Identification of the hyper-variable genomic hotspot for the novel coronavirus SARS-CoV-2

Feng Wen , Hai Yu , Jinyue Guo , Yong Li , Kaijian Luo , Shujian Huang

PII: S0163-4453(20)30108-0
DOI: https://doi.org/10.1016/j.jinf.2020.02.027
Reference: YJINF 4470

To appear in: *Journal of Infection*

Accepted date: 29 February 2020

Identification of the hyper-variable genomic hotspot for the novel coronavirus SARS-CoV-2

Feng Wen[#*1], Hai Yu[*2,5], Jinyue Guo[1], Yong Li[3], Kaijian Luo[4], Shujian Huang[#1]

1. College of Life Science and Engineering, Foshan University, Foshan, 528231, Guangdong, China
2. Shanghai Veterinary Research Institute, Chinese Academy of Agricultural Sciences, Shanghai 200241, China.
3. College of Animal Science and Technology, Jiangxi Agricultural University, Nanchang 330045, Jiangxi, China
4. College of Veterinary Medicine, South China Agricultural University, Guangzhou 510642, Guangdong, China
5. Jiangsu Co-innovation Center for Prevention and Control of Important Animal Infectious Diseases and Zoonoses, Yangzhou 225009, China

#Corresponding address to: Feng Wen, wenfengjlu@163.com, wenf@fosu.edu.cn or Shujian Huang, 617955368@qq.com; College of life science and engineering, Foshan University, No33 guangyun road, Shishan town, Nanhai district, Foshan, 528231, Guangdong, China.

*These authors contribute equally to this work.

Dear editor,

A recent study in this journal studied the genomes of the novel SARS-like coronavirus (SARS-CoV-2) in China and suggested that the SARS-CoV-2 had undergone genetic recombination with SARS-related CoV[1]. By February 14, 2020, a total of 66,576 confirmed cases of COVID-19, people infected with SARS-CoV-2, were reported in China, leading to 1,524 deaths, per the Chinese CDC (http://2019ncov.chinacdc.cn/2019-nCoV/). Several full genomic sequences of this virus have been released for the study of its evolutionary origin and molecular characteristics[2-4]. Here, we analyzed the potential mutations that may have evolved after the virus became epidemic among humans and also the mutations resulting in the human adaptation.

The sequences of BetaCoV were downloaded on February 3, 2020 from the GISAID platform [5]. A total of 58 accessions were available, among which BetaCoV/bat/Yunnan/RaTG13/2013 is a known close relative of SARS-CoV-2. Four accessions, namely, BetaCov/Italy/INM1/2020, BetaCov/Italy/INM2/2020, BetaCoV/Kanagawa/1/2020, and BetaCoV/USA/IL1/2020, were excluded because of the short-truncated sequences or multiple ambiguous nucleotides. A total of 54 accessions (Supplementary table 1) isolated from humans were utilized in the following analysis. The sequences NC_004718.3 of SARS coronavirus [6] genes were utilized to define the protein products of SARS-CoV-2. The protein sequences of ORF1ab, S, E, M, and N genes were translated, and all of the loci without experimental evidences were excluded. First, the protein sequences of SARS-CoV-2 were compared with RaTG13, human SARS (NC_004718.3), bat SARS (DQ022305.2), and human MERS (NC_019843.3) by calculating the similarity in a given sliding window (Figure 1A). The sliding window was set to 500 for ORF1ab and S, and to 50 for proteins E, M, and N considering their short length. SARS-CoV-2 were highly similar to RaTG13 isolated from bats, showing 96% identity based on the whole-nucleotide sequences and 83% based on the protein sequences, suggesting a bat zoonotic origin of SARS-CoV-2. ORF1a, and the head of S seemed to have diverged from other beta coronaviruses.

The molecular phylogenetic tree (Figure 1B) was built by using the maximum likelihood method based on the JTT matrix-based model [7]. It hinted that the protein sequences of SARS-CoV-2 had over 99% similarity. Twenty-eight viruses had shared the same protein sequences, and could be the original strain circulated in the humans. The other viruses had only a few mutations from it. This indicates that the virus could have evolved for only a very short time after gaining the efficient human to human transmissibility, as expected. Next, we analyzed the mutations that occurred after infecting humans (Figure 1C) in order to identify mutations associated with more severe infection. Here, two accessions (BetaCoV/Shenzhen/SZTH-001/2020 and BetaCoV/Shenzhen/SZTH-004/2020) from Shenzhen, which had 5 and 16 mutations, respectively, were excluded, considering the possible experimental issues. All of the mutations only occurred once, so it is possible that all of these mutations occur naturally and are associated with viral survival and infection. Several mutations were clustered in peptides nsp3 and nsp4 of ORF1ab and in the header of S. These results suggested that there had probably been no hyper-variable genomic hotspot in the SARS-CoV-2 population until now.

We compared these results with those of the work of Ceraolo and Giorgi [8], who reported at least two hyper-variable genomic hotspots based on the Shannon entropy of nucleotide sequences. They utilized all of the sequences, while we merged all of the fully identical sequences into one during our Shannon entropy calculation. As shown in Figure 1B, 28 sequences were merged into one in present study because they had been collected in such a short time, so collection time and location could not have produced any large bias. If those identical sequences were calculated individually, any mutations on these 28 sequences would have sharply increased Shannon entropy. The protein sequences were used to exclude any unimportant silent mutations. Finally, the sequences of earliest SARS-CoV-2 were compared with RaTG13 from bats (Figure 1D). Fisher’s exact test with post hoc test suggested that nsp1, nsp3, and nsp15 of ORF1ab and gene S had significantly more mutations than other genes, which might facilitate human adaptation and infection.

S gene encodes spike glycoprotein, which binds host ACE2 receptors and is required for initiation of the infection [9]. They reported that a 193-amino acid fragment was able to bind ACE2 more efficiently than its unmutated counterpart. This region in which spike glycoprotein binds to ACE2 had 21 mutations not found in RaTG13, suggesting their role in the adaptation to human hosts. Peptide nsp1 facilitated viral gene expression and evasion from the host immune response [10]. Peptide nsp3, named papain-like proteinase, was found to be associated with the cleavages, viral replication, and antagonization of innate immune. These two peptides are probably associated with the latent period after infection in humans. Peptide nsp15 acted as uridylate-specific endoribonuclease. These results collectively suggest that peptides nsp1, nsp3, and nsp15 might have unclear but critical roles in this outbreak of SARS-CoV-2.

To summarize, this study confirmed the relationship of SARS-CoV-2 with other beta coronaviruses on the amino acid level. The hyper-variable genomic hotspot has been established in the SARS-CoV-2 population at the nucleotide but not the amino acid level, suggesting that there have been no beneficial mutations. The mutations in nsp1, nsp3, nsp15, and gene S that identified in this study would be associated with the SARS-CoV-2 epidemic and was worthy of further study.

Funding information

This study was supported by Key Laboratory for Preventive Research of Emerging Animal Diseases in Foshan University (KLPREAD201801-06), (KLPREAD201801-10), Youth Innovative Talents Project of Guangdong Province (2018KQNCX280), and National Key Research and Development Project (grant No.2017YFD0500800)

Acknowledgement

We thank the researchers who deposited the SARS-CoV-2 sequences in the GISAID. We thank LetPub for its linguistic assistance during the preparation of this manuscript.

Conflict of interest

All authors declare no conflict of interest.

Reference

1. Zhang, J., et al., *The continuous evolution and dissemination of 2019 novel human coronavirus.* J Infect, 2020.
2. Wu, A., et al., *Genome Composition and Divergence of the Novel Coronavirus (2019-nCoV) Originating in China.* Cell Host Microbe, 2020.
3. Wu, F., et al., *A new coronavirus associated with human respiratory disease in China.* Nature, 2020.
4. Zhou, P., et al., *A pneumonia outbreak associated with a new coronavirus of probable bat origin.* Nature, 2020.

5. Shu, Y. and J. McCauley, *GISAID: Global initiative on sharing all influenza data - from vision to reality*. Euro Surveill, 2017. **22**(13).

6. Marra, M.A., *The Genome Sequence of the SARS-Associated Coronavirus.* Science, 2003. **300**(5624): p. 1399-1404.

7. Jones, D.T., W.R. Taylor, and J.M. Thornton, *The rapid generation of mutation data matrices from protein sequences.* Comput Appl Biosci, 1992. **8**(3): p. 275-82.

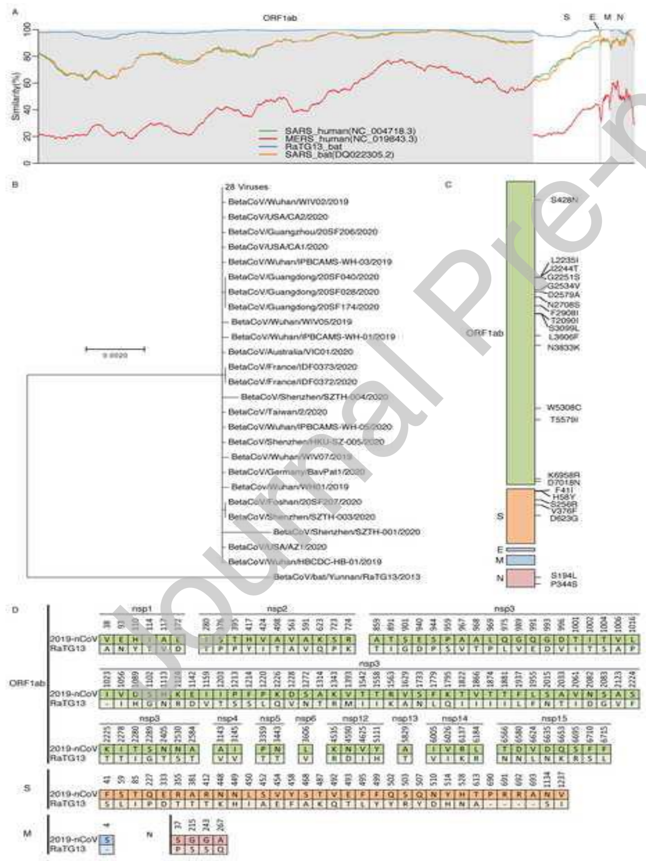
8. Ceraolo, C. and F.M. Giorgi, *Genomic variance of the 2019‐ nCoV coronavirus.* Journal of Medical Virology, 2020.

9. Wong, S.K., et al., *A 193-amino acid fragment of the SARS coronavirus S protein efficiently binds angiotensin-converting enzyme 2.* J Biol Chem, 2004. **279**(5): p. 3197-201.

10. Lokugamage, K.G., et al., *Severe Acute Respiratory Syndrome Coronavirus Protein nsp1 Is a Novel Eukaryotic Translation Inhibitor That Represses Multiple Steps of Translation Initiation.* Journal of Virology, 2012. **86**(24): p. 13598-13608.

Figure legends

Figure 1. (A) The similarity between SARS-CoV-2 and other beta coronaviruses using the sliding window showed that SARS-CoV-2 was similar to bat virus RaTG13. (B) The molecular phylogenetic tree based on protein sequences established the high similarity among SARS-CoV-2 and its near relatives. (C) The mutations that developed after it came to circulate among humans did not include any mutation with high occurrence. (D) The graphs show all of the differences between SARS-CoV-2 and its close relative strains isolated from bats.


A
ORF1ab
S
E
M
N
Similarity(%)
SARS_human(NC_004718.3)
MERS_human(NC_019843.3)
RaTG13_bat
SARS_bat(DQ022305.2)
B
28 Viruses
BetaCoV/Wuhan/WIV02/2019
BetaCoV/USA/CA2/2020
BetaCoV/Guangzhou/20SF206/2020
BetaCoV/USA/CA1/2020
BetaCoV/Wuhan/IPBCAMS-WH-03/2019
BetaCoV/Guangdong/20SF040/2020
BetaCoV/Guangdong/20SF028/2020
BetaCoV/Guangdong/20SF174/2020
BetaCoV/Wuhan/WIV05/2019
BetaCoV/Wuhan/IPBCAMS-WH-01/2019
BetaCoV/Australia/VIC01/2020
BetaCoV/France/IDF0373/2020
BetaCoV/France/IDF0372/2020
BetaCoV/Shenzhen/SZTH-004/2020
BetaCoV/Taiwan/2/2020
BetaCoV/Wuhan/IPBCAMS-WH-05/2020
BetaCoV/Shenzhen/HKU-SZ-005/2020
BetaCoV/Wuhan/WIV07/2019
BetaCoV/Germany/BavPat1/2020
BetaCoV/Wuhan/WH01/2019
BetaCoV/Foshan/20SF207/2020
BetaCoV/Shenzhen/SZTH-003/2020
BetaCoV/Shenzhen/SZTH-001/2020
BetaCoV/USA/AZ1/2020
BetaCoV/Wuhan/HBCDC-HB-01/2019
BetaCoV/bat/Yunnan/RaTG13/2013
C
ORF1ab
S428N
L2235I
I2244T
G2251S
G2534V
D2579A
N2708S
F2908I
T2090I
S3099L
L3606F
N3833K
W5308C
T5579I
K6958R
D7018N
S
F41I
H58Y
S256R
V376F
D623G
E
M
N
S194L
P344S
D
ORF1ab
nsp1
nsp2
nsp3
nsp4
nsp5
nsp6
nsp12
nsp13
nsp14
nsp15
2019-nCoV
RaTG13
S
M
N