Genomic variations of COVID-19 suggest multiple outbreak sources of transmission

Liangsheng Zhang[1, 2*], Jian-Rong Yang[3], Zhenguo Zhang[4*], and Zhenguo Lin[5*]

[1]Fujian Provincial Key Laboratory of Haixia Applied Plant Systems Biology, Key Laboratory of Ministry of Education for Genetics, Breeding and Multiple Utilization of Crops, Fujian Agriculture and Forestry University, Fuzhou, China.

[2]College of Agriculture and Biotechnology, Zhejiang University, Hangzhou, China

[3]Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou, China

[4]Independent Scholar, Irvine, CA, 92612. USA

[5]Department of Biology, Saint Louis University, St. Louis, Missouri, USA

* Correspondence: fafuzhang@163.com, zhangz.sci@gmail.com, Zhenguo.Lin@slu.edu.

Summary

The most important finding of this study is that COVID-19 strains form two well-supported clades (genotype I, or Type I, and Type II). Type II strains were likely evolved from Type I and are more prevalent than Type I among infected patients (68 Type II strains vs 29 Type I strains in total). Our results suggest the outbreak of type II COVID-19 likely occurred in the Huanan market, while the initial transmission of the type I virus to humans probably occurred at a different location in Wuhan. Second, by analyzing the three genomic sites distinguishing Type I and Type II strains, we found that the synonymous changes at two of the three sites confer higher protein translational efficiencies in Type II strains than in Type I strains, which might explain why Type II straints are more prevalent, implying that Type II is more contagious (transmissible) than Type I. These findings could be valuable for the current epidemic prevention and control. The timely sharing of our findings would benefit the public health officials in making policies, diagnosis and treatments.

Introduction

The 2019 novel coronavirus disease (COVID-19, previously known as 2019-nCoV) has been diagnosed in more than 70,000 deaths, more than 2,000 deaths, and more than 10,000 severe cases (http://2019ncov.chinacdc.cn/2019-nCoV/global.html). The current spread trend in China is declining, but it is increasing in other countries. Therefore, it is still challenging to effectively control this ourbreak worldwide. The recent COVID-19 virus was named as SARS-CoV-2, mainly based on its closest

relationship with the SARS-CoV virus. Our recent study showed that SARS-CoV-2 and SARS-CoV have common ancestors, as they form sister groups, and SARS-CoV-2 aggregates with two SARS-like bat viruses [1]. The branch length of the phylogenetic tree of the common ancestor of SARS virus and its recent bat virus (0.03) is short, and the branch length of SARS-CoV-2 and two SARS-like bat viruses is longer (0.09), indicating that there are many viruses in the middle not found. The Yunnan bat coronavirus (BatCoV RaTG13) isolated in 2013 was found to be most closely related to SARS-CoV-2 [2]. The phylogenetic tree of SARS-CoV-2 and their common ancestors of BatCoV RaTG13 has a branch length of only 0.02 (Figure S1). Therefore, we used BatCoV RaTG13 as an outgroup to study the origin and transmission history of SARS-CoV-2. As fears of global pandemic continue to rise, it is necessary to better understand the sources and transmission history of this outbreak and to monitor the changes of genomes for dominant viral strains. These studies are important for public-health officials to prepare better strategies for constraining the outbreak and prevention of further spread.

Data and Methods

We obtained 97 complete genomes of COVID-19 samples from GISAID (www.gisaid.org), NCBI and NMDC (http://nmdc.cn/#/nCov/). Sequence alignment of 97 COVID-19 genomes plus the strain BatCoV RaTG13 used by MAFFT (https://mafft.cbrc.jp/alignment/software/). Genome variable sites of Sequence alignment used the noisy (http://www.bioinf.uni-leipzig.de/Software/noisy/). The three type-specific variants correspond to the genomic positions 8750, 28112, and 29063, respectively; the coordinates are referred to as the sequence MN938384.1. The maximum likelihood (ML) phylogenetic tree used by FastTree (http://meta.microbesonline.org/fasttree/). The tRNA Adaptation Index (tAI) values were computed using Bio::CUA (https://metacpan.org/release/Bio-CUA), and the numbers of human tRNA genes were downloaded from http://gtrnadb.ucsc.edu.

Results and discussions

We obtained 97 complete genomes of COVID-19 samples and inferred their evolutionary relationships based on their genomic variants (Figure 1). Overall, we found only 0 to 3 mutations among the majority of COVID-19 genomes, and there are only 95 variable sites (Figure 1B). Their phylogenetic relationships suggest the presence of two major types of COVID-19, namely Type I and II (Figure 1A). The genomes of the two types mainly differ at three sites (Figure 1B), which are 8750, 28112, and 29063, based on MN938384.1's genome coordinates. Specifically, the nucleotides at the three sites are T, C, and T/C in Type I , and C, T, and C in Type II,

respectively. Based on the nucleotide at the site 29063, the Type I strains can be further divided into Type IA and IB. The number of genomes belonging to Type IA, IB and II are 10, 18, and 69, respectively. This finding suggests that the Type II strains are dominant in the infected populations.

We found that the three sites in Type IA and two in Type IB are identical to those in the BatCoV RaTG13 [2] (Fig. 1B), suggesting that the Type I may be more closely related to the ancestral human-infecting strain than Type II, consistent with a previous report [1]. Therefore, Type II was likely originated from a Type IB strain by accumulating muttaions at 8750 and 28112. Given that the Type I isolates (such as Wuhan/WH04/2020 [3]) have no direct link to Huanan market and that two Type II samples were isolated from the Huanan market (Wuhan/IVDC-HB-envF13-20 and 21), we speculated that the initial transmission of Type I virus to humans might have occurred at another location. Our analysis reinforces earlier reports that some cases had no link to the Huanan market [3-5] and suggests that different transmission sources are associated with different virus strains.

To further understand the functional effects of the three variants, we examined how these genmic variants might affect the translation of virus mRNAs in human cells. The mutations at 8750 and 29063 are synonymous (in gene *orf1ab* and *N*, repectively) and the one at 28112 is nonsynonymous, leading to a change from Leucine to Serine in the gene *ORF8*. Interestingly, we found that the two synonymous changes both confer higher translational efficiencies for the Type II strains than for the Type I ones (Figure 1C), based on the number of tRNA genes matching each codon and tRNA Adaptation Index (tAI) [6]. We speculate that the higher translational efficiencies might have enabled faster production of Type II virus particles, facilitated its spread, and led to its becoming dominant strains, implying that Type II is more contagious (transmissible) than Type I.

Our results above divided the current SARS-CoV-2 into two main types, with three sources of transmission, namely Type IA, Type IB, and Type II (Figure 2). Among them, Type IA is the earliest transmission source, and it did not occur in the Huanan Market, indicating that the original transmission source was not from the Huanan Market. Type II comes from the Huanan Market. As most samples detected belong to Type II, we speculated that type II is the major outbreak source. It is possible that Type IA, Type IB, and Type II may lead to different patient symptoms. It would be valuable to compare the symptoms of patients infected by different types of viruses. Recently, some asymptomatic carriers have been found [7], and it is worth to examine

the specific type of virus they infected and to determine whether the pathogenicity is different among different types of SARS-CoV-2 viruses.

In summary, our analyses show that there are two groups of COVID-19 viruses. Our results suggest the Huanan market is the third transmission source of the outbreak, while initial transmission of the virus to humans likely occurred at a different location. With more sequencing data of 2019-nCoV, we expect a more complete of transmission history to emerge. Our discovery suggests that patients infected with the different groups of viruses may need different treatments, because the Type II of translation is more efficient and may lead to faster onset of illness in infected patients. Comparative studies of the symptoms of patients infected by the two types of 2019-nCoVs will improve our understanding of virulent effects of the three variants. Because virus genomes are vulable for identifying their transmission sources and for monitoring the accumulation of new mutaions, we urge a more rapid sequencing and release of SARS-CoV-2 genomes.

Acknowledgments. We acknowledge the authors and the originating and submitting laboratories of the nucleotide sequences from the Global Initiative on Sharing All Influenza Data's EpiFlu Database, NCBI and NMDC (http://nmdc.cn/#/nCov/)(12 Feb 2020, 98 isolates).

Potential conflicts of interest. All authors: No reported conflicts.

All authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Conflicts that the editors consider relevant to the content of the manuscript have been disclosed.

Figure 1. A phylogenetic tree of the 97 COVID-19 strains and their genomic variants.

A, A maximum likelihood (ML) phylogenetic tree of the human COVID-19 with approximately ML method by FastTree (http://meta.microbesonline.org/fasttree/). The phylogenetic tree was constructed using the sequence alignment shown in **B**. The two groups, Type I and Type II, are colored in blue and red, respectively.

B, Sequence alignment of 97 COVID-19 genomes where only variable sites are shown. Each line corresponds to one branch in the phylogenetic tree to the left. The corresponding sites from the strain BatCoV RaTG13 are shown on the top separated by a red line. Three type-specific variants are marked in red arrows, corresponding to the genomic positions 8750, 28112, and 29063, respectively; the coordinates are referred to the sequence MN938384.1.

C, the codon changes caused by the differences in the three sites. The tAI values were

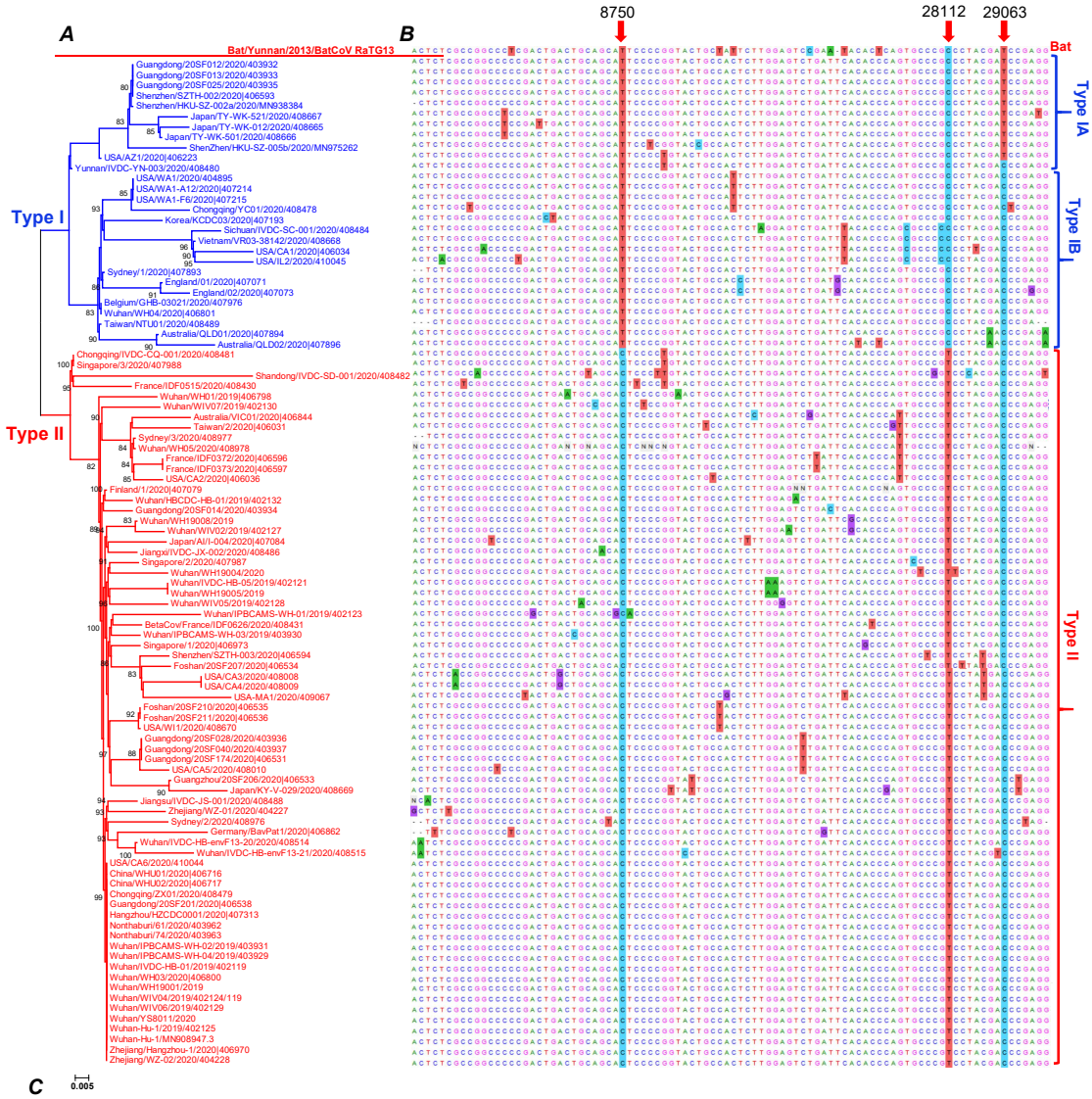
computed using Bio::CUA (https://metacpan.org/release/Bio-CUA), and the numbers of human tRNA genes were downloaded from http://gtrnadb.ucsc.edu.

Figure 2. A simple COVID-19 virus transmission model.

The COVID-19 has at least three sources of transmission, namely Type IA, Type IB and Type II.

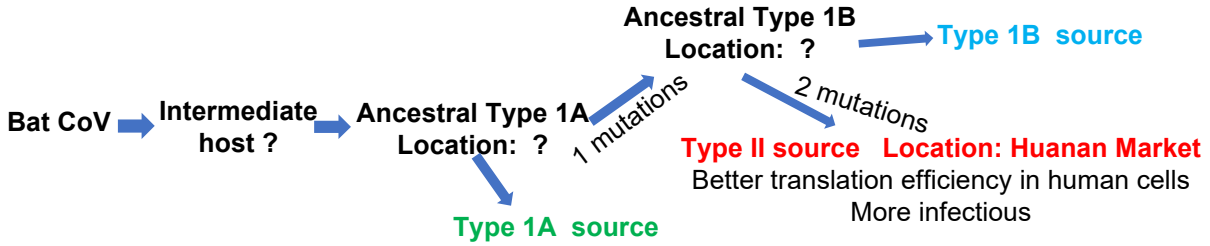
Supplementary Figure 1. The SARS-cov phylogenetic tree uses MERS-CoV as an outgroup.

1. Zhang L., et al., *Origin and evolution of the 2019 novel coronavirus.* Clin Infect Dis, 2020.
2. Zhou, P., et al., *A pneumonia outbreak associated with a new coronavirus of probable bat origin.* Nature, 2020.
3. Lu, R., et al., *Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding.* The Lancet.
4. Huang, C., et al., *Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China.* Lancet, 2020.
5. Li, Q., et al., *Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia.* New England Journal of Medicine, 2020.
6. dos Reis, M., R. Savva, and L. Wernisch, *Solving the riddle of codon usage preferences: a test for translational selection.* Nucleic Acids Res, 2004. **32**(17): p. 5036-44.
7. Bai, Y., et al., *Presumed Asymptomatic Carrier Transmission of COVID-19.* JAMA, 2020.



	Codon	# Anticodon tRNA genes; tAI[a]	Codon	# Anticodon tRNA genes; tAI	Codon	# Anticodon tRNA genes; tAI
SNP position	8750		28112		29063	
RaTG13 (bat)	AGT (Ser)	0; 0.16	TCA (Leu)	4; 0.14	TTT (Phe)	0; 0.20
Type I	AGT (Ser)	0; 0.16	TCA (Leu)	4; 0.14	TTT (Phe)	0; 0.20
Type II	AGC (Ser)	8; 0.28	TTA (Ser)	4; 0.14	TTC (Phe)	10; 0.34

a: this column shows the number of tRNA genes in human genome with anticodons matching the considered codons. tAI is a measure of codon's translational efficiency[5], the higher the more efficient.


Ancestral Type 1B
Location: ?
Type 1B source
Bat CoV
Intermediate host ?
Ancestral Type 1A
Location: ?
1 mutations
2 mutations
Type II source Location: Huanan Market
Better translation efficiency in human cells
More infectious
Type 1A source

EU371559.1 SARS coronavirus ZJ02 complete genome
EU371564.1 SARS coronavirus BJ182-12 complete genome
AY502923.1 SARS coronavirus TW10 complete genome
AY394998.1 SARS coronavirus LC1 complete genome
AY502928.1 SARS coronavirus TW5 complete genome
AY291451.1 SARS coronavirus TW1 complete genome
AY502926.1 SARS coronavirus TW3 complete genome
AY714217.1 SARS Coronavirus CDC*200301157 complete genome
AY323977.2 SARS coronavirus HSR 1 complete genome
AY394978.1 SARS coronavirus GZ-B complete genome
AY559093.1 SARS coronavirus Sin845 complete genome
AY559096.1 SARS coronavirus Sin850 complete genome
AY394991.1 SARS coronavirus HZS2-Fc complete genome
AY394987.1 SARS coronavirus HZS2-Fb complete genome
AY394992.1 SARS coronavirus HZS2-C complete genome
AY394983.1 SARS coronavirus HSZ2-A complete genome
AY394993.1 SARS coronavirus HGZ8L2 complete genome
AY278554.2 SARS coronavirus CUHK-W1 complete genome
AY304488.1 SARS coronavirus SZ16 complete genome
AY304486.1 SARS coronavirus SZ3 complete genome
AY390556.1 SARS coronavirus GZ02 complete genome
AY395003.1 SARS coronavirus ZS-C complete genome
AY394996.1 SARS coronavirus ZS-B complete genome
AY394994.1 SARS coronavirus HSZ-Bc complete genome
AY394985.1 SARS coronavirus HSZ-Bb complete genome
AY394986.1 SARS coronavirus HSZ-Cb complete genome
AY394995.1 SARS coronavirus HSZ-Cc complete genome
AY394999.1 SARS coronavirus LC2 complete genome
AY351680.1 SARS coronavirus ZMY 1 complete genome
FJ882939.1 SARS coronavirus wtic-MB isolate P3pp16 complete genome
FJ882948.1 SARS coronavirus MA15 isolate P3pp3 complete genome
FJ882961.1 SARS coronavirus MA15 isolate P3pp5 complete genome
FJ882952.1 SARS coronavirus MA15 isolate P3pp4 complete genome

SARS-CoV

KY417146.1 Bat SARS-like coronavirus isolate Rs4231 complete genome
KY417150.1 Bat SARS-like coronavirus isolate Rs4874 complete genome
KF367457.1 Bat SARS-like coronavirus WIV1 complete genome
KC881006.1 Bat SARS-like coronavirus Rs3367 complete genome
KC881005.1 Bat SARS-like coronavirus RsSHC014 complete genome
KY417144.1 Bat SARS-like coronavirus isolate Rs4084 complete genome
MK211376.1 Coronavirus BtRs-BetaCoV/YN2018B complete genome
KY417152.1 Bat SARS-like coronavirus isolate Rs9401 complete genome
KY417151.1 Bat SARS-like coronavirus isolate Rs7327 complete genome
KY417145.1 Bat SARS-like coronavirus isolate Rf4092 complete genome
KJ473816.1 BtRs-BetaCoV/YN2013 complete genome
KY770858.1 Bat coronavirus isolate Anlong-103 complete genome
KY770859.1 Bat coronavirus isolate Anlong-112 complete genome
MK211378.1 Coronavirus BtRs-BetaCoV/YN2018D complete genome
KY417143.1 Bat SARS-like coronavirus isolate Rs4081 complete genome
KY417149.1 Bat SARS-like coronavirus isolate Rs4255 complete genome
KY417148.1 Bat SARS-like coronavirus isolate Rs4247 complete genome
KY417147.1 Bat SARS-like coronavirus isolate Rs4237 complete genome
MK211375.1 Coronavirus BtRs-BetaCoV/YN2018A complete genome
KY417142.1 Bat SARS-like coronavirus isolate As6526 complete genome
MK211377.1 Coronavirus BtRs-BetaCoV/YN2018C complete genome
KP886808.1 Bat SARS-like coronavirus YNLF 31C complete genome
KP886809.1 Bat SARS-like coronavirus YNLF 34C complete genome
KJ473815.1 BtRs-BetaCoV/GX2013 complete genome
DQ071615.1 Bat SARS coronavirus Rp3 complete genome
JX993988.1 Bat coronavirus Cp/Yunnan2011 complete genome
KF569996.1 Rhinolophus affinis coronavirus isolate LYRa11 complete genome
MK211374.1 Coronavirus BtRl-BetaCoV/SC2018 complete genome
JX993987.1 Bat coronavirus Rp/Shaanxi2011 complete genome
KJ473811.1 BtRf-BetaCoV/JL2012 complete genome
KY770860.1 Bat coronavirus isolate Jiyuan-84 complete genome
KJ473813.1 BtRf-BetaCoV/SX2013 complete genome
KJ473812.1 BtRf-BetaCoV/HeB2013 complete genome
DQ648856.1 Bat coronavirus
DQ412042.1 Bat SARS coronavirus Rf1 complete genome
KJ473814.1 BtRs-BetaCoV/HuB2013 complete genome
DQ648857.1 Bat coronavirus
DQ412043.1 Bat SARS coronavirus Rm1 complete genome
GQ153542.1 Bat SARS coronavirus HKU3-7 complete genome
GQ153543.1 Bat SARS coronavirus HKU3-8 complete genome
GQ153547.1 Bat SARS coronavirus HKU3-12 complete genome
DQ084199.1 bat SARS coronavirus HKU3-2 complete genome
GQ153539.1 Bat SARS coronavirus HKU3-4 complete genome
GQ153541.1 Bat SARS coronavirus HKU3-6 complete genome
GQ153540.1 Bat SARS coronavirus HKU3-5 complete genome
GQ153548.1 Bat SARS coronavirus HKU3-13 complete genome
GQ153546.1 Bat SARS coronavirus HKU3-11 complete genome
GQ153545.1 Bat SARS coronavirus HKU3-10 complete genome
GQ153544.1 Bat SARS coronavirus HKU3-9 complete genome
FJ211859.1 Recombinant coronavirus clone Bat SARS-CoV complete sequence
DQ022305.2 Bat SARS coronavirus HKU3-1 complete genome
DQ084200.1 bat SARS coronavirus HKU3-3 complete genome
USA/AZ1/2020|406223
Guangdong/20SF012/2020/403932
Guangdong/20SF013/2020/403933
Guangdong/20SF025/2020/403935
Shenzhen/SZTH-002/2020|406593
USA/WA1-F6/2020|407215
USA/WA1-A12/2020|407214
USA/WA1/2020/404895
USA/CA1/2020|406034
Korea/KCDC03/2020|407193
Wuhan/WH04/2020|406801
England/02/2020|407073
England/01/2020|407071
Australia/VIC01/2020|406844
France/IDF0372/2020|406596
France/IDF0373/2020|406597
Taiwan/2/2020|406031
USA/CA2/2020|406036
Wuhan/HBCDC-HB-01/2019/402132
Finland/1/2020|407079
Guangdong/20SF014/2020/403934
Nonthaburi/61/2020/403962
Nonthaburi/74/2020/403963
Wuhan/IPBCAMS-WH-04/2019/403929
Wuhan/IVDC-HB-01/2019/402119
Wuhan/WIV04/2019/402124/119
Wuhan/WIV06/2019/402129
Wuhan-Hu-1/2019/402125
Zhejiang/WZ-02/2020/404228
Wuhan/IPBCAMS-WH-02/2019/403931
Wuhan/IPBCAMS-WH-01/2019/402123
Wuhan/WIV05/2019/402128
Guangdong/20SF174/2020|406531
Guangdong/20SF028/2020/403936
Guangdong/20SF040/2020/403937
Zhejiang/WZ-01/2020/404227
Germany/BavPat1/2020|406862
Guangdong/20SF201/2020|406538
Zhejiang/Hangzhou-1/2020|406970
Wuhan/YS8011/2020
Singapore/1/2020|406973
Foshan/20SF211/2020|406536
Foshan/20SF210/2020|406535
Wuhan/WIV07/2019/402130
Guangzhou/20SF206/2020|406533
Wuhan/IPBCAMS-WH-03/2019/403930
Wuhan/WH19008/2019
Wuhan/WIV02/2019/402127
Japan/AI/I-004/2020|407084
China/WHU01/2020|406716
China/WHU02/2020|406717
Wuhan/WH19001/2019
Foshan/20SF207/2020|406534
Shenzhen/SZTH-003/2020|406594
Wuhan/WH01/2019|406798
Wuhan/WH03/2020|406800
Wuhan/WH19005/2019
Wuhan/IVDC-HB-05/2019/402121
Wuhan/WH19004/2020
Shenzhen/HKU-SZ-002a/2020/MN938384
ShenZhen/HKU-SZ-005b/2020/MN975262
Yunnan/bat/RaTG13/2013/402131

SARS-CoV-2

MG772933.1 Bat SARS-like coronavirus isolate bat-SL-CoVZC45 complete genome
MG772934.1 Bat SARS-like coronavirus isolate bat-SL-CoVZXC21 complete genome
GU190215.1 Bat coronavirus BM48-31/BGR/2008 complete genome
KY352407.1 Severe acute respiratory syndrome-related coronavirus strain BtKY72 complete genome
KJ477103.2 Middle East respiratory syndrome-related coronavirus isolate NRCE-HKU270 complete genome
KF192507.1 Middle East respiratory syndrome coronavirus complete genome
MF598663.1 Middle East respiratory syndrome-related coronavirus strain camel/UAE B73 2015 complete genome
KF600652.1 Middle East respiratory syndrome coronavirus isolate Riyadh 2 2012 complete genome
KJ614529.1 Human betacoronavirus 2c Jordan-N3/2012 isolate MG167 complete genome
JX869059.2 Human betacoronavirus 2c EMC/2012 complete genome
MH734115.1 Middle East respiratory syndrome-related coronavirus isolate MERS-CoV camel/Kenya/C1272/2018 complete genome
MH734114.1 Middle East respiratory syndrome-related coronavirus isolate MERS-CoV camel/Kenya/C1215/2018 complete genome
KU740200.1 Middle East respiratory syndrome coronavirus isolate MERS CoV/camel/Egypt/NRCE-NC163/2014 partial genome
MG923466.1 Middle East respiratory syndrome-related coronavirus isolate MERS-CoV camel/Ethiopia/AAU-EPHI-HKU4412/2017 complete genome
MG923468.1 Middle East respiratory syndrome-related coronavirus isolate MERS-CoV camel/Ethiopia/AAU-EPHI-HKU4458/2017 complete genome
MG923467.1 Middle East respiratory syndrome-related coronavirus isolate MERS-CoV camel/Ethiopia/AAU-EPHI-HKU4448/2017 complete genome
MK564474.1 Middle East respiratory syndrome-related coronavirus isolate camel/MERS/Amibara/118/2017 complete genome
MK564475.1 Middle East respiratory syndrome-related coronavirus isolate camel/MERS/Amibara/126/2017 complete genome

MERS-CoV

0.05