Genomic variations of SARS-CoV-2 suggest multiple outbreak sources of transmission

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

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

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

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

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

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

Summary

We examined 169 genomes of SARS-CoV-2 and found that they can be classified into two major genotypes, Type I and Type II. Type I can be further divided into Type IA and IB. Our phylogenetic analysis showed that the Type IA resembles the ancestral SARS-CoV-2 most. Type II was likely evolved from Type I and predominant in the infections. Our results suggest that Type II SARS-CoV-2 was the source of the outbreak in the Wuhan Huanan market and it was likely originated from a super-spreader. The outbreak caused by the Type I virus should have occurred somewhere else, because the patients had no direct link to the market. Furthermore, by analyzing three genomic sites that distinguish Type I and Type II strains, we found that 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 strains are predominant, implying that Type II is more contagious (transmissible) than Type I. These findings could be valuable for the current epidemic prevention and control.

Introduction

The outbreak of the coronavirus disease 2019 (COVID-19), caused by a novel coronavirus named “Severe Acute Respiratory Syndrome CoronaVirus 2 (SARS-CoV-2) ”, has now been detected in over 60 nations and lead to more than 80,000 confirmed cases and ~3,000 deaths (http://2019ncov.chinacdc.cn/2019-nCoV/global.html). Although the spread of SARS-CoV-2 in China seems to be largely contained, the confirmed cases outside China have been rising. Therefore, the risk of global pandemic of SARS-CoV-2 is still

growing. Currently, the outbreak sources and the transmission history of SARS-CoV-2 is far from well-understood. Revealing the evolutionary history of different SARS-CoV-2 samples allows us to infer the virus transmission routes and to identify novel mutations associated with the transmissions. Such information will be valuable for vaccine development and disease control.

Our recent study showed that SARS-CoV-2 form a sister group to two SARS-like bat viruses (bat-SL-CoVZC45 and bat-SL-CoVZXC21) [1] that were collected in Zhoushan, Zhejiang Province, China, from 2015 to 2017 [2]. A subsequently released sequence of a bat coronavirus (BatCoV RaTG13) isolated in Yunnan in 2013 was found to be more similar to the sequences of SARS-CoV-2 [3], which is confirmed by our phylogenetic analysis based on whole genome sequences (Figure S1). Therefore, we used BatCoV RaTG13 and the two SARS-like bat viruses (bat-SL-CoVZC45 and bat-SL-CoVZXC21) as an outgroup to study the origin and transmission history of SARS-CoV-2.

Materials and Methods

The complete genomes of SARS-CoV-2 samples were obtained from GISAID (www.gisaid.org), NCBI and NMDC (http://nmdc.cn/#/nCov/). Alignment of the complete genome sequences and that of BatCoV RaTG13 was carried out by MAFFT [4].

We identified genome variable sites from the sequence alignment using Noisy [5]. Variants that are present in at least two samples were used in our subsequence analysis. Some samples have unusually large number of mutation sites, which could be due to due to sequencing or assembly errors. These sequences were also excluded in our analysis.

We inferred the phylogeny of the SARS-CoV-2 isolates based on the variable sites using the maximum likelihood (ML) method by FastTree [6]. The tRNA Adaptation Index (tAI), an indicator of codon translational efficiency, was computed using the tool Bio::CUA (https://metacpan.org/release/Bio-CUA), and the numbers of human tRNA genes for each codon (used for computing tAI) were downloaded from http://gtrnadb.ucsc.edu.

Results and discussions

We obtained 169 complete genomes of SARS-CoV-2 samples and identified variable sites based on whole genome alignment (Figure 1). In most cases, different SARS-CoV-2 genomes only differ in 0 to 3 sites. A total number of 207 variable sites were identified (Figure S2). It shows that the amplitude of mutations is small, and

their evolution history can be inferred from these variable sites.

Their phylogenetic relationships reveal two major genotypes from the SARS-CoV-2 samples, namely Type I and II (Figures 1A, S3 and S4). The genomes of the two types mainly differ at three sites (Figure 1B), which are 8750, 28112, and 29063, using the genome coordinates of sample MN938384.1 as a reference. 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 11, 37, and 121, respectively.

We found that the three sites in Type IA and two in Type IB are identical to those in three bat viruses BatCoV RaTG13 [3], bat-SL-CoVZC45 and bat-SL-CoVZXC21[2](Figure 1B), suggesting that the Type I may be more closely related to the ancestral human-infecting strain than Type II. Therefore, in the principle of parsimony, Type IA resembles the most ancestral lineage of SARS-CoV-2, and Type IB was derived from Type IA by a new mutation at site 29063. In addition, Type II may have originated from a Type IB strain by accumulating mutations at sites 8750 and 28112.

Two environmental samples isolated from the Huanan market (Wuhan/IVDC-HB-envF13-20 and 21) belong to Type II, and no samples from Type I have a direct link to the Huanan market (such as Wuhan/WH04/2020 [7]). These observations suggest that the outbreak in the Huanan market was triggered by the Type II virus and that the initial transmission of Type I viruses to humans might have occurred somewhere else in Wuhan, probably preceding the outbreak of Type II in the Huanan market. Our speculation is in line with earlier reports that 14 of the first reported 41 cases had no link to the Huanan market [7-9].

In addition to the above-mentioned transmission sources, we also found that there are several large clades in Type II (Group 1-3). The Group-1 of Type II are identical, suggesting they were likely originated from the same transmission source. Group-2 and Group-3 share at least one mutation, indicating that they shared the same transmission source. As only a small portion of SARS-CoV-2 isolates have been sequenced, the actual numbers of infected patients in each group are expected to be much larger. We speculate that each group was likely from a source of super transmission or super-spreader.

Because Type II strains account for most of the sampled genomes, it is reasonable to speculate that Type II is more contagious, assuming no strong sampling bias. To better understand why the Type II virus is more prevalent, we then focused on the three variable genomic sites that distinguish Type II from Type I strains. First, we noticed the mutations at 8750 and 29063 are synonymous (in gene *orf1ab* and *N*, respectively) 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 Type I (Figure 2), based on the number of tRNA genes matching each codon and tRNA Adaptation Index (tAI) [10]. We speculate that the higher translational efficiencies might have enabled faster production of Type II virus particles, facilitated its spread, and led to its dominance in infections, implying that Type II is more contagious (transmissible) than Type I. If our speculation is correct, this would provide guidelines for treating Type I and II patients differently.

In summary, our results illustrate the presence of two major genotypes of SARS-CoV-2, suggesting of at least two, possibly three, major outbreak sources (Figure 3). The outbreak in the Huanan market may not be the initiation transmission of SARS-CoV-2 to human, and thus the location of initial transmission to humans remain to be determined. As most samples detected belong to Type II, it indicates that Type II is more transmissible. We identified three genomic variants that separate the two types of SARS-CoV-2. Our data suggest that two variants are linked to improved translation efficiency in Type II. The impacts of these genomic variants could be further evaluated by comparing the symptoms of patients infected by each type of viruses. Particularly, some asymptomatic carriers have been recently found [11]. It is worth examining whether these asymptomatic cases are enriched in Type I. With more sequencing data of SARS-CoV-2, we expect a more complete understanding of transmission history.

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/).

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

Figure 1. A phylogenetic tree of the 169 SARS-CoV-2strains and their genomic variants.

A, A maximum likelihood (ML) phylogenetic tree of the human SARS-CoV-2 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 the SARS-CoV-2 genomes where only major 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.

Figure 2. The codon changes caused by the three variant sites in Figure 1. 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 3. A simple SARS-CoV-2 virus transmission model.

The SARS-CoV-2 may have two or three sources of transmissions, corresponding to the three genotypes we identified here, Type IA, Type IB and Type II.

Supplementary Figure 1. The SARS-CoV and SARS-CoV-2 phylogenetic tree uses MERS-CoV as an outgroup.

Supplementary Figure 2. Sequence alignment of 169 COVID-19 genomes where only the variable sites are shown.

Supplementary Figure 3. A maximum likelihood (ML) phylogenetic tree of the human COVID-19 virus. The phylogenetic tree was constructed using the genome sequence alignment. The two groups, Type I and Type II, are colored in blue and red, respectively.

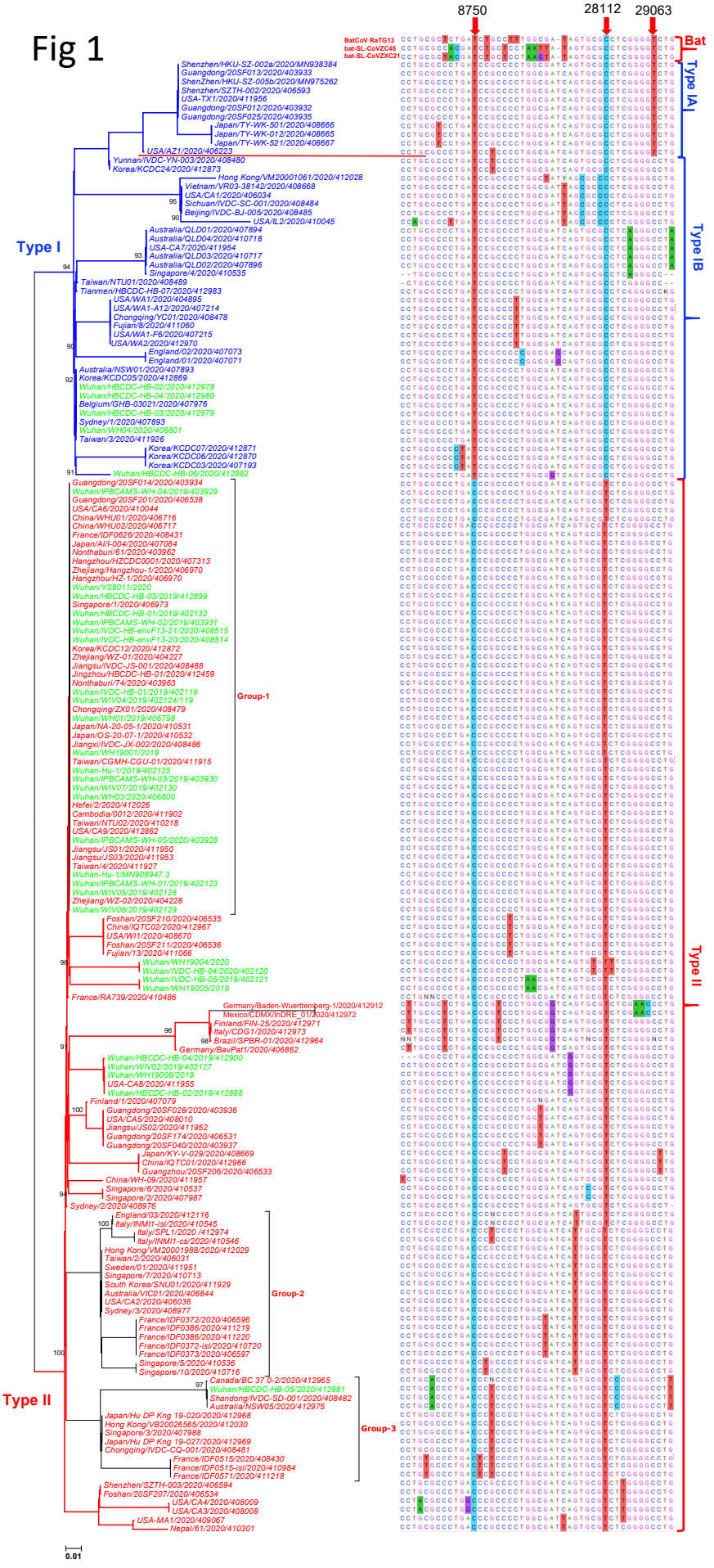
Supplementary Figure 4. A maximum likelihood (ML) phylogenetic tree of the human COVID-19 virus. The phylogenetic tree was constructed using the variable sites shown in **Supplementary Figure 2**. The two groups, Type I and Type II, are colored in blue and red, respectively.

1. Zhang L., et al., *Origin and evolution of the 2019 novel coronavirus.* Clin Infect Dis, 2020.
2. Hu, D., et al., *Genomic characterization and infectivity of a novel SARS-like coronavirus in Chinese bats.* Emerg Microbes Infect, 2018. **7**(1): p. 154.
3. Zhou, P., et al., *A pneumonia outbreak associated with a new coronavirus of probable bat origin.* Nature, 2020.
4. Yamada, K.D., K. Tomii, and K. Katoh, *Application of the MAFFT sequence alignment program to large data-reexamination of the usefulness of chained guide trees.* Bioinformatics, 2016. **32**(21): p. 3246-3251.
5. Dress, A.W., et al., *Noisy: identification of problematic columns in multiple sequence alignments.* Algorithms Mol Biol, 2008. **3**: p. 7.
6. Price, M.N., P.S. Dehal, and A.P. Arkin, *FastTree 2--approximately maximum-likelihood trees for large alignments.* PLoS One, 2010. **5**(3): p. e9490.
7. Lu, R., et al., *Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding.* The Lancet.
8. Huang, C., et al., *Clinical features of patients infected with 2019 novel coronavirus in Wuhan,*

China. Lancet, 2020.

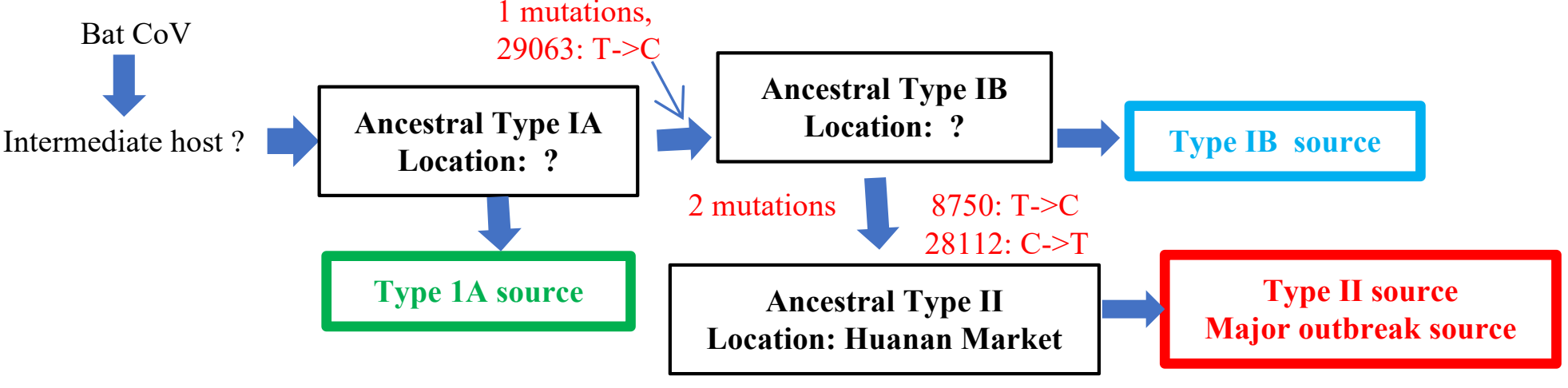
9. Li, Q., et al., *Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia.* New England Journal of Medicine, 2020.
10. dos Reis, M., R. Savva, and L. Wernisch, *Solving the riddle of codon usage preferences: a test for translational selection.* Nucleic Acids Res, 2004. **32**(17): p. 5036-44.
11. Bai, Y., et al., *Presumed Asymptomatic Carrier Transmission of COVID-19.* JAMA, 2020.

Fig 1


8750
28112
29063
BatCoV RaTG13
bat-SL-CoVZC45
bat-SL-CoVZXC21
Bat
Type IA
Type IB
Type II
Type I
Type II
Group-1
Group-2
Group-3
Shenzhen/HKU-SZ-002a/2020/MN938384
Guangdong/20SF013/2020/403933
ShenZhen/HKU-SZ-005b/2020/MN975262
Shenzhen/SZTH-002/2020/406593
USA-TX1/2020/411956
Guangdong/20SF012/2020/403932
Guangdong/20SF025/2020/403935
Japan/TY-WK-501/2020/408666
Japan/TY-WK-012/2020/408665
Japan/TY-WK-521/2020/408667
USA/AZ1/2020/406223
Yunnan/IVDC-YN-003/2020/408480
Korea/KCDC24/2020/412873
Hong Kong/VM20001061/2020/412028
Vietnam/VR03-38142/2020/408668
USA/CA1/2020/406034
Sicilia/IVDC-SC-001/2020/408484
Beijing/IVDC-BJ-005/2020/408485
USA/IL2/2020/410045
Australia/QLD01/2020/407894
Australia/QLD04/2020/410718
USA-CA7/2020/411954
Australia/QLD03/2020/410717
Australia/QLD02/2020/407896
Singapore/4/2020/410535
Taiwan/NTU01/2020/408489
Tianmen/HBCDC-HB-07/2020/412983
USA/WA1/2020/404895
USA/WA1-A12/2020/407214
Chongqing/YC01/2020/408478
Fujian/8/2020/411060
USA/WA1-F6/2020/407215
USA/WA2/2020/412970
England/02/2020/407073
England/01/2020/407071
Australia/NSW01/2020/407893
Korea/KCDC05/2020/412869
Wuhan/HBCDC-HB-02/2020/412978
Wuhan/HBCDC-HB-04/2020/412980
Belgium/GHB-03021/2020/407976
Wuhan/HBCDC-HB-03/2020/412979
Sydney/1/2020/407893
Wuhan/WH04/2020/406801
Taiwan/3/2020/411926
Korea/KCDC07/2020/412871
Korea/KCDC06/2020/412870
Korea/KCDC03/2020/407193
Wuhan/HBCDC-HB-06/2020/412982
Guangdong/20SF014/2020/403934
Wuhan/IPBCAMS-WH-01/2019/403929
Guangdong/20SF201/2020/406538
USA/CA6/2020/410044
China/WHU01/2020/406716
China/WHU02/2020/406717
France/IDF0626/2020/408431
Japan/AI/I-004/2020/407084
Nonthaburi/61/2020/403962
Hangzhou/HZCDC0001/2020/407313
Zhejiang/Hangzhou-1/2020/406970
Hangzhou/HZ-1/2020/406970
Wuhan/YS8011/2020
Wuhan/HBCDC-HB-03/2019/412899
Singapore/1/2020/406973
Wuhan/HBCDC-HB-01/2019/402132
Wuhan/IPBCAMS-WH-02/2019/403931
Wuhan/IVDC-HB-envF13-21/2020/408515
Wuhan/IVDC-HB-envF13-22/2020/408514
Korea/KCDC12/2020/412872
Zhejiang/WZ-01/2020/404227
Jiangsu/IVDC-JS-001/2020/408488
Jingzhou/HBCDC-HB-01/2020/412459
Nonthaburi/74/2020/403963
Wuhan/IVDC-HB-01/2019/402119
Wuhan/WIV04/2019/402124/119
Chongqing/ZX01/2020/408479
Wuhan/WH01/2019/406798
Japan/NA-20-05-1/2020/410531
Japan/OS-20-07-1/2020/410532
Jiangxi/IVDC-JX-002/2020/408486
Wuhan/WH19001/2019
Taiwan/CGMH-CGU-01/2020/411915
Wuhan-Hu-1/2019/402125
Wuhan/IPBCAMS-WH-02/2019/403930
Wuhan/WIV07/2019/402130
Wuhan/WH03/2020/406800
Hefei/2/2020/412026
Cambodia/0012/2020/411902
Taiwan/NTU02/2020/410218
USA/CA9/2020/412862
Wuhan/IPBCAMS-WH-05/2020/403928
Jiangsu/JS01/2020/411950
Jiangsu/JS03/2020/411953
Taiwan/4/2020/411927
Wuhan-Hu-1/MN908947.3
Wuhan/IPBCAMS-WH-01/2019/402123
Wuhan/WIV05/2019/402128
Zhejiang/WZ-02/2020/404228
Wuhan/WIV06/2019/402129
Foshan/20SF210/2020/406535
China/IQTC02/2020/412967
USA/WI1/2020/408670
Foshan/20SF211/2020/406536
Fujian/13/2020/411066
Wuhan/WH19004/2020
Wuhan/IVDC-HB-04/2020/402120
Wuhan/IVDC-HB-05/2019/402121
Wuhan/WH19005/2019
France/RA739/2020/410486
Germany/Baden-Wuerttemberg-1/2020/412912
Mexico/CDMX/InDRE_01/2020/412972
Finland/FIN-25/2020/412971
Italy/CDG1/2020/412973
Brazil/SPBR-01/2020/412964
Germany/BavPat1/2020/406862
Wuhan/HBCDC-HB-04/2019/412900
Wuhan/WIV02/2019/402127
Wuhan/WH19008/2019
USA-CA8/2020/411955
Wuhan/HBCDC-HB-02/2019/412898
Finland/1/2020/407079
Guangdong/20SF208/2020/403936
USA/CA5/2020/408010
Jiangsu/JS02/2020/411952
Guangdong/20SF174/2020/406531
Guangdong/20SF040/2020/403937
Japan/KY-V-029/2020/408669
China/IQTC01/2020/412966
Guangzhou/20SF206/2020/406533
China/WH-09/2020/411957
Singapore/6/2020/410537
Singapore/2/2020/407987
Sydney/2/2020/408976
England/03/2020/412116
Italy/INMI1-is/2020/410545
Italy/SPL1/2020 /412974
Italy/INMI1-cs/2020/410546
Hong Kong/VM20001988/2020/412029
Taiwan/2/2020/406033
Sweden/01/2020/411951
Singapore/7/2020/410713
South Korea/SNU01/2020/411929
Australia/VIC01/2020/406844
USA/CA2/2020/406036
Sydney/3/2020/408977
France/IDF0372/2020/406596
France/IDF0386/2020/411219
France/IDF0386/2020/411220
France/IDF0372-is/2020/410720
France/IDF0373/2020/406597
Singapore/5/2020/410536
Singapore/10/2020/410716
Canada/BC 37 0-2/2020/412965
Wuhan/HBCDC-HB-05/2020/412981
Shandong/IVDC-SD-001/2020/408482
Australia/NSW05/2020/412975
Japan/Hu DP Kng 19-020/2020/412968
Hong Kong/VB20026565/2020/412030
Singapore/3/2020/407988
Japan/Hu DP Kng 19-027/2020/412969
Chongqing/IVDC-CQ-001/2020/408481
France/IDF0515/2020/408430
France/IDF0515-is/2020/410984
France/IDF0571/2020/411218
Shenzhen/SZTH-003/2020/406594
Foshan/20SF207/2020/406534
USA/CA4/2020/408009
USA-CA3/2020/408008
USA-MA1/2020/409067
Nepal/61/2020/410301
0.01

	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[10], the higher the more efficient.


Bat CoV
Intermediate host ?
Ancestral Type IA
Location: ?
Type 1A source
1 mutations,
29063: T->C
Ancestral Type IB
Location: ?
Type IB source
2 mutations
8750: T->C
28112: C->T
Ancestral Type II
Location: Huanan Market
Type II source
Major outbreak source