doi:10.20944/preprints202006.0044.v1

Major Concerns on the Identification of Bat Coronavirus Strain RaTG13 and Quality of Related Nature Paper

Authors:

Xiaoxu Sean Lin, Global Health Knowledge Exchange Inc., Silver spring, MD, U.S.A.
Shizhong Chen, Genestitute, San Diego, CA, U.S.A

Abstract

A recent manuscript (Zhou, P. et al. "A pneumonia outbreak associated with a new coronavirus of probable bat origin", Nature 579, 270–273 (2020). https://doi.org/10.1038/s41586-020-2012-7) from Wuhan Institute of Virology claimed the identification of a bat coronavirus, RaTG13, which showed 96.2% genome homology with SARS-CoV-2. In this paper, we raise the puzzling observations surrounding the identification, characterization, unique genome features of this RaTG13 strain, as well as its 100% nucleotide identity in partial RdRp gene with another bat coronavirus strain BtCoV/4991. And the paper presented premature hypothesis of potential bat origin of SARS-CoV-2 while RaTG13 strain was not successfully isolated. We also present the concerns on the methodology, data quality and experiment procedures described in this paper. We call for the authors to provide additional data, to share related samples to be verified and further characterized by other scientists.

Keywords: Epidemiology; COVID-19; coronavirus; bat; RaTG13; BtCoV/4991; SARS-CoV-2; Pangolin Coronavirus; next generation sequencing

----

After COVID-19 outbreak started in China, one group of Chinese scientists, including Peng Zhou and Zhengli Shi from Wuhan Institute of Virology (WIV), published an impactful article in Nature on February 3, 2020[1]. The paper was first submitted on January 20, 2020. It claimed that full-length coronavirus genome sequences were obtained from five patients and the viral genome is 96.2% identical at the whole-genome level to a bat coronavirus RaTG13 strain. This paper was highly cited since its publication, as it is the only one that presented this unique bat coronavirus strain, which serves as support for a potential viral zoonotic transmission from bat to human.

However, after careful reading into this paper, the origin, identification and characterization of BatCoV RaTG13 strain emerges as outstanding questions. Some experimental methodology, data quality, and experimental procedures described in this paper are concerning and warrant further validation as well.

The mysterious origin of bat coronavirus strain RaTG13

Shi's Nature paper[1] stated that "We then found that a short region of RNA-dependent RNA polymerase (RdRp) from a bat coronavirus (BatCoV RaTG13)—which was previously detected in

doi:10.20944/preprints202006.0044.v1

Rhinolophus affinis from Yunnan province—showed high sequence identity to 2019-nCoV. We carried out full-length sequencing on this RNA sample (GISAID accession number EPI_ISL_402131). Simplot analysis showed that 2019-nCoV was highly similar throughout the genome to RaTG13 (Fig. 1c), with an overall genome sequence identity of 96.2%."

According to the information on GISAID regarding RaTG13, this bat coronavirus strain was collected in July 2013, nearly 7 years ago. Zhengli Shi's previous publications related to bat coronaviruses have identified a total of 365 bat coronavirus strains from Yunnan province, from a total of 1981 bat samples[2-6]. However, all these publications did not mention this unique strain RaTG13 despite high-profile publications describing single coronavirus discoveries[2]. In addition, Shi's previous study[5,6] of BatCoV RdRp in 2016 did not report about this RaTG13 strain, but highlighted another bat coronavirus strain, BtCoV/4991, which was also identified in the same bat species of *Rhinolophus affinis*[5].

What is most unusual is that the short region of RdRp gene that was used to distinguish different lineages of bat coronaviruses in the phylogenetic analysis[5] showed 100% nucleotide identity between BtCoV/4991 and RaTG13[7]. This raises the serious question whether RaTG13 and BtCoV/4991 are the same strain, as this 100% identity is not at the amino acid level, but at nucleotide level. If these two were indeed two separate strains of bat coronaviruses, then Shi's group should also report, or even first find out, that BtCoV/4991 showing high similarity with SARS-CoV-2 RdRp, as BtCoV/4991 RdRp sequence was previously sequenced and submitted to GenBank (Accession number KP876546) in 2016. And if they are the same strain, what was the rationale to designate two separate names to the same thing?

In addition, through GenBank blast analysis, we found that BtCoV/4991 partial RdRp gene sequence has 98.65-98.92% nucleotide homology with that of SARS-CoV-2 strains, but only 87% homology with two other bat SARS-like coronavirus strains (bat-SL-CoVZXC21, GenBank: MG772934.1; bat-SL-CoVZC45, GenBank: MG772933.1) that have relatively high genome homology (89%) with SARS-CoV-2. Therefore, this unique feature of very high homology in RdRp gene should warrant more studies on BtCoV/4991 strain, yet this strain was not mentioned at all in this Nature paper[1].

Currently, RaTG13 is the only BatCoV that shows as high as 96% full genome homology with SARS-CoV-2, as BtCoV/4991 full genome sequence was not available. The sequence of RaTG13 was a total outlier in the phylogenetic analysis when compared to other bat coronaviruses. Jiumeng Sun et al.[8] found that in the maximum likelihood phylogenetic analysis a middle segment of the SARS-CoV-2 genome (from nt 13522 to nt 23686) and RaTG13 does not cluster with Sarbecovirus, a subgenus of the betacoronavirus that bat SARS-like coronaviruses belong to. Lam et al.[9] found six pangolin coronavirus sequences with 85.5% to 92.4% homology to SARS-CoV-2, which is less than the 96% homology that RaTG13 has. However, the pangolin coronavirus spike protein sequences shared five key amino acids in the receptor-binding domain (RBD) with that of SARS-CoV-2, while RaTG13 only shared one key amino acid in its RBD with SARS-CoV-2. Meanwhile, only SARS-CoV-2 has a G/C rich polybasic furin-cleavage site at the S1/S2 junction in the spike protein, while RaTG13 or pangolin coronaviruses do not have this cleavage site. No mention of this polybasic cleavage site was reported in Shi's Nature paper[1], despite it being a major feature that differentiates SARS-CoV-2 from RaTG13 spike proteins. Therefore, these unique features of the RaTG13 strain sequence make the SARS-CoV-2 origin and immediate animal reservoir issues deeply compounded.

As it is a routine task of Shi's lab to study bat coronavirus spike proteins and their RBD domains, it is odd that RaTG13 or BtCoV/4991 was allegedly not pursued by the Shi group for a period of nearly seven years, to further characterize their S proteins. In 2013, her group published an article in Nature describing the discovery of two bat coronaviruses, Rs3367 and SHC014, that share considerable sequence similarities in the RBD region with SARS[2]. However, while studies with RBD motifs in Rs3367 and SHC014 made breakthroughs in coronavirus research, it is rather unusual that RaTG13 with unique features failed to trigger any interest within the Shi group so far. Even in the most recent 2020 publication[10] by Shi's group, SHC014 and other bat SARS-like coronaviruses, but not RaTG13 or BtCoV/4991, were used to study the interaction among S proteins and bat ACE2 variants, while making pseudoviruses with the S protein from RaTG13 or BtCoV/4991 would not be challenging for her lab.

The only functional characterization experiment related to RaTG13 was a structure study[11] using the synthetic S protein based on the RaTG13 S gene sequence in the GenBank (accession number QHR63300.2), not using any RaTG13 virus sample. In addition, the affinity of S protein from RaTG13 with human ACE2 needs to be characterized as well, since affinity of human ACE2 is 4.5 times higher for SARS-CoV-2 than for SARS-CoV[12]. If the affinity of RaTG13 S protein is similar to SARS CoV-2, this suggests human transmission is readily possible from a RaTG13 "sister virus" (or progenitor virus of SARS-CoV-2). If RaTG13 S protein affinity is much lower, this suggests significant adaptation could have been required between the progenitor virus and SARS-CoV-2 to gain tight human ACE2 interaction capacity.

In essence, only the genome sequence of RaTG13 has been made available so far. Key information related to the identification and isolation of this RaTG13 virus strain are missing in Shi's Nature paper[1]. A series of important questions regarding RaTG13 still remain to be answered: What bat tissue/organ samples were collected in 2013 and then subjected to viral isolation or sequencing to obtain this BatCoV RaTG13? Did RaTG13 cause diseases in this bat species of *Rhinolophus affinis?* Was the bat sample for RaTG13 collected in the same cave as that of BtCoV/4991?

In addition, this Nature paper[1] did not mention any efforts to rule out the possibilities that the RaTG13-related bat sample collected in 2013 might have been mixed with other bat samples, or the bat was infected with two different strains of coronaviruses, as co-infections with two or more strains of coronaviruses in bats was not a rare event[5, 13]. If the current RaTG13 sequence identified was indeed a mixed sequence of two strains of bat coronavirus, due to random sequencing of bat samples without prior isolation of the virus, then all the current phylogenetic analyses involving RaTG13 were futile and subjected to correction.

Considering the above-mentioned strange features in RaTG13 genome as well as the fact that no prior or recent studies have been conducted using live viral stocks of this virus, concerns regarding the history and existence of the BatCoV RaTG13 strain is reasonable and legitimate. This is an important issue because the existence of this unique BatCoV RaTG13 is significantly involved in the analyses of the evolutionary relationship among SARS-CoV-2, bat SARS-like coronaviruses and other pangolin coronaviruses. If the authors had collected these two specimens (for RaTG13 & BtCoV/4991) that were outliers to other bat coronaviruses in one location in Yunnan, it would suggest that maybe other closely related sister viruses also exist in the same region. This is critical to know as these SARS-CoV-2 “sister viruses” could pose significant threat of another global pandemic and could provide key information on the evolution origin of the SARS-CoV-2 virus. Furthermore, additional virus sequences in the SARS-CoV-2, RaTG13, or BtCoV/4991 sublineage would aid understanding of the origin of the insertion between S1 and S2 of a furin cleavage site in SARS-CoV-2 that is associated with increased pathogenicity[14].

Under the openness and ethics guidelines for scientific publications, particularly in Nature and given the magnitude of the pandemic’s impact, the Shi’s team has the obligation to provide samples of RaTG13 & BtCoV/4991 for other scientists to conduct independent verification experiments and further characterization of this RaTG13 or BtCoV/4991 virus strain.

Concerns on methodology, data quality and experiment procedures

Shi’s Nature paper[1] mentioned that “Of the 10,038,758 total reads—of which 1,582 total reads were retained after filtering of reads from the human genome—1,378 (87.1%) sequences matched the sequence of SARSr-CoV (Fig. 1a). By de novo assembly and targeted PCR, we obtained a 29,891-base-pair CoV genome that shared 79.6% sequence identity to SARS-CoV BJ01 (GenBank accession number AY278488.2). High genome coverage was obtained by remapping the total reads to this genome (Extended Data Fig. 1). This sequence has been submitted to GISAID (https://www.gisaid.org/) (accession number EPI_ISL_402124).”

However, the methods described here to obtain full SARS-CoV-2 sequence have major flaws. First, characterization of novel viruses from patient samples using next-generation sequencing (NGS) technology must overcome the challenges posed by the high degree of genetic diversity observed across most virus families, especially for RNA viruses. Due to the error rate in RNA viral replication, what existed in a patient sample are usually viral quasispecies or mixed populations. Therefore, the method of random sequencing plus de novo assembly used in this study should only be used as the initial characterization. What is needed is to redo the NGS using the isolated virus stock so that the volume of raw reads related to target sequence would be significantly enhanced[15]. Then, reference assembly (using a SARS-CoV strain or Bat SARS-like CoV strain as reference) can be applied to obtain comprehensive coverage of the full viral genome with ample depth. Therefore, in this study, using only 1,378 reads from random amplification to conduct de novo assembly and getting near complete genome coverage (as shown in Extended Data Fig.1) for such a large RNA virus genome (near 30K in total length) is beyond miracle. Meanwhile, for regions with high chance of mutations such as spike protein open reading frame, very deep coverage of raw reads is often needed to ensure the accuracy of sequencing data. Random sequencing from patient samples would not work for genome regions with high variations, yet the paper did not mention any extra efforts being applied to address such concerns. The accuracy of the full genome sequences obtained in this study should be seriously challenged.

In the "Extended Data Fig. 6: Isolation and antigenic characterization of 2019-nCoV" of this paper[1], it did mention carrying out "metagenomics analysis of supernatants from Vero E6 cell cultures". Then, if NGS was conducted using viral supernatants from cell culture, the authors need to explain why those sets of NGS data was not presented in the paper, but instead NGS data from random sequencing of patient samples was presented. Therefore, the SARS-CoV-2 genome sequence submitted to the GISAID (accession number EPI_ISL_402124) in this study needs to be verified for its quality.

Meanwhile, Shi's Nature paper[1] mentioned that "four more full-length genome sequences of 2019-nCoV (WIV02, WIV05, WIV06 and WIV07) (GISAID accession numbers EPI_ISL_402127–402130) that were more than 99.9% identical to each other were subsequently obtained from four additional patients using next-generation sequencing and PCR (Extended Data Table 2)." However, since these full-length genome sequences were obtained without isolating the viruses from the patient samples (as explained in the footnote for Table 2)[1], the quality of the genome sequences could be compromised. To ensure the quality of data submitted to GISAID, Zhengli Shi team should provide raw NGS reads to open platforms so that other scientists could review and re-analyze the raw sequencing data related to these important COVID-19 patient samples.

Meanwhile, on the official website of Wuhan Institute of Virology, the Institute leadership published an open letter to all its staffs and graduate students on February 17, 2020. In this open letter, it stated that the SARS-CoV-2 strain was isolated on January 5, 2020 and its full genome sequence was obtained as early as January 2, 2020[16]. However, regarding the procedures for viral isolation, the Nature paper[1] also described in the Method section that "the culture supernatant was examined for the presence of virus by qRT–PCR methods developed in this study". This indicated that the viral isolation could NOT be initiated at the same time with genome sequencing experiments from patient samples, because the qRT-PCR experiments would need specific primers and probes that could only be designed and produced after the genome sequencing was completed. This would suggest that Shi's team only had as little as 2-3 days to obtain the viral isolate. Therefore, the experiment procedures described in this paper were surely rushed and needs further validation.

Conclusion:

In summary, as a study that isolated first SARS-CoV-2 strain and identified a bat coronavirus with high homology to SARS-CoV-2, it is critical that all data relating to viral genomes are of top quality, since many studies used or might use them as reference sequences. Meanwhile, although the leadership in Wuhan Institute of Virology might highlight their impressive speed to complete all related experiments described in the paper[16], the accurate patient sample collection date and sequencing data with better quality needs to be recorded in the scientific paper.

In addition, there have been no studies on RaTG13's infectivity in bat/human cells or in animal models, its interactions with antibodies or antiviral drugs. There is lack of understandings on RaTG13's virulence, transmissibility, pathogenicity, immune epitopes, immune evasion mechanism, etc. This was because WIV did not isolate the RaTG13 virus and does not have any related viral stocks, if the statement from Dr. Yanyi Wang (the director of WIV) in a recent TV interview[17] was accurate.

Therefore, a careful examination of the related RaTG13 samples and raw data sets of its genome sequencing are warranted to exclude any possibilities of errors or the potential co-infection of two different strains of coronavirus. And the authors need to clearly explain the relationship between RaTG13 and BtCoV/4991, whether they were the same strain or two closely related strains.

This paper was rushed to make a premature connection between bat coronavirus and SARS-CoV-2, drawing a potential bat origin scenario to support SARS-CoV-2 zoonotic transmission from bat to human. However, this connection was based on a potential bat coronavirus strain RaTG13, that may not truly exist, considering its key information missing: such as no related bat sample description, no sequencing procedure details published, confusion/identity issue with BtCoV/4991 strain, unusual sequence features, no viral isolation and related characterization, et al.

In light of these concerns, we call for the retraction of this Nature paper[1] to further verify the sequencing data, patient sample collection date and provide more information regarding the origin, identification and characterization of this BatCoV RaTG13. Proper verification should involve Dr. Zhengli Shi sending the RaTG13 and BtCoV/4991-related bat samples to other non-collaborating laboratories to be analyzed independently. And this Nature paper[1] should be cautious on making the "probable bat origin" hypothesis before RaTG13 existence could be confirmed.

Reference:

1. Zhou, P., Yang, X., Wang, X. et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579, 270–273 (2020). https://doi.org/10.1038/s41586-020-2012-7
2. Ge, X., Li, J., Yang, X. et al. Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. Nature 503, 535–538 (2013). https://doi.org/10.1038/nature12711

3. Luo Y, Li B, Jiang RD, et al. Longitudinal Surveillance of Betacoronaviruses in Fruit Bats in Yunnan Province, China During 2009-2016. Virol Sin. 2018;33(1):87-95. https://doi:10.1007/s12250-018-0017-2
4. Hu B, Zeng LP, Yang XL, et al. Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus. PLoS Pathog. 2017;13(11): e1006698. Published 2017 Nov 30. https://doi:10.1371/journal.ppat.1006698
5. Ge, X., Wang, N., Zhang, W. et al. Coexistence of multiple coronaviruses in several bat colonies in an abandoned mineshaft. Virol. Sin. 31, 31–40 (2016). https://doi.org/10.1007/s12250-016-3713-9
6. Wang MN, Zhang W, Gao YT, et al. Longitudinal surveillance of SARS-like coronaviruses in bats by quantitative real-time PCR. Virol Sin. 2016;31(1):78-80. https://doi:10.1007/s12250-015-3703-3
7. Rahalkar, M.C.; Bahulikar, R.A. Understanding the Origin of BatCoVRaTG13, a Virus Closest to SARS-CoV-2. Preprints 2020, 2020050322. https://doi:10.20944/preprints202005.0322.v1
8. Sun, J., He, W. T., Wang, L., Lai, A., Ji, X., Zhai, X., Li, G., Suchard, M. A., Tian, J., Zhou, J., Veit, M., & Su, S. (2020). COVID-19: Epidemiology, Evolution, and Cross-Disciplinary Perspectives. Trends in molecular medicine, 26(5), 483–495. https://doi.org/10.1016/j.molmed.2020.02.008
9. Lam, T.T., Shum, M.H., Zhu, H. et al. Identifying SARS-CoV-2 related coronaviruses in Malayan pangolins. Nature (2020). https://doi.org/10.1038/s41586-020-2169-0
10. Guo H, Hu B, Shi Z, et al. Evolutionary arms race between virus and host drives genetic diversity in bat SARS related coronavirus spike genes. bioRxiv 2020.05.13.093658; https://doi.org/10.1101/2020.05.13.093658
11. Shang, J., Ye, G., Shi, K. et al. Structural basis of receptor recognition by SARS-CoV-2. Nature 581, 221–224 (2020). https://doi.org/10.1038/s41586-020-2179-y
12. Walls A, Park Y, M. Veesler A, et al. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. Cell, Volume 181, Issue 2, 2020, 281-292.e6, https://doi.org/10.1016/j.cell.2020.02.058.
13. Drexler J, Gloza-Rausch F, Drosten C, et al. Genomic Characterization of Severe Acute Respiratory Syndrome-Related Coronavirus in European Bats and Classification of Coronaviruses Based on Partial RNA-Dependent RNA Polymerase Gene Sequences. Journal of Virology Oct 2010, 84 (21) 11336-11349; https://DOI:10.1128/JVI.00650-10
14. Coutard B, Valle C, de Lamballerie X, Canard B, Seidah NG, Decroly E. The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade. Antiviral Res. 2020;176:104742. https://doi:10.1016/j.antiviral.2020.104742

15. Park WB, Kwon NJ, Oh MD, et al. Virus Isolation from the First Patient with SARS-CoV-2 in Korea. J Korean Med Sci. 2020 Feb 24; 35(7): e84. https://doi:10.3346/jkms.2020.35.e84
16. “致全所职工和研究生的一封信 - 中国科学院武汉病毒研究所”, February 19, 2020. http://www.whiov.ac.cn/tzgg_105342/202002/t20200219_5502325.html
17. Exclusive: Dir. of Wuhan Institute of Virology on COVID-19. CGTN, May 24, 2020. https://www.youtube.com/watch?v=bRuzsPA4Ukw