



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

Roshan Kumar,^a Helianthous Verma,^b Nirjara Singhvi,^c Utkarsh Sood,^d Vipin Gupta,^e Mona Singh,^e Rashmi Kumari,^f Princy Hira,^g Shekhar Nagar,^c Chandni Talwar,^c Namita Nayyar,^h Shailly Anand,ⁱ Charu Dogra Rawat,^b Mansi Verma,^h Ram Krishan Negi,^c Yogendra Singh,^c  Rup Lal^d

^aP.G. Department of Zoology, Magadh University, Bodh Gaya, Bihar, India

^bDepartment of Zoology, Ramjas College, University of Delhi, Delhi, India

^cDepartment of Zoology, University of Delhi, Delhi, India

^dThe Energy and Resources Institute, New Delhi, India

^ePhiXGen Private Limited, Gurugram, Haryana, India

^fDepartment of Zoology, College of Commerce, Arts & Science, Patliputra University, Patna, Bihar, India

^gDepartment of Zoology, Maitreyi College, University of Delhi, New Delhi, India

^hDepartment of Zoology, Sri Venkateswara College, University of Delhi, New Delhi, India

ⁱDepartment of Zoology, Deen Dayal Upadhyaya College, University of Delhi, New Delhi, India

ABSTRACT The outbreak of coronavirus disease 2019 (COVID-19) that started in Wuhan, China, in December 2019 has spread worldwide, emerging as a global pandemic. The severe respiratory pneumonia caused by novel severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has so far claimed more than 0.38 million lives and has impacted human lives worldwide. However, as the novel SARS-CoV-2 virus displays high transmission rates, the underlying genomic severity is required to be fully understood. We studied the complete genomes of 95 SARS-CoV-2 strains from different geographical regions worldwide to uncover the pattern of the spread of the virus. We show that there is no direct transmission pattern of the virus among neighboring countries, suggesting that its spread is a result of travel of infected humans to different countries. We revealed unique single nucleotide polymorphisms (SNPs) in nonstructural protein 13 (nsp13), nsp14, nsp15, and nsp16 (ORF1b polyproteins) and in the S-protein within 10 viral isolates from the United States. These viral proteins are involved in RNA replication and binding with the human receptors, indicating that the viral variants that are circulating in the population of the United States are different from those circulating in the populations of other countries. In addition, we found an amino acid addition in nsp16 (mRNA cap-1 methyltransferase) of a U.S. isolate (GenBank accession no. [MT188341.1](https://doi.org/10.1128/MT188341.1)) leading to a shift in the amino acid frame from position 2540 onward. Through comparative structural analysis of the wild-type and mutant proteins, we showed that this addition of a phenylalanine residue renders the protein in the mutant less stable, which might affect mRNA cap-1 methyltransferase function. We further analyzed the SARS-CoV-2–human interactome, which revealed that the interferon signaling pathway is targeted by orf1ab during infection and that it also interacts with NF- κ B-repressing factor (NKRF), which is a potential regulator of interleukin-8 (IL-8). We propose that targeting this interaction may subsequently improve the health condition of COVID-19 patients. Our analysis also emphasized that SARS-CoV-2 manipulates spliceosome machinery during infection; hence, targeting splicing might affect viral replication. In conclusion, the replicative machinery of SARS-CoV-2 is targeting interferon and the notch signaling pathway along with spliceosome machinery to evade host challenges.

IMPORTANCE The COVID-19 pandemic continues to storm the world, with over 6.5 million cases worldwide. The severity of the disease varies with the territories and is

Citation Kumar R, Verma H, Singhvi N, Sood U, Gupta V, Singh M, Kumari R, Hira P, Nagar S, Talwar C, Nayyar N, Anand S, Rawat CD, Verma M, Negi RK, Singh Y, Lal R. 2020. Comparative genomic analysis of rapidly evolving SARS-CoV-2 reveals mosaic pattern of phylogeographical distribution. *mSystems* 5:e00505-20. <https://doi.org/10.1128/mSystems.00505-20>.

Editor Ileana M. Cristea, Princeton University

Copyright © 2020 Kumar et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

Address correspondence to Rup Lal, ruplal@gmail.com.

Received 5 June 2020

Accepted 14 July 2020

Published 28 July 2020

mainly influenced by population density and age factor. In this study, we analyzed the transmission pattern of 95 SARS-CoV-2 genomes isolated from 11 different countries. Our study also revealed several nonsynonymous mutations in ORF1b and S-proteins and the impact on their structural stability. Our analysis showed the manipulation of host system by viral proteins through SARS-CoV-2–human protein interactome, which can be useful to understand the impact of virus on human health.

KEYWORDS COVID-2019, SARS-CoV-2, viruses

Since the current outbreak of pandemic coronavirus disease 2019 (COVID-19) caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), assessment of the biogeographical pattern of SARS-CoV-2 isolates and the mutations present at the nucleotide and protein levels is of high interest to many research groups (1–3). Coronaviruses (CoVs), members of *Coronaviridae* family, order *Nidovirales*, have been known to be human pathogens for the last 6 decades (4). Their targets are not limited just to humans but also extend to other mammals and birds (5). Coronaviruses have been classified in the alphacoronavirus, betacoronavirus, gammacoronavirus, and deltacoronavirus groups (6), among which the members of the first two groups are known to infect mammals whereas those of the latter two primarily infect bird species (7). Symptoms in humans range from common cold to respiratory and gastrointestinal distress of various intensities. In the past, more-severe forms caused major outbreaks that included infections by severe acute respiratory syndrome coronavirus (SARS-CoV) (outbreak in 2003 in China) and Middle East respiratory syndrome coronavirus (MERS-CoV) (outbreak in 2012 in the Middle East) (8). Among mammals, bats have been shown to host coronaviruses, acting as their natural reservoirs, through which the viruses may be transmitted to humans via an intermediate host (9). However, SARS-CoV and MERS-CoV were found to have been transmitted from the intermediate hosts palm civets and camel, respectively (10, 11). Though it is not yet clear which animal served as the intermediate host for transmission of SARS-CoV-2 from bats to humans, it was suggested previously that it was most likely a warm-blooded vertebrate (12, 13).

The inherently high recombination frequency and mutation rates of coronavirus genomes enable their easy transmission among different hosts. Structurally, they are positive-sense single-stranded RNA (ssRNA) virions with characteristic spikes projecting from the surface of the capsid coating (14, 15). The spherical capsid and spikes give them a crown-like appearance, due to which the viruses were named “corona,” meaning “crown” or “halo” in Latin. Their genome is nearly 30 kb in length, largest among the RNA viruses, with a 5′ cap and a 3′ poly(A) tail, for translation (16). Coronaviruses consist of four main proteins, the spike (S), membrane (M), envelope (E), and nucleocapsid (N) proteins. The spike protein (~150 kDa) mediates its attachment to host receptor proteins (17). The membrane protein (~25 to 30 kDa) attaches with nucleocapsid and maintains the curvature of the virus membrane (18). The envelope protein (8 to 12 kDa) is responsible for the pathogenesis of the virus as it eases the assembly and release of virion particles and also has ion channel activity as an integral membrane protein (19). Nucleocapsid, the fourth protein, helps in the packaging of virus particles into capsids and promotes formation of the replicase-transcriptase complex (RTC) (20).

The outbreak of novel betacoronavirus (2019-nCoV), or SARS-CoV-2, in December 2019 in Wuhan, China, has shown devastating effects worldwide. It has emerged as a reason for concern not only for its pneumonia-like symptoms but also for its asymptomatic nature and the challenges encountered in efforts to contain it (21, 22). The World Health Organization (WHO) declared COVID-19 a pandemic on 11 March 2020; however, by the time it was declared a pandemic, more than 118,000 cases of the coronavirus illness had been reported from over 113 countries and territories around the world (WHO Situation Report 51). So far, it has affected more than 200 countries and territories, and the number of deaths resulting from the disease have exceeded 0.38 million worldwide. Virtually all human lives have been impacted, with no foreseeable end of the pandemic. SARS-CoV-2 is assumed to have originated from bats, which

serve as a reservoir host of the virus (9). Similar mutation patterns in Bat-SARS-CoV RaTG13 and SARS-CoV-2 were also recently revealed, but the data set was limited to 21 strains, including a few SARS-CoV-2 strains and neighboring strains (9). Numerous studies have now reported the genome composition and divergence patterns of SARS-CoV-2 (3, 23). 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-based and single nucleotide polymorphism (SNP)-based phylogeny, we attempted to uncover the transmission pattern of the virus across the affected countries, which was not known earlier. We analyzed the open reading frames (ORFs) of the isolates to reveal unique point mutations and amino acid substitutions/additions in the isolates from the United States. In addition, we analyzed the gene/protein mutations in these novel strains and estimated the direction of selection to decipher their evolutionary divergence rate. Further, we also established the interactome of SARS-CoV-2 with the human host proteins to predict the functional implications represented by the viral infected host cells. The results obtained from the analyses indicate different variants of SARS-CoV-2 isolates, with an inherent capability of unique mutations and evolving viral replication system enabling adaptation to human hosts. To our knowledge, this is the first study to demonstrate the biogeographical distribution pattern of this emerging pathogen coupled with the high rate of mutations.

RESULTS AND DISCUSSION

General genomic attributes of SARS-CoV-2. In this study, we analyzed a total of 95 SARS-CoV-2 strains (available on 19 March 2020) isolated and sequenced between December 2019 and March 2020 from 11 different countries, namely, the United States ($n = 52$ isolates), China ($n = 30$), Japan ($n = 3$), India ($n = 2$), Taiwan ($n = 2$), Australia ($n = 1$), Brazil ($n = 1$), Italy ($n = 1$), Nepal ($n = 1$), South Korea ($n = 1$), and Sweden ($n = 1$). A total of 68 strains were isolated from either oronasopharynges or lungs, while two of them were isolated from feces, suggesting both respiratory and gastrointestinal SARS-CoV-2 connections (Table 1). No information concerning the source of isolation of the remaining isolates was available. The average genome size and GC content were found to be $29,879 \pm 26.6$ bp and $37.99\% \pm 0.018\%$, respectively. All these isolates were found to harbor 9 open reading frames (ORFs) coding for ORF1a (13,218-bp) and ORF1b (7,788-bp) polyproteins, surface glycoprotein or S-protein (3,822 bp), ORF3a protein (828 bp), membrane glycoprotein (M-protein) (669 bp), ORF6 protein (186 bp), ORF7a protein (366 bp), ORF8 protein (366 bp), and nucleocapsid phosphoprotein (N-protein) (1,260 bp), which agrees with a recently published study (24). ORF1a harbors 12 nonstructural proteins (nsp), namely, nsp1, nsp2, nsp3 (papain-like protease or PLpro domain), nsp4, nsp5 (3C-like protease [3CLpro]), nsp6, nsp7, nsp8, nsp9, nsp10, nsp11, and nsp12 (RNA-dependent RNA polymerase [RdRp]) (24). Similarly, ORF1b contains four putative nonstructural proteins, namely, nsp13 (helicase or Hel), nsp14 (3'-to-5' exoribonuclease or ExoN), nsp15, and nsp16 (mRNA cap-1 methyltransferase).

Phylogenomic analysis: defining evolutionary relatedness. Our analysis revealed that strains of human infecting SARS-CoV-2 are novel and highly similar ($>99.9\%$). A recent study established the closest neighbor of SARS-CoV-2 to be SARSr-CoV-RaTG13, a bat coronavirus (25). As COVID-19 transitioned from epidemic to pandemic due to the extremely contagious nature of the SARS-CoV-2, it was interesting to delineate the relationship between strains and their geographical locations. In this study, we employed two methods to delineate the phylogenomic relatedness of the isolates: analyses of the core genome (Fig. 1) and single nucleotide polymorphisms (SNPs) (Fig. 2A). The phylogenies obtained were annotated with the country of isolation of each strain (Fig. 1). Using GrapeTree, we identified strains which were diverging from the common core population (Fig. 1B). For example, two isolates from the United States (GenBank accession no. [MN994468](#) and [MT163716](#)) and one each from Australia ([MT007544](#)), Italy ([MT066156](#)), Sweden ([MT093571](#)), South Korea ([MT039890](#)), and Brazil ([MT126808](#)) were

TABLE 1 General genomic attributes of SARS-CoV-2 strains

Strain no.	Accession no.	Virus (SARS-CoV-2)	Country of origin	Genome size (bp)	GC%	Isolation source(s)	Date of isolation
1	LC528232.1	Hu/DP/Kng/19-020	Japan	29,902	37.98	Oronasopharynx	10 February 2020
2	LC528233.1	Hu/DP/Kng/19-027	Japan	29,902	38.02	Oronasopharynx	10 February 2020
3	LC529905.1	TKYE6182_2020	Japan	29,903	37.97	NA ^a	January 2020
4	LR757995.1	Wuhan seafood market pneumonia virus	China (Wuhan)	29,872	38	NA	5 January 2020
5	MT163720.1	WA8-UW5/human/2020/USA	United States	29,732	37.97	NA	1 March 2020
6	LR757998.1	Wuhan seafood market pneumonia virus	China (Wuhan)	29,866	37.99	NA	26 December 2019
7	MN908947.3	Wuhan-Hu-1	China	29,903	37.97	NA	December 2019
8	MN938384.1	2019-nCoV_HKU-SZ-002a_2020	China (Shenzhen)	29,838	38.02	Oronasopharynx	10 January 2020
9	MN975262.1	2019-nCoV_HKU-SZ-005b_2020	China	29,891	37.98	Oronasopharynx	11 January 2020
10	MN985325.1	2019-nCoV/USA-WA1/2020	United States	29,882	38	Oronasopharynx	19 January 2020
11	MN988668.1	2019-nCoV WHU01	China	29,881	38	NA	2 January 2020
12	MN988669.1	2019-nCoV WHU02	China	29,881	38	NA	2 January 2020
13	MN988713.1	2019-nCoV/USA-IL1/2020	United States	29,882	37.99	Lung, oronasopharynx	21 January 2020
14	MN994467.1	2019-nCoV/USA-CA1/2020	United States	29,882	38	Oronasopharynx	23 December 2019
15	MN994468.1	2019-nCoV/USA-CA2/2020	United States	29,883	37.99	Oronasopharynx	22 January 2020
16	MN996527.1	WIV02	China	29,825	38.02	Lung	30 December 2019
17	MN996528.1	WIV04	China	29,891	37.99	Lung	30 December 2019
18	MN996529.1	WIV05	China	29,852	38.02	Lung	30 December 2019
19	MN996530.1	WIV06	China	29,854	38.03	Lung	30 December 2019
20	MN996531.1	WIV07	China	29,857	38.02	Lung	30 December 2019
21	MN997409.1	2019-nCoV/USA-AZ1/2020	United States	29,882	37.99	Feces	22 January 2020
22	MT007544.1	Australia/VIC01/2020	Australia	29,893	37.97	NA	25 January 2020
23	MT012098.1	SARS-CoV-2/29/human/2020/IND	Kerala, India	29,854	38.02	Oronasopharynx	27 January 2020
24	MT019529.1	BetaCoV/Wuhan/IPBCAMS-WH-01/2019	China	29,899	37.98	Lung	23 December 2019
25	MT019530.1	BetaCoV/Wuhan/IPBCAMS-WH-02/2019	China	29,889	38	Lung	30 December 2019
26	MT019531.1	BetaCoV/Wuhan/IPBCAMS-WH-03/2019	China	29,899	37.98	Lung	30 December 2019
27	MT019532.1	BetaCoV/Wuhan/IPBCAMS-WH-04/2019	China	29,890	37.99	Lung	30 December 2019
28	MT019533.1	BetaCoV/Wuhan/IPBCAMS-WH-05/2020	China	29,883	37.99	Lung	1 January 2020
29	MT020880.1	2019-nCoV/USA-WA1-A12/2020	United States	29,882	38	Oronasopharynx	25 January 2020
30	MT020881.1	2019-nCoV/USA-WA1-F6/2020	United States	29,882	38	Oronasopharynx	25 January 2020
31	MT027062.1	2019-nCoV/USA-CA3/2020	United States	29,882	38	Oronasopharynx	29 January 2020
32	MT027063.1	2019-nCoV/USA-CA4/2020	United States	29,882	38	Oronasopharynx	29 January 2020
33	MT027064.1	2019-nCoV/USA-CA5/2020	United States	29,882	37.99	Oronasopharynx	29 January 2020
34	MT039873.1	HZ-1	China	29,833	38.02	Lung, Oronasopharynx	20 January 2020
35	MT039887.1	2019-nCoV/USA-W11/2020	United States	29,879	38	Oronasopharynx	31 January 2020
36	MT039888.1	2019-nCoV/USA-MA1/2020	United States	29,882	37.99	Oronasopharynx	29 January 2020
37	MT039890.1	SNU01	South Korea	29,903	37.96	NA	January 2020
38	MT044257.1	2019-nCoV/USA-IL2/2020	United States	29,882	38	Lung, Oronasopharynx	28 January 2020
39	MT044258.1	2019-nCoV/USA-CA6/2020	United States	29,858	38	Oronasopharynx	27 January 2020
40	MT049951.1	SARS-CoV-2/Yunnan-01/human/2020/CHN	China	29,903	37.97	Lung, Oronasopharynx	17 January 2020
41	MT050493.1	SARS-CoV-2/166/human/2020/IND	Kerala, India	29,851	38.01	Oronasopharynx	31 January 2020
42	MT066156.1	SARS-CoV-2/NM	Italy	29,867	38.01	Lung, Oronasopharynx	30 January 2020
43	MT066175.1	SARS-CoV-2/NTU01/2020/TWN	Taiwan	29,870	38.01	NA	31 January 2020
44	MT066176.1	SARS-CoV-2/NTU02/2020/TWN	Taiwan	29,870	38.01	NA	5 February 2020
45	MT072688.1	SARS0CoV-2/61-TW/human/2020/ NPL	Nepal	29,811	38.02	Oronasopharynx	13 February 2020
46	MT093571.1	SARS-CoV-2/01/human/2020/SWE	Sweden	29,886	38	NA	7 February 2020
47	MT093631.2	SARS-CoV-2/WH-09/human/2020/CHN	China	29,860	38.02	Oronasopharynx	8 January 2020
48	MT106052.1	2019-nCoV/USA-CA7/2020	United States	29,882	37.99	Oronasopharynx	6 February 2020
49	MT106053.1	2019-nCoV/USA-CA8/2020	United States (CA)	29,882	38	Oronasopharynx	10 February 2020
50	MT106054.1	2019-nCoV/USA-TX1/2020	United States (TX)	29,882	38	Lung, Oronasopharynx	11 February 2020
51	MT118835.1	2019-nCoV/USA-CA9/2020	United States (CA)	29,882	38	Lung	23 February 2020
52	MT121215.1	SARS-CoV-2/SH01/human/2020/CHN	China	29,945	37.91	Oronasopharynx	2 February 2020
53	MT123290.1	SARS-CoV-2/IQTC01/human/2020/CHN	China	29,891	38	Oronasopharynx	5 February 2020
54	MT123291.2	SARS-CoV-2/IQTC02/human/2020/CHN	China	29,882	37.99	Lung	29 January 2020
55	MT123292.2	SARS-CoV-2/QT	China	29,923	38.02	Lung, Oronasopharynx	27 January 2020
56	MT123293.2	SARS-CoV-2/IQTC03/human/2020/CHN	China	29,871	38	Feces	29 January 2020
57	MT126808.1	SARS-CoV-2/SPO2/human/2020/BRA	Brazil	29,876	38	Oronasopharynx	28 February 2020
58	MT135041.1	SARS-CoV-2/105/human/2020/CHN	China:Beijing	29,903	37.97	NA	26 January 2020
59	MT135042.1	SARS-CoV-2/231/human/2020/CHN	China:Beijing	29,903	37.97	NA	28 January 2020

(Continued on next page)

TABLE 1 (Continued)

Strain no.	Accession no.	Virus (SARS-CoV-2)	Country of origin	Genome size (bp)	GCC%	Isolation source(s)	Date of isolation
60	MT135043.1	SARS-CoV-2/233/human/2020/CHN	China:Beijing	29,903	37.97	NA	28 January 2020
61	MT135044.1	SARS-CoV-2/235/human/2020/CHN	China:Beijing	29,903	37.97	NA	28 January 2020
62	MT152824.1	SARS-CoV-2/WA2/human/2020/USA	United States (WA)	29,878	38	Mid-nasal swab	24 February 2020
63	MT159705.1	2019-nCoV/USA-CruiseA-7/2020	United States	29,882	37.99	Oronasopharynx	17 February 2020
64	MT159706.1	2019-nCoV/USA-CruiseA-8/2020	United States	29,882	38	Oronasopharynx	17 February 2020
65	MT159707.1	2019-nCoV/USA-CruiseA-10/2020	United States	29,882	38	Oronasopharynx	17 February 2020
66	MT159708.1	2019-nCoV/USA-CruiseA-11/2020	United States	29,882	38	Oronasopharynx	17 February 2020
67	MT159709.1	2019-nCoV/USA-CruiseA-12/2020	United States	29,882	38	Oronasopharynx	20 February 2020
68	MT159710.1	2019-nCoV/USA-CruiseA-9/2020	United States	29,882	38	Oronasopharynx	17 February 2020
69	MT159711.1	2019-nCoV/USA-CruiseA-13/2020	United States	29,882	38	Oronasopharynx	20 February 2020
70	MT159712.1	2019-nCoV/USA-CruiseA-14/2020	United States	29,882	37.99	Oronasopharynx	25 February 2020
71	MT159713.1	2019-nCoV/USA-CruiseA-15/2020	United States	29,882	38	Oronasopharynx	18 February 2020
72	MT159714.1	2019-nCoV/USA-CruiseA-16/2020	United States	29,882	38	Oronasopharynx	18 February 2020
73	MT159715.1	2019-nCoV/USA-CruiseA-17/2020	United States	29,882	38	Oronasopharynx	24 February 2020
74	MT159716.1	2019-nCoV/USA-CruiseA-18/2020	United States	29,867	38	Oronasopharynx	24 February 2020
75	MT159717.1	2019-nCoV/USA-CruiseA-1/2020	United States	29,882	37.99	Oronasopharynx	17 February 2020
76	MT159718.1	2019-nCoV/USA-CruiseA-2/2020	United States	29,882	37.99	Oronasopharynx	18 February 2020
77	MT159719.1	2019-nCoV/USA-CruiseA-3/2020	United States	29,882	38	Oronasopharynx	18 February 2020
78	MT159720.1	2019-nCoV/USA-CruiseA-4/2020	United States	29,882	37.99	Oronasopharynx	21 February 2020
79	MT159721.1	2019-nCoV/USA-CruiseA-5/2020	United States	29,882	38	Oronasopharynx	21 February 2020
80	MT159722.1	2019-nCoV/USA-CruiseA-6/2020	United States	29,882	37.99	Oronasopharynx	21 February 2020
81	MT163716.1	SARS-CoV-2/WA3-UW1/human/2020/USA	United States (WA)	29,903	37.95	NA	27 February 2020
82	MT163717.1	SARS-CoV-2/WA4-UW2/human/2020/USA	United States (WA)	29,897	37.97	NA	28 February 2020
83	MT163718.1	SARS-CoV-2/WA6-UW3/human/2020/USA	United States (WA)	29,903	37.97	NA	29 February 2020
84	MT163719.1	SARS-CoV-2/WA7-UW4/human/2020/USA	United States (WA)	29,903	37.97	NA	1 March 2020
85	LR757996.1	Wuhan seafood market pneumonia virus	China (Wuhan)	29,732	37.96	NA	1 January 2020
86	MT184907.1	2019-nCoV/USA-CruiseA-19/2020	United States	29,882	38	Oronasopharynx	18 February 2020
87	MT184908.1	2019-nCoV/USA-CruiseA-21/2020	United States	29,880	38	Oronasopharynx	17 February 2020
88	MT184909.1	2019-nCoV/USA-CruiseA-22/2020	United States	29,882	38	Oronasopharynx	21 February 2020
89	MT184910.1	2019-nCoV/USA-CruiseA-23/2020	United States	29,882	37.99	Oronasopharynx	18 February 2020
90	MT184911.1	2019-nCoV/USA-CruiseA-24/2020	United States	29,882	37.97	Oronasopharynx	17 February 2020
91	MT184912.1	2019-nCoV/USA-CruiseA-25/2020	United States	29,882	38	Oronasopharynx	17 February 2020
92	MT184913.1	2019-nCoV/USA-CruiseA-26/2020	United States	29,882	37.99	Oronasopharynx	24 February 2020
93	MT188339.1	USA/MN3-MDH3/2020	United States (MN)	29,783	38.01	Oronasopharynx	7 March 2020
94	MT188340.1	USA/MN2-MDH2/2020	United States (MN)	29,845	37.98	Oronasopharynx	9 March 2020
95	MT188341.1	USA/MN1-MDH1/2020	United States (MN)	29,835	37.99	Oronasopharynx	5 March 2020

found to have diverged from a recent common ancestor (Fig. 1B). Further, the phylogenetic clustering was found majorly concordant by both the core-genome-based (Fig. 1A) and the SNP-based (Fig. 2A) methods. The strains formed a monophyletic clade, in which [MT039890](#) (Sweden) and [MT093571](#) (South Korea) were the most diverged. Focusing on the edge connection between the neighboring countries from which the transmission is more likely to occur, we noted that a strain from Taiwan ([MT066176](#)) clustered closely with another strain from China ([MT121215](#)). With the exception of those two strains, we did not find any connection between strains of neighboring countries. Thus, most strains belonging to the same country clustered distantly from each other and showed relatedness to strains isolated from distant geographical locations (Fig. 1A; see also Fig. 2A). For instance, a SARS-CoV-2 strain isolated from Nepal ([MT072688](#)) clustered with a strain from the United States ([MT039888](#)). Also, strains from Wuhan (GenBank accession no. [LR757998](#) and [LR757995](#)), where the virus was originated, showed highest identity with the United

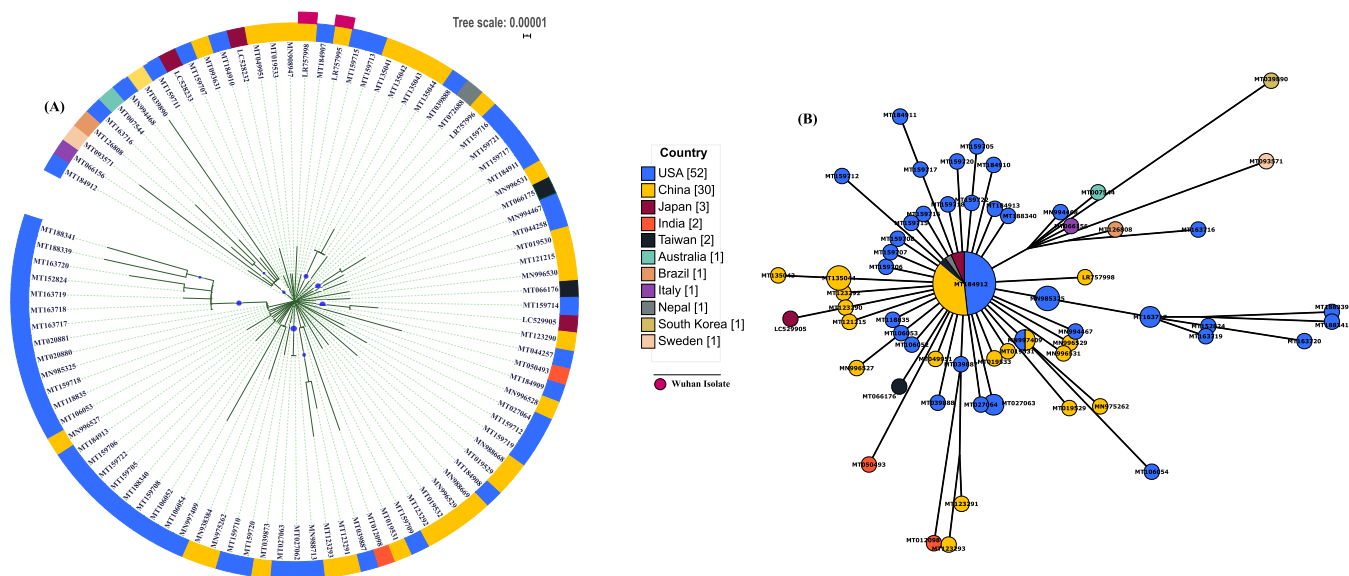


FIG 1 (A) Core genome-based phylogenetic analysis of SARS-CoV-2 isolates using the maximum likelihood method based on the Tamura-Nei model. The analysis involved 95 SARS-CoV-2 sequences with a total of 28,451 nucleotide positions. Bootstrap values of more than 70% are shown on branches as blue dots with sizes corresponding to the bootstrap values. The colored circle represents the country of origin of each isolate. The two isolates from Wuhan are marked separately on the outer side of the ring. (B) 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.

States strains as well as the Chinese strains; strains MT012098 and MT050493 from India (26) clustered closely with China and U.S. strains, respectively (Fig. 1A; see also Fig. 2A). Similarly, an Australian strain (MT007544) showed close clustering with a U.S. strain (Fig. 1A; see also Fig. 2A) and one strain from Taiwan (MT066175) clustered with highly similar Chinese isolates (Fig. 2A). Isolates from Italy (MT012098) and Brazil (MT126808) clustered with different U.S. strains (Fig. 1A; see also Fig. 2A). Notably, isolates from same country or geographical location formed a mosaic pattern of phylogenetic placements of isolates from those countries. For viral transmission, contact between the individuals is also an important factor, due to which the spread of identical strains across the border of neighboring countries is supposedly more likely. But we obtained a pattern where the Indian strains showed highest similarity with the United States and China strains, the Australian strains with the United States strains, and the Italian and Brazilian strains with strains isolated from the United States, among others. This depicts the viral spread across different communities. However, as SARS-CoV-2 genomes were available mostly from the United States and China, sampling biases are evident in the analyzed data set available on NCBI. Thus, it is plausible for strains from other countries to show the highest similarity to strains from these two countries. In the near future, as more and more genome sequences become available from different geographical locations, more-accurate patterns representing their relatedness across the globe will become available.

Recent studies have also focused on the roles of synanthropic animals in transmission of SARS-COV-2 where epitopes of spike and nucleocapsid proteins of taxonomically related coronaviruses of animal species living in close vicinity to humans were compared with those of SARS-COV-2 (27, 28). Here, we also compared the spike and nucleocapsid proteins of other synanthropic animals with those of 95 SARS-CoV-2 strains (see Fig. S1A and B in the supplemental material). Phylogenetic clustering revealed that all the SARS-CoV-2 strains formed a clear-cut separate cluster (Fig. S1A and 1B). The results obtained suggested that the homology of spike and nucleocapsid proteins of SARS-CoV-2 with those of other animal coronaviruses was accountable, but not much higher. This supports the hypothesis that the synanthropic animals may not act as a reservoir for the infection but may rather provide a stimulus for an immune response which may help the organism to fight against SARS-CoV-2 (27, 28). On the

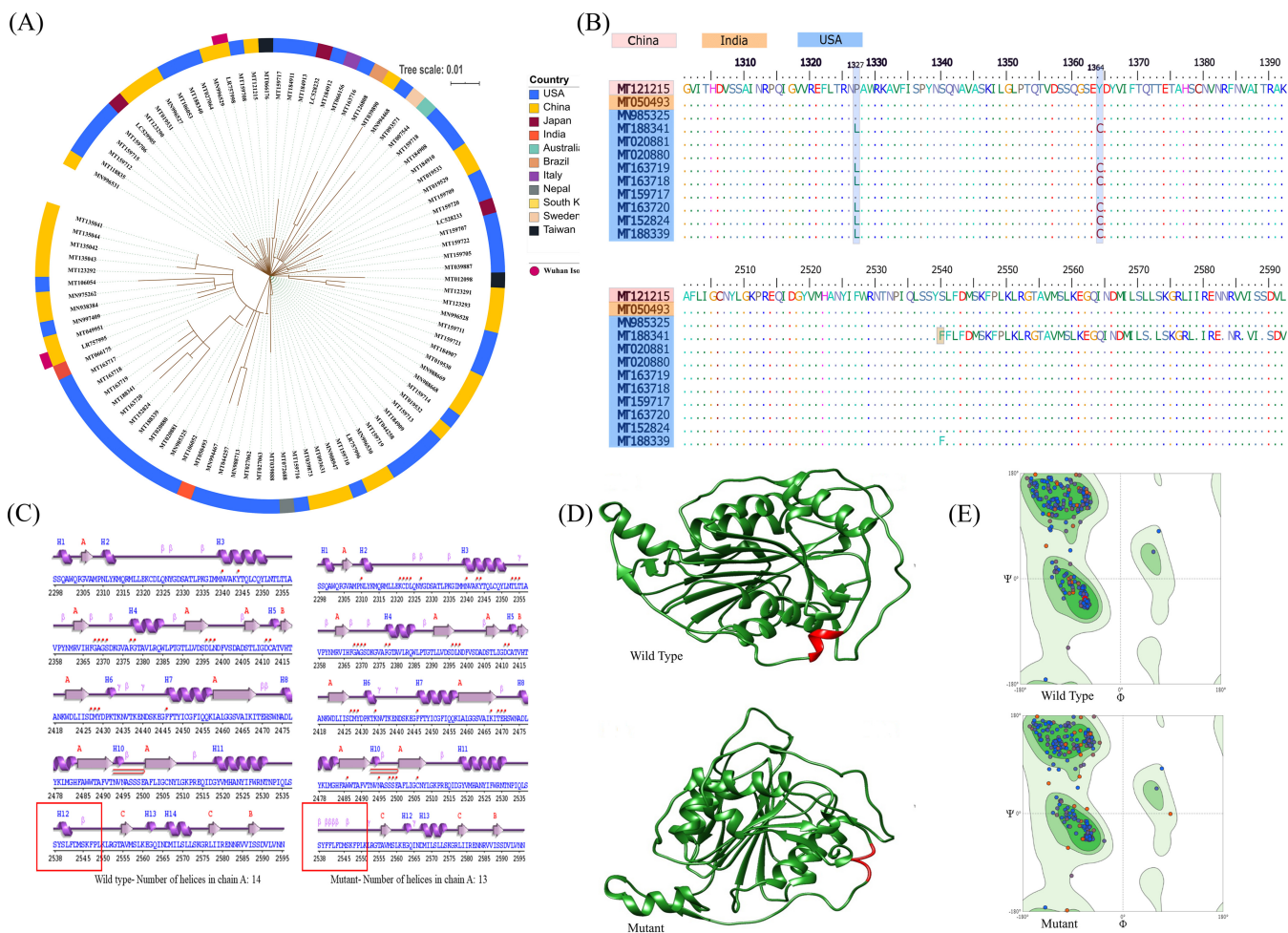


FIG 2 (A) SNP-based phylogeny of SARS-CoV-2 isolates. Highly similar genomes of coronaviruses were taken as the input by Parsnp. Whole-genome alignments were made using libMUSCLE aligner and the annotated genome of MT121215 strain as the reference. Parsnp identifies the maximal unique matches (MUMs) among the query genomes provided in a single directory. As only the genomes corresponding to a specified MUM index (MUMI) distance threshold are recruited, option -c was used to force inclusion of all the strains. The output phylogeny based on single nucleotide polymorphisms was obtained following variant calling on core-genome alignment. (B) Multiple-sequence alignment of ORF1b protein showing amino acid substitutions at three positions: P1327L, Y1364C, and S2540F. The isolate USA/MN1-MDH1/2020 (MT188341) showed an amino acid addition leading to a change in an amino acid frame from position 2540 onward. (C and D) 2D and 3D structures for nsp16 in the wild-type strain (MT121215) and the mutant strain (MT188341) predicted using PDBsum and the SWISS-MODEL. (E) Ramachandran plot of the predicted wild-type and mutant proteins, where the green region represents a most-favored region whereas the light green area denotes an allowed region. The white zone represents a generously allowed region.

other hand, the recurring exposure might lead to eliciting overresponsiveness by the immune system and thus might worsen the symptoms.

SNPs in the SARS-CoV-2 genomes. SNPs in all predicted ORFs in each genome were analyzed using SARS-CoV-2/SH01/human/2020/CHN as a reference. SNPs were assayed using maximum unique matches between the genomes of coronavirus. We observed that the strains isolated from the United States (MT188341, MN985325, MT020881, MT020880, MT163719, MT163718, MT163717, MT152824, MT163720, and MT188339) are the most evolved and that they carry set of unique point mutations (Table 2) in nsp13, nsp14, nsp15, nsp16 (present in the orf1b polyprotein region), and S-protein (Table 2; see also Table S1 in the supplemental material). Most of these mutated proteins are nonstructural proteins (NSP) functionally involved in forming viral replication-transcription complexes (RTC) (29). For instance, nsp13 belongs to helicase superfamily 1 and is putatively involved in viral RNA replication through RNA-DNA duplex unwinding (30) whereas nsp14 and nsp15 represent an exoribonuclease and an endoribonuclease, respectively (31, 32). nsp16 functions as mRNA cap-1 methyltransferase (33). All these proteins contain SNPs at several positions (Table 2), which

TABLE 2 Major mutations present in different isolates of SARS-CoV-2 at different locations

Strain(s) with major mutation(s)	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

^aNA, information not available.

indicates that the viral machinery for its RNA replication and processing had evolved to the utmost in strains from the United States compared to those from the other countries. Further, we analyzed the SNPs at the protein level; interestingly, there were amino acid substitutions at P1327L, Y1364C, and S2540F in the ORF1b protein in U.S. isolates.

One isolate, namely, USA0/MN1-MDH1/2020 (MT188341), carried an amino acid addition at position 2540 leading to a shift in the amino acid frame from there onward (Fig. 2B). The insertion of phenylalanine in the sequence might result in an increased size of the mutant residue compared to the wild-type residue, which might affect the functioning of nsp16 (2'-O-MTase). To further analyze the structural heterogeneity resulting from the insertion of phenylalanine, we predicted the two-dimensional (2D) and 3D structures of the wild type (MT121215) and the mutant (MT188341) based on homology modeling with identity levels of 100% for the reference sequence (template: 6w61.1.A) and 99.66% for the mutant sequence. Interestingly, the protein structures were found to be different (Fig. 2C and D). We predicted the motifs in these strains and observed that one motif from mutant strain at positions 243 to 246 was missing compared to the wild type (Fig. 2C). Thus, the results suggested the presence of 14 helical regions in the wild-type strain, whereas only 13 helices were present in the mutant type (Fig. 2C). The region of mutation/insertion that lacked the helical region was found to harbor beta sheets (Fig. 2C). Further, the 3D structure analysis results showing sequence identity above 99% revealed that the model was constructed with a high confidence value (Fig. 2D). Thereafter, we validated the structures with the help of Ramachandran plot analysis. In case of the wild type, the analysis predicted the presence of 97.97% residues in the favored region, in contrast to the presence of 95.98% residues in the favored region in the mutant (Fig. 2E). Thus, we concluded that the wild-type structure is more stable than the mutant structure. Further, the results of the protein stability analysis showed that this mutation could decrease the stability of this protein with possible effects on size and hydrophobicity. This mutant residue was found to be more hydrophobic than the wild-type residue, which would impact hydrogen bond formation. The residue in the wild type is buried in the core of a domain. The differences between the wild-type and mutant residues might disturb the core structure of this domain. Thus, in our opinion the mutation may lead to a conformational change in the protein structure and affect the functioning of nsp16.

NF- κ B signaling during SARS-CoV-2 infection. It is now well established that COVID-19 pathogenesis is driven by profound cytokines responses such as those of interleukin-6 (IL-6), IL-8, tumor necrosis factor (TNF), IL-1 β , granulocyte colony-stimulating factor (G-CSF), granulocyte-macrophage colony-stimulating factor (GM-CSF), etc. (41). Reports suggest that a phase II clinical trial is focusing on neutralizing IL-8 in order to improve the health condition of COVID-19 patients (<https://clinicaltrials.gov/ct2/show/NCT04347226>). We also found that SARS-CoV-2 protein (orf1ab) interacts with NF- κ B-repressing factor (NKRF), which is a potential regulator of IL-8; thus, targeting this interaction may subsequently improve the health condition of COVID-19 patients.

Further, heterogeneous RNA molecules such as snoRNA/small Cajal body-specific RNA (scaRNA) and snRNA, which are located in the Cajal body (in regions within the nucleus that are enriched in RNAs and proteins) have been reported to facilitate the activity of 2'-O-ribose-methylated nucleotides and pseudouridines in the RNA polymerase II-transcribed U1, U2, U4 and U5 spliceosomes (42). Through our interactome studies, we found that, remarkably, the multiple spliceosome components [SLU7, poly(U)-binding-splicing factor 60 (PUF60), SRSF protein kinase 1 (SRPK1), SRSF5, LUC7L2, U2AF1L5, SNRNP70, HNRNPUL2, etc.] of the host interact with viral proteins (Fig. 3; see also Table S2). Although previous studies showed that the role of interactions between spliceosome components and host proteins was substantial in SARS-CoV infections and other coronavirus infections (43, 44), with this study, we confirmed their role in novel coronavirus (SARS-CoV-2) infection, too. We found that orf1ab interacts with SRSF protein kinase 1 (SRPK1), which plays a central role in splicing and is known to phosphorylate serine/arginine-rich splicing factor 1 (SRSF1) (45). SRSF1 regulates the accuracy of splicing and also that of alternative splicing. PUF60, another component of the spliceosome, was found to interact with orf1ab, which plays a vital role in pre-mRNA splicing and 3' end processing (46). It promotes the splicing of introns in a cooperative manner with another splicing factor, U2AF2 (46). The U4/U6-U5 tri-snRNP complex is involved in spliceosome assembly, and one component, U4/U6 small nuclear ribonucleoprotein Prp3 (PRPF3) (47), was found to interact with orf1ab. The results of our analysis emphasized that SARS-CoV-2 manipulates spliceosome machinery during infection; hence, targeting splicing might affect viral replication. Recently, Bojkova et al. also showed that addition of spliceosome inhibitor Pladeinolide-B into SARS-CoV-2-infected human Caco-2 cells significantly inhibited viral replication (48). Thus, targeting the splicing could be another prospective drug discovery.

Further, there are reports which suggest that targeting notch signaling could be a way to prevent SARS-CoV-2 infection, as notch-mediated downregulation of furin (a host protease) levels was found to interfere with entrance of the virus into the host cell (49). Interestingly, we showed that nsp8 interacts with POGLUT2, POGLUT3 and POFUT1, which regulate the transport of notch1 and notch3 to the plasma membrane and fucosylation of notch1 protein, thereby modulating the notch signaling (50, 51).

Membrane (M) and envelope (E) proteins are structural proteins which are crucial for viral assembly and pathogenesis. Our interactome studies showed that M and E proteins of SARS-CoV-2 interact with multiple host proteins. Membrane protein and nsp8 have been found to interact with protein transport protein Sec16A (Fig. S2; see also Table S2), which mediates endoplasmic reticulum (ER) membrane insertion of SARS-CoV-2 proteins, critical for cotranslational entry into secretory pathways (52). Further, the data also showed that M and nsp8 proteins interact with host protein responsible in cristae formation, which suggests the possibility of mitochondrial fusion-mediated downregulation of host cell interferon gamma responses similar to that seen with SARS-CoV (53). A similar form of mitochondrial dysfunction in SARS-CoV-2 infected cells was also reported previously (54).

Furthermore, we showed that the E-protein interacts with bromodomain proteins (BRD4), which is in agreement with a previous study in which the researchers showed that SARS-CoV-2 envelop interacted with bromodomain proteins BRD2 and BRD4 to regulate gene transcription (52). Unfortunately, the later stages of COVID-19 infection result in development of a hypoxic condition which leads to progression of ARDS (acute

respiratory distress syndrome) and toxic encephalopathy (55). Our interactome study revealed that ORF7a potentially regulates such hypoxic conditions by interacting with host proteins. Studies have shown that kidney involvement is frequent in COVID-19 patients and even acute kidney injury is common in critically ill patients (56). Our interactome analysis showed an interaction of ORF7a and nsp8 with NPEPPS, a puromycin-sensitive amino peptidase, which is commonly used as a biomarker against damaged kidneys (57). Our results also indicated the interaction of ORF7a and nsp8 with multiple proteasome-related proteins such as PSMD6, PSMD7, PSMD2, and PSMD13, which is in agreement with a previous study in which the authors revealed the interactions by overexpressing the SARS-CoV-2 gene in HEK293 cells (58). Taking the results together, these proteins regulate multiple cellular pathways, such as interleukin pathways, which may be a reason for the high plasma concentrations of cytokines such as IL-2, IL-7, IL-10, and IL-6 in critically ill COVID-19 patients (59), and indeed, the cytokine storm is a major cause of inflammatory cascades during COVID-19 infection (60).

It is well known that spike (S) glycoproteins facilitate the entry of SARS-CoV-2 into host cells by binding with cellular receptor angiotensin-converting enzyme 2 (ACE2) (2, 61, 62). This protein has been the most highly studied protein so far, and our interactome results are in consensus with the previous reports. Moreover, the studies on HEK293T cells have shown that the S-proteins of SARS-CoV-2 were entirely processed at the S1 and S2 sites during biosynthesis in the Golgi compartment (2). We also found strong interactions among S-protein, ZDHHC5, and GOLGA7 (Fig. S2), where the latter two formed a palmitoyltransferase complex involved in palmitoylation process and in transport from the Golgi compartment to the cell surface. S-proteins also interact with LDHB (lactate dehydrogenase B), a hallmark of inflammation during COVID-19 infection. Recent studies of critically ill COVID-19 patients showed high levels of lactate dehydrogenase (63, 64). It is presumed that increased LDH levels lead to the complexity of disease manifestations accompanied by respiratory failure due to prolonged severe inflammatory responses. Being the linchpin in the process, it is considered to represent a possible biomarker for diagnostic testing to identify persons infected with SARS-CoV-2 (63–65).

Taking the results together, this study integrated the host-pathogen protein interaction network using data sets generated from different studies (35, 36, 52) and showed that SARS-CoV-2 viral proteins discretely manipulate the host system for their own survival and pathogenicity.

Direction of selection of SARS-CoV-2 genes. Our analysis revealed that ORF8 (121 amino acids [aa]) ($dN/dS = 35.8$), along with ORF3a (275 bp), ($dN/dS = 8.95$) showed the highest dN/dS values among the nine ORFs and that ORF8 and ORF3a thus have much greater numbers of nonsynonymous substitutions than the synonymous substitution (Fig. 4). dN/dS values that are much higher than 1 are indicative of a strongly divergent lineage (66). Thus, both of these proteins are evolving under conditions of high selection pressure and represent ORFs that are highly divergent across strains. Two other proteins, ORF1ab polyprotein ($dN/dS = 0.996$ and 0.575) and S-protein ($dN/dS = 0.88$) might confer a selective advantage with respect to host challenges and survival. The dN/dS rate values of nearly 1 and greater than 1 suggest that the strains are coping well with the challenges that they encounter, i.e., with immune responses and the inhibitory environment of host cells (67). The other gene clusters, namely, those corresponding to M-protein and orf1a polyprotein, did not possess at least three unique sequences, which is a condition that is necessary for the analysis; hence, they should be similar across the strains. The two ORF1ab polyprotein genes that code for protein translation and posttranslation modification were found to have evolved to perform active translation and to facilitate the multiplication and growth of virus inside the host. Similarly, the S-protein which helps in the entry of virus into the host cells by facilitating penetration of the cell membrane was found to be involved in the acceleration toward positive selection, confirming the ability of the enzyme to initiate the

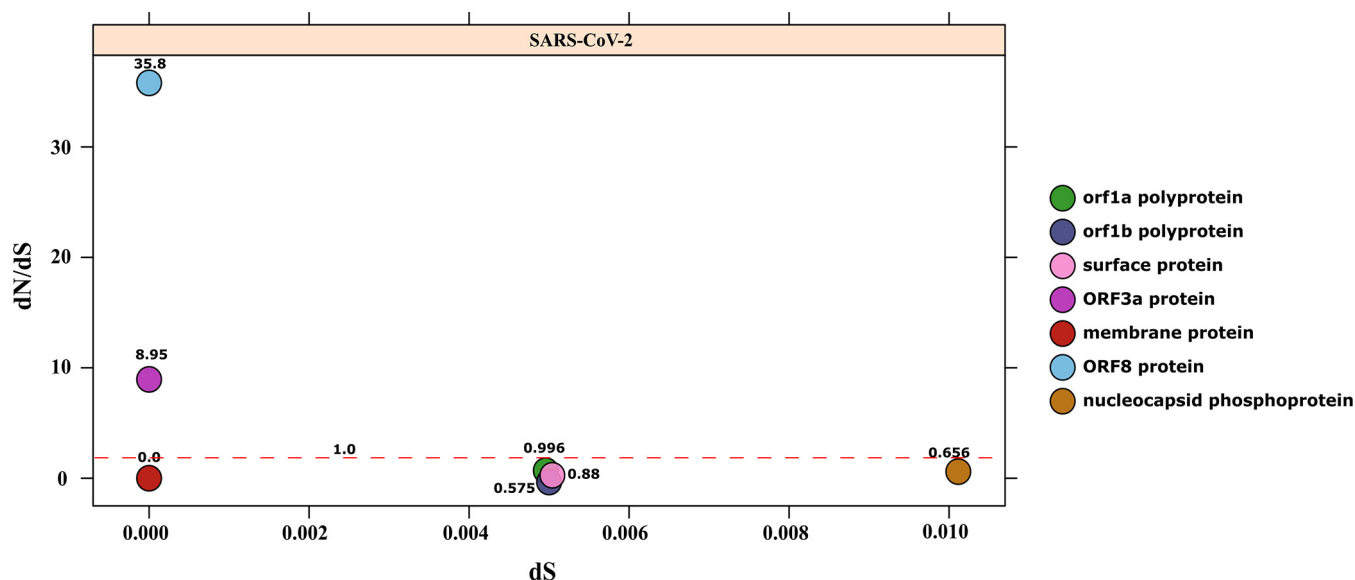


FIG 4 Estimation of purifying natural selection pressure in nine coding sequences of SARS-CoV-2. dN/dS values are plotted as a function of dS .

infection. Another positively diversifying N protein gene codes for nucleocapsid formation, which protects the genetic material of virus from host immune responses such as cellular protease activity. Overall, the data indicate that the growth-related and multiplication-related genes are evolving at a high rate. The results revealing other proteins with dN/dS values equal to zero suggest a conserved repertoire of genes.

Conclusions. As COVID-19 continues to impact virtually all human lives worldwide due to its extremely contagious nature, it has spiked the interest of scientific community all over the world in better understanding of the pathogenesis of the novel SARS-CoV-2 virus. In this study, analyses were performed on the genomes of the novel SARS-CoV-2 isolates recently reported from different countries to understand the viral pathogenesis. We observed no direct pattern of transmission of the novel SARS-CoV-2 in the neighboring countries through our analyses of the phylogenomic relatedness of geographical isolates. Isolates collected from same locations, for instance, isolates from the United States and isolates from China, were phylogenetically distant. Thus, there appears to be a mosaic pattern of transmission indicating that it represents the result of travel of infected humans among different countries. As COVID-19 transitioned from epidemic to pandemic within a short time, such a result does not look surprising on the basis of the genome structures of the viral isolates. The genomes of six isolates, specifically, from the United States, were found to harbor unique amino acid SNPs and showed amino acid substitutions in ORF1b protein and S-protein, while one of them also harbored an amino acid addition. The ORF1ab polyprotein and S-protein were also found to have dN/dS values approaching 1 and thus might confer a selective advantage enabling the virus to evade host response mechanisms. Thus, these proteins are rapidly evolving and are involved in viral replication. Therefore, these mutations cannot be neglected while developing a vaccine. Further, the SARS-CoV-2–human interactome revealed that its pathogenicity is mediated by a surge in proinflammatory cytokine levels. It is predicted that the major mechanism of immune pathogenicity in SARS-CoV-2 includes alteration of the host cell environment by disintegration of signal transduction pathways and immunity evasion by several protection mechanisms. Our results provide insights into COVID-19 genomes and interactomes which may be useful to understand the impact of virus on human health.

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 19 March 2020 from the NCBI database, and, based on quality assessment, two genomes with multiple Ns were removed from the study. Further, the genomes were annotated using Prokka (68). 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 (GenBank accession number [MT121215](#)), and ORFs were predicted against the formatted database using Prokka (-gcode 1) (68). Further, the GC content information was generated using the QUAST standalone tool (69).

Phylogenetic analysis. To infer the phylogeny, the core gene alignment was generated using MAFFT (70) within the Roary Package (71). Further, the phylogeny was inferred using the maximum likelihood method and the Tamura-Nei model (72) at the 1,000-bootstrap level in MEGAX (73) and was visualized in interactive Tree of Life (iTOL) (74) and GrapeTree (75).

We also constructed the gene phylogeny using nucleocapsid and spike proteins of 95 SARS-CoV-2 strains along with members of the *Coronaviridae* family present in synanthropic animals. The multiple-sequence alignment was performed using the MUSCLE (76) aligner, and the phylogeny was constructed using the neighbor joining method and MEGAX software (73). Further, the trees were visualized in interactive Tree of Life (iTOL) (74).

Single nucleotide polymorphism and structural analysis. To determine each single nucleotide polymorphism (SNP), whole-genome alignments were made using the libMUSCLE aligner. For this, we used annotated GenBank of SARS-CoV-2/SH01/human/2020/CHN (GenBank accession no. MT121215) as the reference in the parsnp tool of Harvest suite (77). As only genomes within a specified MUMI distance threshold are recruited, we used option -c to force inclusion of all the strains. For the output, it produced a core-genome alignment, variant calls, and a phylogeny based on single nucleotide polymorphisms. The SNPs were further visualized in GIngr, a dynamic visual platform (77). Further, the tree was visualized in interactive Tree of Life (iTOL) (74).

The 3D structures for nsp13 and nsp16 were predicted using the amino acid sequence from the wild-type reference (MT121215) and mutants MT163719 and MT188341, respectively. The 3D structures were predicted using the ExPASy Web interface tool, the SwissModel server homology modeling pipeline (78). The structures were subjected to energy minimization using UCSF Chimera v.1.13.1 software (79). The predicted models were subjected to validation using a Ramachandran plot of the proteins and the structure assessment tool in SWISS-MODEL. The structures were compared to assess the effect of mutation, and the levels of stability were compared using the HOPE (80), iPBA (81), and I-mutant v3.0 (82) servers. The secondary protein structure motifs were created by PDB-sum using v3.0 of Gail Hutchinson's PROMOTIF program (83, 84) and compared for changes. The active sites and motifs were predicted for both models by the use of the PROSITE online server of ExPasy (<https://prosite.expasy.org/>). The 3D structures of the predicted models were visualized using UCSF Chimera software v.1.13.1 (79).

SARS-CoV-2 protein annotation and host-pathogen interactions. The SARS-CoV-2/SH01/human/2020/CHN virus genome having accession no. [MT121215.1](#) was used for protein-protein network analysis. The data corresponding to the interaction of SARS-COV-2 and human host were extracted from the IntAct database with high-confidence values ranging between 0.74 and 0.97 (MIscore) (34–36). The network was visualized using Cytoscape v3.7.2 (85) and was analyzed to gain insights into the network topology using Network Analyzer, a plugin of Cytoscape. Network topology results gave an overview of network topological features, including diameter, degree distribution, shortest path distribution, and clustering coefficient of the interaction network. The network was investigated with the power law equation, and degree-based analyses were carried out. Further, the human proteins interacting with individual viral proteins were subjected to functional annotation. Gene ontology (GO) analysis was performed using ClueGo (86), selecting the Kyoto Encyclopedia of Genes and Genomes (KEGG) (87, 88), Gene Ontology—biological function database, and Reactome Pathways (89) databases. The ClueGo parameters were as follows: Go Term Fusion selected; pathways or terms of the associated genes, ranked based on the *P* value corrected with Bonferroni stepdown (*P* values of <0.05); GO tree interval, all levels; GO term minimum number of genes, 3; threshold, 4% of genes per pathway; kappa score, 0.42. The GO term value was reduced to 1 gene for the S and E proteins. Gene ontology terms are presented as nodes and clustered together based on the similarity of genes corresponding to each term or pathway.

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 (32) and further processed for removal of stop codons using HyPhy v2.2.4 (90). The single-likelihood ancestor counting (SLAC) method was used in Datamonkey v2.0 (91) (<http://www.datamonkey.org/slac>) to calculate the *dN/dS* value for each orthologous gene cluster. The *dN/dS* values were plotted in R (R Development Core Team, 2015).

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

FIG S1, PDF file, 0.04 MB.

FIG S2, EPS file, 0.7 MB.

TABLE S1, DOCX file, 0.01 MB.

TABLE S2, XLSX file, 0.02 MB.

ACKNOWLEDGMENTS

R. Kumar acknowledges Magadh University, Bodh Gaya, for providing support. R. Lal and U. Sood also acknowledge The National Academy of Sciences, India, for support

under the NASI-Senior Scientist Platinum Jubilee Fellowship Scheme. N. Singhvi, S. Nagar, and C. Talwar acknowledge the Council of Scientific and Industrial Research (CSIR), New Delhi, for doctoral fellowships. H. Verma and C. D. Rawat thank Ramjas College, University of Delhi, Delhi, for providing support. V. Gupta and M. Singh acknowledge Phixgen Pvt. Ltd. for research fellowships. P. Hira thanks Maitreyi College, University of Delhi, Delhi, for providing support. Y. Singh acknowledges support for a J. C. Bose fellowship (SERB).

R. Lal, R. Kumar, H. Verma, V. Gupta, and U. Sood conceived and designed the study. R. Kumar, H. Verma, N. Singhvi, U. Sood, V. Gupta, M. Singh, S. Nagar, and P. Hira executed the analysis and prepared figures. R. Kumar, H. Verma, R. Kumari, N. Singhvi, U. Sood, V. Gupta, M. Singh, S. Nagar, P. Hira, C. Talwar, N. Nayyar, S. Anand, C. D. Rawat, and M. Verma wrote the manuscript with contributions from all of us. Y. Singh and R. K. Negi provided guidance from time to time.

We declare that we have no conflict of interest.

REFERENCES

- Wang Q, Zhang Y, Wu L, Niu S, Song C, Zhang Z, Lu G, Qiao C, Hu Y, Yuen KY, Wang Q, Zhou H, Yan J, Qi J. 2020. Structural and functional basis of SARS-CoV-2 entry by using human ACE2. *Cell* 181:894–904.e9. <https://doi.org/10.1016/j.cell.2020.03.045>.
- Walls AC, Park YJ, Tortorici MA, Wall A, McGuire AT, Veesler D. 2020. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell* 181:281–292.e6. <https://doi.org/10.1016/j.cell.2020.02.058>.
- Wu A, Peng Y, Huang B, Ding X, Wang X, Niu P, Meng J, Zhu Z, Zhang Z, Wang J, Sheng J, Quan L, Xia Z, Tan W, Cheng G, Jiang T. 2020. Genome composition and divergence of the novel coronavirus (2019-nCoV) originating in China. *Cell Host Microbe* 27:325–328. <https://doi.org/10.1016/j.chom.2020.02.001>.
- Tyrrell DA, Bynoe ML. 1966. Cultivation of viruses from a high proportion of patients with colds. *Lancet* i:76–77. [https://doi.org/10.1016/s0140-6736\(66\)92364-6](https://doi.org/10.1016/s0140-6736(66)92364-6).
- Woo PC, Lau SK, Lam CS, Lau CC, Tsang AK, Lau JH, Bai R, Teng JL, Tsang CC, Wang M, Zheng BJ, Chan KH, Yuen KY. 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. <https://doi.org/10.1128/JVI.06540-11>.
- Li F. 2016. Structure, function, and evolution of coronavirus spike proteins. *Annu Rev Virol* 3:237–261. <https://doi.org/10.1146/annurev-virology-110615-042301>.
- Tang Q, Song Y, Shi M, Cheng Y, Zhang W, Xia XQ. 2015. Inferring the hosts of coronavirus using dual statistical models based on nucleotide composition. *Sci Rep* 5:17155. <https://doi.org/10.1038/srep17155>.
- Fehr AR, Perlman S. 2015. Coronaviruses: an overview of their replication and pathogenesis. *Methods Mol Biol* 1282:1–23. https://doi.org/10.1007/978-1-4939-2438-7_1.
- Lv L, Li G, Chen J, Liang X, Li Y. 2020. Comparative genomic analysis revealed specific mutation pattern between human coronavirus SARS-CoV-2 and Bat-SARSr-CoV RaTG13. *bioRxiv* <https://doi.org/10.1101/2020.02.27.969006>.
- Lau SK, Woo PC, Li KS, Huang Y, Tsoi HW, Wong BH, Wong SS, Leung SY, Chan KH, Yuen KY. 2005. Severe acute respiratory syndrome coronavirus-like virus in Chinese horseshoe bats. *Proc Natl Acad Sci U S A* 102:14040–14045. <https://doi.org/10.1073/pnas.0506735102>.
- Meyer B, Muller MA, Corman VM, Reusken CB, Ritz D, Godeke GJ, Lattwein E, Kallies S, Siemens A, van Beek J, Drexler JF, Muth D, Bosch BJ, Wernery U, Koopmans MP, Wernery R, Drosten C. 2014. Antibodies against MERS coronavirus in dromedary camels, United Arab Emirates, 2003 and 2013. *Emerg Infect Dis* 20:552–559. <https://doi.org/10.3201/eid2004.131746>.
- Zhang C, Zheng W, Huang X, Bell EW, Zhou X, Zhang Y. 2020. Protein structure and sequence reanalysis of 2019-nCoV genome refutes snakes as its intermediate host and the unique similarity between its spike protein insertions and HIV-1. *J Proteome Res* 19:1351–1360. <https://doi.org/10.1021/acs.jproteome.0c00129>.
- Bajaj A, Purohit HJ. 2020. Understanding SARS-CoV-2: genetic diversity, transmission and cure in human. *Indian J Microbiol* 60:398–401. <https://doi.org/10.1007/s12088-020-00869-4>.
- Neuman BW, Adair BD, Yoshioka C, Quispe JD, Orca G, Kuhn P, Milligan RA, Yeager M, Buchmeier MJ. 2006. Supramolecular architecture of severe acute respiratory syndrome coronavirus revealed by electron cryomicroscopy. *J Virol* 80:7918–7928. <https://doi.org/10.1128/JVI.00645-06>.
- Barcena M, Oostergetel GT, Bartelink W, Faas FG, Verkleij A, Rottier PJ, Koster AJ, Bosch BJ. 2009. Cryo-electron tomography of mouse hepatitis virus: insights into the structure of the coronavirus. *Proc Natl Acad Sci U S A* 106:582–587. <https://doi.org/10.1073/pnas.0805270106>.
- Chen Y, Liu Q, Guo D. 2020. Emerging coronaviruses: genome structure, replication, and pathogenesis. *J Med Virol* 92:418–423. <https://doi.org/10.1002/jmv.25681>.
- Collins AR, Knobler RL, Powell H, Buchmeier MJ. 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. [https://doi.org/10.1016/0042-6822\(82\)90095-2](https://doi.org/10.1016/0042-6822(82)90095-2).
- Neuman BW, Kiss G, Kunding AH, Bhella D, Baksh MF, Connelly S, Droese B, Klaus JP, Makino S, Sawicki SG, Siddell SG, Stamou DG, Wilson IA, Kuhn P, Buchmeier MJ. 2011. A structural analysis of M protein in coronavirus assembly and morphology. *J Struct Biol* 174:11–22. <https://doi.org/10.1016/j.jsb.2010.11.021>.
- Ruch TR, Machamer CE. 2012. The coronavirus E protein: assembly and beyond. *Viruses* 4:363–382. <https://doi.org/10.3390/v4030363>.
- McBride R, van Zyl M, Fielding BC. 2014. The coronavirus nucleocapsid is a multifunctional protein. *Viruses* 6:2991–3018. <https://doi.org/10.3390/v6082991>.
- Yu X, Yang R. 2020. COVID-19 transmission through asymptomatic carriers is a challenge to containment. *Influenza Other Respir Viruses* 14:474–475. <https://doi.org/10.1111/irv.12743>.
- Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, Wang W, Song H, Huang B, Zhu N, Bi Y, Ma X, Zhan F, Wang L, Hu T, Zhou H, Hu Z, Zhou W, Zhao L, Chen J, Meng Y, Wang J, Lin Y, Yuan J, Xie Z, Ma J, Liu WJ, Wang D, Xu W, Holmes EC, Gao GF, Wu G, Chen W, Shi W, Tan W. 2020. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet* 395:565–574. [https://doi.org/10.1016/S0140-6736\(20\)30251-8](https://doi.org/10.1016/S0140-6736(20)30251-8).
- Sah R, Rodriguez-Morales AJ, Jha R, Chu DKW, Gu H, Peiris M, Bastola A, Lal BK, Ojha HC, Rabaan AA, Zambrano LI, Costello A, Morita K, Pandey BD, Poon L. 2020. Complete genome sequence of a 2019 novel coronavirus (SARS-CoV-2) strain isolated in Nepal. *Microbiol Resour Announc* 9:e00169-20. <https://doi.org/10.1128/MRA.00169-20>.
- Ren LL, Wang YM, Wu ZQ, Xiang ZC, Guo L, Xu T, Jiang YZ, Xiong Y, Li YJ, Li XW, Li H, Fan GH, Gu XY, Xiao Y, Gao H, Xu JY, Yang F, Wang XM, Wu C, Chen L, Liu YW, Liu B, Yang J, Wang XR, Dong J, Li L, Huang CL, Zhao JP, Hu Y, Cheng ZS, Liu LL, Qian ZH, Qin C, Jin Q, Cao B, Wang JW. 2020. Identification of a novel coronavirus causing severe pneumonia in human: a descriptive study. *Chin Med J (Engl)* 133:1015–1024. <https://doi.org/10.1097/CM9.0000000000000722>.

25. Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. 2020. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nat Microbiol* 5:536–544. <https://doi.org/10.1038/s41564-020-0695-z>.
26. Yadav PD, Potdar VA, Choudhary ML, Nyayanit DA, Agrawal M, Jadhav SM, Majumdar TD, Shete-Aich A, Basu A, Abraham P, Cherian SS. 2020. Full-genome sequences of the first two SARS-CoV-2 viruses from India. *Indian J Med Res* 151:200–209. https://doi.org/10.4103/ijmr.IJMR_663_20.
27. Tilocca B, Soggiu A, Sanguinetti M, Musella V, Britti D, Bonizzi L, Urbani A, Roncada P. 2020. Comparative computational analysis of SARS-CoV-2 nucleocapsid protein epitopes in taxonomically related coronaviruses. *Microbes Infect* 22:188–194. <https://doi.org/10.1016/j.micinf.2020.04.002>.
28. Tilocca B, Soggiu A, Musella V, Britti D, Sanguinetti M, Urbani A, Roncada P. 2020. Molecular basis of COVID-19 relationships in different species: a one health perspective. *Microbes Infect* 22:218–220. <https://doi.org/10.1016/j.micinf.2020.03.002>.
29. Snijder EJ, Decroly E, Ziebuhr J. 2016. The nonstructural proteins directing coronavirus RNA synthesis and processing. *Adv Virus Res* 96:59–126. <https://doi.org/10.1016/bs.aivir.2016.08.008>.
30. Jang KJ, Jeong S, Kang DY, Sp N, Yang YM, Kim DE. 2020. A high ATP concentration enhances the cooperative translocation of the SARS coronavirus helicase nsp13 in the unwinding of duplex RNA. *Sci Rep* 10:4481. <https://doi.org/10.1038/s41598-020-61432-1>.
31. Becares M, Pascual-Iglesias A, Nogales A, Sola I, Enjuanes L, Zuñiga S. 2016. Mutagenesis of coronavirus nsp14 reveals its potential role in modulation of the innate immune response. *J Virol* 90:5399–5414. <https://doi.org/10.1128/JVI.03259-15>.
32. Athmer J, Fehr AR, Grunewald M, Smith EC, Denison MR, Perlman S. 2017. In situ tagged nsp15 reveals interactions with coronavirus replication/Transcription complex-associated proteins. *mBio* 8:e02320-16. <https://doi.org/10.1128/mBio.02320-16>.
33. von Grotthuss M, Wyrwicz LS, Rychlewski L. 2003. mRNA cap-1 methyltransferase in the SARS genome. *Cell* 113:701–702. [https://doi.org/10.1016/S0092-8674\(03\)00424-0](https://doi.org/10.1016/S0092-8674(03)00424-0).
34. Hermjakob H, Montecchi-Palazzi L, Lewington C, Mudali S, Kerrien S, Orchard S, Vingron M, Roehert B, Roepstorff P, Valencia A, Margalit H, Armstrong J, Bairoch A, Cesareni G, Sherman D, Apweiler R. 2004. IntAct: an open source molecular interaction database. *Nucleic Acids Res* 32:D452–D455. <https://doi.org/10.1093/nar/gkh052>.
35. Gao Y, Yan L, Huang Y, Liu F, Zhao Y, Cao L, Wang T, Sun Q, Ming Z, Zhang L, Ge J, Zheng L, Zhang Y, Wang H, Zhu Y, Zhu C, Hu T, Hua T, Zhang B, Yang X, Li J, Yang H, Liu Z, Xu W, Guddat LW, Wang Q, Lou Z, Rao Z. 2020. Structure of the RNA-dependent RNA polymerase from COVID-19 virus. *Science* 368:779–782. <https://doi.org/10.1126/science.abb7498>.
36. Xia S, Liu M, Wang C, Xu W, Lan Q, Feng S, Qi F, Bao L, Du L, Liu S, Qin C, Sun F, Shi Z, Zhu Y, Jiang S, Lu L. 2020. Inhibition of SARS-CoV-2 (previously 2019-nCoV) infection by a highly potent pan-coronavirus fusion inhibitor targeting its spike protein that harbors a high capacity to mediate membrane fusion. *Cell Res* 30:343–355. <https://doi.org/10.1038/s41422-020-0305-x>.
37. Khailany RA, Safdar M, Ozaslan M. 2020. Genomic characterization of a novel SARS-CoV-2. *Gene Rep* 113:100682. <https://doi.org/10.1016/j.genrep.2020.100682>.
38. Goncalves A, Burckstummer T, Dixit E, Scheicher R, Gorna MW, Karayel E, Sugar C, Stukalov A, Berg T, Kralovics R, Planyavsky M, Bennett KL, Colinge J, Superti-Furga G. 2011. Functional dissection of the TBK1 molecular network. *PLoS One* 6:e23971. <https://doi.org/10.1371/journal.pone.0023971>.
39. Wang C, Chen T, Zhang J, Yang M, Li N, Xu X, Cao X. 2009. The E3 ubiquitin ligase Nrdp1 ‘preferentially’ promotes TLR-mediated production of type I interferon. *Nat Immunol* 10:744–752. <https://doi.org/10.1038/ni.1742>.
40. Tetsuka T, Uranishi H, Imai H, Ono T, Sonta S, Takahashi N, Asamitsu K, Okamoto T. 2000. Inhibition of nuclear factor-kappaB-mediated transcription by association with the amino-terminal enhancer of split, a Groucho-related protein lacking WD40 repeats. *J Biol Chem* 275:4383–4390. <https://doi.org/10.1074/jbc.275.6.4383>.
41. Schett G, Sticherling M, Neurath MF. 2020. COVID-19: risk for cytokine targeting in chronic inflammatory diseases? *Nat Rev Immunol* 20:271–272. <https://doi.org/10.1038/s41577-020-0312-7>.
42. Darzacq X, Jady BE, Verheggen C, Kiss AM, Bertrand E, Kiss T. 2002. Cajal body-specific small nuclear RNAs: a novel class of 2'-O-methylation and pseudouridylation guide RNAs. *EMBO J* 21:2746–2756. <https://doi.org/10.1093/emboj/21.11.2746>.
43. Neuman BW, Joseph JS, Saikatendu KS, Serrano P, Chatterjee A, Johnson MA, Liao L, Klaus JP, Yates JR, III, Wuthrich K, Stevens RC, Buchmeier MJ, Kuhn P. 2008. Proteomics analysis unravels the functional repertoire of coronavirus nonstructural protein 3. *J Virol* 82:5279–5294. <https://doi.org/10.1128/JVI.02631-07>.
44. Jourdan SS, Osorio F, Hiscox JA. 2012. An interactome map of the nucleocapsid protein from a highly pathogenic North American porcine reproductive and respiratory syndrome virus strain generated using SILAC-based quantitative proteomics. *Proteomics* 12:1015–1023. <https://doi.org/10.1002/pmic.201100469>.
45. Ghosh G, Adams JA. 2011. Phosphorylation mechanism and structure of serine-arginine protein kinases. *FEBS J* 278:587–597. <https://doi.org/10.1111/j.1742-4658.2010.07992.x>.
46. Page-McCaw PS, Amonlirdviman K, Sharp PA. 1999. PUF60: a novel U2AF65-related splicing activity. *RNA* 5:1548–1560. <https://doi.org/10.1017/s1355838299991938>.
47. Agafonov DE, Dybkov O, Hofele RV, Liu WT, Urlaub H, Lührmann R, Stark H. 2016. Molecular architecture of the human U4/U6.U5 tri-snRNP. *Science* 351:1416–1420. <https://doi.org/10.1126/science.122085>.
48. Břkova D, Klann K, Koch B, Wiedera M, Krause D, Cinatl C, Sandra C, Cinatl J, Münch C. 2020. SARS-CoV-2 infected host cell proteomics reveal potential therapy targets. Preprint <https://doi.org/10.21203/rs.3.rs-17218/v1>.
49. Rizzo P, Vicieli Dalla Sega F, Fortini F, Marracino L, Rapezzi C, Ferrari R. 2020. COVID-19 in the heart and the lungs: could we “Notch” the inflammatory storm? *Basic Res Cardiol* 115:31. <https://doi.org/10.1007/s00395-020-0791-5>.
50. Takeuchi H, Schneider M, Williamson DB, Ito A, Takeuchi M, Handford PA, Hattiwanger RS. 2018. Two novel protein O-glycosyltransferases that modify sites distinct from POGlut1 and affect Notch trafficking and signaling. *Proc Natl Acad Sci U S A* 115:E8395–E8402. <https://doi.org/10.1073/pnas.1804005115>.
51. McMillan BJ, Zimmerman B, Egan ED, Lofgren M, Xu X, Hesser A, Blacklow SC. 2017. Structure of human POFUT1, its requirement in ligand-independent oncogenic Notch signaling, and functional effects of Dowling-Degos mutations. *Glycobiology* 27:777–786. <https://doi.org/10.1093/glycob/cwx020>.
52. Gordon DE, Jang GM, Bouhaddou M, Xu J, Obernier K, White KM, O’Meara MJ, Rezelj VV, Guo JZ, Swaney DL, Tummino TA, Hüttenhain R, Kaake RM, Richards AL, Tutuncuoğlu B, Fousnard H, Batra J, Haas K, Modak M, Kim M, Haas P, Polacco BJ, Braberg H, Fabius JM, Eckhardt M, Soucheyra M, Bennett MJ, Cakir M, McGregor MJ, Li Q, Meyer B, Roesch F, Vallet T, Mac Kain A, Miorin L, Moreno E, Naing ZZC, Zhou Y, Peng S, Shi Y, Zhang Z, Shen W, Kirby IT, Melnyk JE, Chorba JS, Lou K, Dai SA, Barrio-Hernandez I, Memon D, Hernandez-Armenta C, et al. 2020. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* 583:459–468. <https://doi.org/10.1038/s41586-020-2286-9>.
53. Shi CS, Qi HY, Boularan C, Huang NN, Abu-Asab M, Shelhamer JH, Kehrl JH. 2014. SARS-coronavirus open reading frame-9b suppresses innate immunity by targeting mitochondria and the MAVS/TRAF3/TRAF6 signalosome. *J Immunol* 193:3080–3089. <https://doi.org/10.4049/jimmunol.1303196>.
54. Singh K, Chen Y-C, Judy JT, Seifuddin F, Tunc I, Pirooznia M. 2020. Network analysis and transcriptome profiling identify autophagic and mitochondrial dysfunctions in SARS-CoV-2 infection. *bioRxiv* <https://doi.org/10.1101/2020.05.13.092536>.
55. Mason RJ. 2020. Pathogenesis of COVID-19 from a cell biology perspective. *Eur Respir J* 55:2000607. <https://doi.org/10.1183/13993003.00607-2020>.
56. Ronco C, Reis T, Husain-Syed F. 2020. Management of acute kidney injury in patients with COVID-19. *Lancet Respir Med* 8:738–742. [https://doi.org/10.1016/S2213-2600\(20\)30229-0](https://doi.org/10.1016/S2213-2600(20)30229-0).
57. Holdt B, Peters E, Nagel HR, Steiner M. 2008. An automated assay of urinary alanine aminopeptidase activity. *Clin Chem Lab Med* 46:537–540. <https://doi.org/10.1515/CCLM.2008.103>.
58. Li J, Guo M, Tian X, Liu C, Wang X, Yang X, Wu P, Xiao Z, Qu Y, Yin Y, Fu J, Zhu Z, Liu Z, Peng C, Zhu T, Liang Q. 2020. Virus-host interactome and proteomic survey of PMBCs from COVID-19 patients reveal potential virulence factors influencing SARS-CoV-2 pathogenesis. *bioRxiv* <https://doi.org/10.1101/2020.03.31.019216>.

59. Zhong J, Tang J, Ye C, Dong L. 2020. The immunology of COVID-19: is immune modulation an option for treatment? *Lancet Rheumatol* 2:e428–e436. [https://doi.org/10.1016/S2665-9913\(20\)30120-X](https://doi.org/10.1016/S2665-9913(20)30120-X).
60. Coperchini F, Chiovato L, Croce L, Magri F, Rotondi M. 2020. The cytokine storm in COVID-19: an overview of the involvement of the chemokine/chemokine-receptor system. *Cytokine Growth Factor Rev* 53:25–32. <https://doi.org/10.1016/j.cytogfr.2020.05.003>.
61. Hoffmann M, Kleine-Weber H, Schroeder S, Krüger N, Herrler T, Erichsen S, Schiergens TS, Herrler G, Wu N-H, Nitsche A, Müller MA, Drosten C, Pöhlmann S. 2020. SARS-CoV-2 cell entry depends on ACE2 and TM-PRSS2 and is blocked by a clinically proven protease inhibitor. *Cell* 181:271–280.e8. <https://doi.org/10.1016/j.cell.2020.02.052>.
62. Yan R, Zhang Y, Li Y, Xia L, Guo Y, Zhou Q. 2020. Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. *Science* 367:1444–1448. <https://doi.org/10.1126/science.abb2762>.
63. Zhou F, Yu T, Du R, Fan G, Liu Y, Liu Z, Xiang J, Wang Y, Song B, Gu X, Guan L, Wei Y, Li H, Wu X, Xu J, Tu S, Zhang Y, Chen H, Cao B. 2020. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *Lancet* 395:1054–1062. [https://doi.org/10.1016/S0140-6736\(20\)30566-3](https://doi.org/10.1016/S0140-6736(20)30566-3).
64. Huang X, Wei F, Yang Z, Li M, Liu L, Chen K. 2020. Lactose dehydrogenase in patients with severe COVID-19: a meta-analysis of retrospective study. *Prehosp Disaster Med* <https://doi.org/10.1017/S1049023X20000576>.
65. Cheng MP, Papenburg J, Desjardins M, Kanjilal S, Quach C, Libman M, Dittrich S, Yansouni CP. 2020. Diagnostic testing for severe acute respiratory syndrome-related coronavirus-2: a narrative review. *Ann Intern Med* <https://doi.org/10.7326/M20-1301>.
66. Kryazhimskiy S, Plotkin JB. 2008. The population genetics of *dN/dS*. *PLoS Genet* 4:e1000304. <https://doi.org/10.1371/journal.pgen.1000304>.
67. Kosakovsky Pond SL, Frost SD. 2005. Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Mol Biol Evol* 22:1208–1222. <https://doi.org/10.1093/molbev/msi105>.
68. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30:2068–2069. <https://doi.org/10.1093/bioinformatics/btu153>.
69. Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 29:1072–1075. <https://doi.org/10.1093/bioinformatics/btt086>.
70. Nakamura T, Yamada KD, Tomii K, Katoh K. 2018. Parallelization of MAFFT for large-scale multiple sequence alignments. *Bioinformatics* 34:2490–2492. <https://doi.org/10.1093/bioinformatics/bty121>.
71. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, Fookes M, Falush D, Keane JA, Parkhill J. 2015. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 31:3691–3693. <https://doi.org/10.1093/bioinformatics/btv421>.
72. Tamura K, 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. <https://doi.org/10.1093/oxfordjournals.molbev.a040023>.
73. Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for bigger datasets. *Mol Biol Evol* 33:1870–1874. <https://doi.org/10.1093/molbev/msw054>.
74. Letunic I, 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–W245. <https://doi.org/10.1093/nar/gkw290>.
75. Zhou Z, Alikhan NF, Sergeant MJ, Luhmann N, Vaz C, Francisco AP, Carrico JA, Achtman M. 2018. GrapeTree: visualization of core genomic relationships among 100,000 bacterial pathogens. *Genome Res* 28:1395–1404. <https://doi.org/10.1101/gr.232397.117>.
76. Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792–1797. <https://doi.org/10.1093/nar/gkh340>.
77. Treangen TJ, Ondov BD, Koren S, Phillippy AM. 2014. The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biol* 15:524. <https://doi.org/10.1186/s13059-014-0524-x>.
78. Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, Heer FT, de Beer TAP, Rempfer C, Bordoli L, Lepore R, Schwede T. 2018. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res* 46:W296–W303. <https://doi.org/10.1093/nar/gky427>.
79. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. 2004. UCSF Chimera—a visualization system for exploratory research and analysis. *J Comput Chem* 25:1605–1612. <https://doi.org/10.1002/jcc.20084>.
80. Venselaar H, Te Beek TA, Kuipers RK, Hekkelman ML, Vriend G. 2010. Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces. *BMC Bioinformatics* 11:548. <https://doi.org/10.1186/1471-2105-11-548>.
81. Gelly JC, Joseph AP, Srinivasan N, de Brevern AG. 2011. iPBA: a tool for protein structure comparison using sequence alignment strategies. *Nucleic Acids Res* 39:W18–W23. <https://doi.org/10.1093/nar/gkr333>.
82. Capriotti E, Fariselli P, Casadio R. 2005. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res* 33:W306–W310. <https://doi.org/10.1093/nar/gki375>.
83. Laskowski RA. 2001. PDBsum: summaries and analyses of PDB structures. *Nucleic Acids Res* 29:221–222. <https://doi.org/10.1093/nar/29.1.221>.
84. Laskowski RA, Jabłońska J, Právda L, Vařeková RS, Thornton JM. 2018. PDBsum: structural summaries of PDB entries. *Protein Sci* 27:129–134. <https://doi.org/10.1002/pro.3289>.
85. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 13:2498–2504. <https://doi.org/10.1101/gr.1239303>.
86. Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, Fridman WH, Pages F, Trajanoski Z, Galon J. 2009. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* 25:1091–1093. <https://doi.org/10.1093/bioinformatics/btp101>.
87. Kanehisa M, Goto S. 2000. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 28:27–30. <https://doi.org/10.1093/nar/28.1.27>.
88. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. 2016. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* 44:D457–D462. <https://doi.org/10.1093/nar/gkv1070>.
89. Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, Haw R, Jassal B, Korninger F, May B, Milacic M, Roca CD, Rothfels K, Sevilla C, Shamovsky V, Shorsler S, Varusai T, Viteri G, Weiser J, Wu G, Stein L, Hermjakob H, D'Eustachio P. 2018. The Reactome Pathway Knowledgebase. *Nucleic Acids Res* 46:D649–D655. <https://doi.org/10.1093/nar/gkx1132>.
90. Pond SL, Frost SD, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* 21:676–679. <https://doi.org/10.1093/bioinformatics/bti079>.
91. Weaver S, Shank SD, Spielman SJ, Li M, Muse SV, Kosakovsky Pond SL. 2018. Datamonkey 2.0: a modern Web application for characterizing selective and other evolutionary processes. *Mol Biol Evol* 35:773–777. <https://doi.org/10.1093/molbev/msx335>.