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

Liangsheng Zhang¹,²*, Jian-Rong Yang³, Zhenguozhao Zhang⁴*, and Zhenguolin⁵*

¹Fujian 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.
²College of Agriculture and Biotechnology, Zhejiang University, Hangzhou, China
³Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou, China
⁴Independent Scholar, Irvine, CA, 92612. USA
⁵Department of Biology, Saint Louis University, St. Louis, Missouri, USA

* Correspondence: fafuzhang@163.com, zhangz.sci@gmail.com, Zhenguolin@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 strains 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 outbreak 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://gtmadb.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 mutations 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 genomic variants might affect the translation of virus mRNAs in human cells. 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 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 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 valuable for identifying their transmission sources and for monitoring the accumulation of new mutations, 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://gtmadb.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.

Bat CoV → Intermediate host?

Ancestral Type 1A
Location: ?
1 mutations

Type 1A source

Type 1B
Location: ?
2 mutations

Type 1B source
Location: Huanan Market
Better translation efficiency in human cells
More infectious

CC-BY-NC-ND 4.0 International license
It is made available under a (which was not certified by peer review)
The copyright holder for this preprint this version posted February 26, 2020. ; https://doi.org/10.1101/2020.02.25.20027953
doi: medRxiv preprint