx	20/02/20 20
70	MT159712.1	2019- nCoV/USA- CruiseA- 14/2020	USA	2988 2	37.99	Oronasopharynx	25/02/20 20
71	MT159713.1	2019- nCoV/USA- CruiseA- 15/2020	USA	2988 2	38	Oronasopharynx	18/02/20 20
72	MT159714.1	2019- nCoV/USA- CruiseA- 16/2020	USA	2988 2	38	Oronasopharynx	18/02/20 20
73	MT159715.1	2019- nCoV/USA- CruiseA-	USA	2988 2	38	Oronasopharynx	24/02/20 20

		17/2020					
74	MT159716.1	2019- nCoV/USA- CruiseA- 18/2020	USA	2986 7	38	Oronasopharynx	24/02/20 20
75	MT159717.1	2019- nCoV/USA- CruiseA- 1/2020	USA	2988 2	37.99	Oronasopharynx	17/02/20 20
76	MT159718.1	2019- nCoV/USA- CruiseA- 2/2020	USA	2988 2	37.99	Oronasopharynx	18/02/20 20
77	MT159719.1	2019- nCoV/USA- CruiseA- 3/2020	USA	2988 2	38	Oronasopharynx	18/02/20 20
78	MT159720.1	2019- nCoV/USA- CruiseA- 4/2020	USA	2988 2	37.99	Oronasopharynx	21/02/20 20
79	MT159721.1	2019- nCoV/USA- CruiseA- 5/2020	USA	2988 2	38	Oronasopharynx	21/02/20 20
80	MT159722.1	2019- nCoV/USA- CruiseA- 6/2020	USA	2988 2	37.99	Oronasopharynx	21/02/20 20
81	MT163716.1	SARS- CoV- 2/WA3-	USA:W A	2990 3	37.95	NA	27/02/20 20

		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	2988 2	38	Oronasopharynx	21/02/2020
89	MT184910.1	2019-nCoV/USA-CruiseA-23/2020	USA	2988 2	37.99	Oronasopharynx	18/02/2020
90	MT184911.1	2019-nCoV/USA-CruiseA-24/2020	USA	2988 2	37.97	Oronasopharynx	17/02/2020
91	MT184912.1	2019-nCoV/USA-CruiseA-25/2020	USA	2988 2	38	Oronasopharynx	17/02/2020
92	MT184913.1	2019-nCoV/USA-CruiseA-26/2020	USA	2988 2	37.99	Oronasopharynx	24/02/2020
93	MT188339.1	USA/MN3-MDH3/2020	USA:M N	2978 3	38.01	Oronasopharynx	07/03/2020
94	MT188340.1	USA/MN2-MDH2/2020	USA:M N	2984 5	37.98	Oronasopharynx	09/03/2020
95	MT188341.1	USA/MN1-MDH1/2020	USA:M N	2983 5	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. <u>2</u>	266-13468, 13468- 21555	QII57165.1	86
Nsp1	265-804	NP_828860. <u>2</u>	266-805		84.44
Nsp2	805-2718	NP_828861. <u>2</u>	806-2719		68.34
Nsp3/PL-PRO	2719-8484	NP_828862. <u>2</u>	2720-8554		75.77
Nsp4	8485-9984	NP_904322. <u>1</u>	8555-10054		<80
Nsp5/3CLp	9985- 10902	NP_828863. <u>1</u>	10055- 10972		<90
Nsp6	10903- 11772	NP_828864. <u>1</u>	10973- 11842		88.15
Nsp7	11773- 12021	NP_828865. <u>1</u>	11843- 12091		98.80
Nsp8	12022- 12615	NP_828866. <u>1</u>	12092- 12685		97.47
Nsp9	12616- 12954	NP_828867. <u>1</u>	12686- 13024		97.35
Nsp10	12955- 13371	NP_828868. <u>1</u>	13025- 13441		97.12
Nsp12 (RdRp)	13372- 13398, 13398- 16166	NP_828869. <u>1</u>	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			