The emergence of SARS-CoV-2 by an unusual genome reconstitution

Seong-Tshool Hong (✉ seonghong@jbnu.ac.kr)

1. Department of Biomedical Sciences and Institute for Medical Science, Jeonbuk National University Medical School, Jeonju, Jeonbuk 54907, South Korea.

Md. Mehedi Hassan

1. Department of Biomedical Sciences and Institute for Medical Science, Jeonbuk National University Medical School, Jeonju, Jeonbuk 54907, South Korea. 2. JINIS BDRD institute, JINIS Biopharmaceuticals Inc., 224 Wanjusandan 6-Ro, Bongdong, Wanju, Jeonbuk 55315, South Korea.

Shirina Sharmin

1. Department of Biomedical Sciences and Institute for Medical Science, Jeonbuk National University Medical School, Jeonju, Jeonbuk 54907, South Korea.

Jinny Hong

3. SNJ Pharma Inc., 1124 W. Carson St, MRL Bldg 3F, Torrance, CA 90502, USA.

Hoi-Seon Lee

4. Department of Bioenvironmental Chemistry, Jeonbuk National University, Jeonju, Jeonbuk 54896, South Korea.

Hyeon Jin Kim

2. JINIS BDRD institute, JINIS Biopharmaceuticals Inc., 224 Wanjusandan 6-Ro, Bongdong, Wanju, Jeonbuk 55315, South Korea.

Short Report

Keywords: SARS-CoV-2, prototype, genome, phylogenetic analysis.

Posted Date: June 5th, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-33201/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

SARS-CoV-2 has been spreading remarkably fast around the world since its emergence while the origin of the virus remains ambiguous. Here, we constructed all of the original prototype genome sequences of SARS-CoV-2 by selecting the common nucleotide among the different virus strains with species. Phylogenetic analysis on the prototype sequences showed that SARS-CoV-2 was a direct descendant of Bat-CoV and was closely related to Pan-CoV, Bat-SL-CoV, and SARS-CoV. The pairwise comparison of SARS-CoV-2 with Bat-CoV showed an unusual replacement of the motif consisting of 7 amino acids within the spike protein of SARS-CoV-2. Database searches showed that the motif originated from a surface protein of *Plasmodium malariae*, suggesting that the SARS-CoV-2 was emerged after acquiring the motif of the malaria surface protein.

Main Text

Coronaviruses (CoVs) are prevalent causes of the common cold widely present in nature with a broad spectrum of hosts. Although they frequently infect humans, the natural host of CoVs are animals and thereby all of the CoVs for the common colds, HCoV-229E, HCoV-NL63, HCoV-OC43, and HCoV-HKU1, have zoonotic origins (1). The genomes of CoVs consist of a single-stranded, positive RNA of 26,000 to 32,000 base pairs and a variable number (from 6 to 11) of open reading frames (2). Because of their unusually large size with the characteristics of the RNA genomes, CoVs can frequently mutate to escape their natural hosts, causing serious diseases to humans. Outbreaks of SARS-CoV in 2003 and MERS-CoV in 2012 are well-known examples. Currently, another very serious pathogenic novel coronavirus, SARS-CoV-2, has emerged and caused a global pandemic.

The genome sequence of SARS-CoV-2 (3, 4) suggested that SARS-CoV-2 stemmed from coronaviruses of bat or pangolin (5, 6). Because of the ambiguity of the origin of SARS-CoV-2, we adopted a different approach to investigate the origin of SARS-CoV-2. First, we pooled out all of the available genome sequences of SARS-CoV-2 as well as its related viral species by the BLAST sequence similarity from all of the publicly available genome databases (Table S1). Since genetic mutation or recombination occurs on an individual basis and these events cannot occur at the same position repeatedly, the prototype sequence of the genome within a species is always the most prevalent base among the members. Therefore, we constructed the original prototype aa sequences of SARS-CoV-2 and its related species, Bat-CoV, Pan-CoV, Bat-SL-CoV, and SARS-CoV (Fig. S1 & Table S2). The prototype sequences for each viral species were determined by choosing the most prevalent nucleotide bases in each position within the same species from pairwise comparison. The prototype sequences of the five viral species were directly used for phylogenetic analysis by the neighbor-joining method using Unipro UGENE bioinformatics toolkits (7). As shown in Figure-1A, SARS-CoV-2 was a direct descendant of Bat-CoV and was closely related to Pan-CoV, Bat-SL-CoV, and SARS-CoV. The sequence alignment of the prototype amino acid sequences by using multiple sequence alignments (MSA) using MAFFT program (8) showed that the amino acid (aa) sequence of SARS-CoV-2 shared 98.67% sequence similarity with Bat-CoV while 94.51%, 94.35% and 82.93% of the sequences of SARS-CoV-2 were identical with that of Pan-CoV, Bat-

SL-CoV, and SARS-CoV respectively (Table 1). The aa sequence similarities were well-matched with their phylogenetic distances. The zoonotic origin of SARS-CoV-2 is under debate to be either Bat-CoV or Pan-CoV. The current argument on the origin of SARS-CoV-2 would not be surprising since phylogenetic analyses can generate different results if individual genome sequences were used. Considering the sequence variability of RNA genomes as in the case of SARS-CoV-2, we believe that the prototype sequence analysis approach would generate a more precise result than conventional approaches in phylogenetic analysis for species consisting of RNA genome.

To investigate how SARS-CoV-2 originated, we compared pairwise differences between the prototype aa sequences of SARS-CoV-2 and Bat-CoV. The pairwise aa sequence analysis showed that both of the two virus genomes encode 14 genes (Fig. 1B). The aa sequence mismatches were randomly distributed except for the consecutive 7 aa alteration (439 NNLDISKV 445) in the spike protein. Since genetic mutation is a random process, the mutated points are expected to be distributed randomly throughout the genome. Considering this nature of mutations, the consecutive 7-aa sequence difference in the spike protein was peculiar, indicating that an unusual event happened during the emergence of SARS-CoV-2 from Bat-CoV. Hence, 3D models of the spike proteins were made to analyze the consequence of alteration to the structure of the spike protein during the emergence of SARS-CoV-2. The spike protein of coronavirus (CoV) is the surface protein that binds to a receptor on the host cell surface. The spike protein consists of three large domains: a large ectodomain, a single-pass transmembrane anchor, and a short intracellular tail (9). The ectodomain is further divided into a receptor-binding subunit S1 and a membrane-fusion subunit S2. Coronavirus first binds to a receptor on the host cell surface through its S1 subunit and then fuses the viral and the host membranes through its S2 subunit, meaning that the S1 domain plays the most critical role for the virus's invasion into its host (10, 11). As shown in Figure-1C, altering the consecutive 7-aa on the outer layer of the spike protein's S1 domain did not affect the overall structural integrity of the spike protein. The 3D modeling also showed that the consecutive 7-aa alteration resulted in a new motif to occupy more space in the S1 domain than its original structure by transforming a partial α -helical structure to a random coil. Since the S1 domain acts as the binding domain for the entry of the virus, enlargement of its space as well as adaptation of different conformation in the outermost layer surface of the S1 domain would endow SARS-CoV-2 to expand its host range.

The unusual pattern of genetic alteration motivated us to investigate the origin of the 7-aa motif (439 NNLDISKV 445). To trace down its origin, we overlappingly defragmented the aa sequence of the spike protein of SARS-CoV-2 into 9 aa units. The fragmented aa sequence matched with all of the available protein sequences. All of the aa sequences matched only with the spike proteins of coronaviruses except for the 7-aa motif. Very surprisingly, the NNLDISKV motif perfectly matched only with a motif of a surface protein (GenBank Accession : XP_028861348.1) of *Plasmodium malariae* which causes malaria to humans. The NNLDISKV motif was located at aa 449-455 of a membrane-bound surface protein of *P. malariae* (Fig. 2).

The clinical presentation of COVID-19 very differs from that of other coronaviruses (12). The acquisition of the *P. malariae* gene in the spike protein seems to explain why. *Plasmodium* species, which is

responsible for malaria, infect liver cells and erythrocytes to promote blood clotting and damages in the heart, liver, or kidney. Unlike other coronaviral infections, it is very interesting to note that symptoms similar to the case of the *P. malariae* infection were reported in COVID–19. Red cell distribution width (RDW) of erythrocytes has been reported to be enlarged after the malarial invasion because the growth of the malaria parasite causes the cells to be enlarged (13, 14). Foy et al., recently reported the observation of the same phenomenon in COVID–19 that elevated RDW is correlated with increased mortality risk in COVID–19 (15). Not only is the enlargement of erythrocytes in COVID–19 correlated with malaria infection, but also clinical presentations such as blood clotting and damages in the heart, liver, or kidney were also observed in COVID–19 (16–18). Furthermore, the acquisition of the *P. malariae* gene seems to explain why hydroxychloroquine, an anti-malaria drug for *P. malariae* infection, shows some efficacy in COVID–19 (18–20). Considering these facts, the investigation of different kinds of anti-malaria drugs for COVID–19 treatment needs to be pursued.

According to the current outbreak statistic of COVID–19 (<https://www.worldometers.info/coronavirus/>), a racial background affects the morbidity and mortality of COVID–19. It seems to be that the morbidity and mortality by COVID–19 are much more serious in the Western world than in Eastern countries. Multisystem inflammatory syndrome in children never reported and the morbidity and mortality by COVID–19 are much milder in Eastern countries. It would be very interesting to investigate the role of the NNLSKV motif among different racial backgrounds including ACE–2 since the genetic polymorphism of ACE–2 is known to differ among different ethnic and racial groups (21).

References & Notes

1. W. Ye, S. Yuan, K. S. Yuen, S. Y. Fung, C. P. Chan, D. Y. Jin, Zoonotic origins of human coronaviruses. *Int. J. Biol. Sci.* **16**, 1686-1697 (2020).
2. Song, Y. Xu, L. Bao, L. Zhang, P. Yu, Y. Qu, H. Zhu, W. Zhao, Y. Han, C. Qin, From SARS to MERS, thrusting coronaviruses into the spotlight. *Viruses*. **11** (2019), p. E59
3. Wu, Y. Peng, B. Huang, X. Ding, X. Wang, P. Niu, J. Meng, Z. Zhu, Z. Zhang, J. Wang, J. Sheng, L. Quan, Z. Xia, W. Tan, G. Cheng, T. Jiang, Genome composition and divergence of the novel coronavirus (2019-nCoV) originating in China. *Cell Host Microbe*. **27**, 325-328 (2020).
4. Wu, S. Zhao, B. Yu, Y. M. Chen, W. Wang, Z. G. Song, Y. Hu, Z. W. Tao, J. H. Tian, Y. Y. Pei, M. L. Yuan, Y. L. Zhang, F. H. Dai, Y. Liu, Q. M. Wang, J. J. Zheng, L. Xu, E. C. Holmes, Z. Zhang, Author Correction: A new coronavirus associated with human respiratory disease in China. *Nature*. **580** (7803):E7 (2020).
5. G. Andersen, A. Rambaut, W. I. Lipkin, E. C. Holmes, Garry RF. The proximal origin of SARS-CoV-2. *Nat. Med.* **26**, 450-452 (2020).
6. T. Lam, M. H. Shum, H. C. Zhu, Y. G. Tong, X. B. Ni, Y. S. Liao, W. Wei, W. Y. Cheung, Li, L. F. Li, G. M. Leung, E. C. Holmes, Y. L. Hu, Y. Guan, Identifying SARS-CoV-2 related coronaviruses in Malayan pangolins. *Nature* (2020) DOI: 10.1038/s41586-020-2169- 0.

7. Okonechnikov, O. Golosova, M. Fursov, the UGENE team. Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics*. **28**: 1166-1167 (2012).
8. Katoh, K. Misawa, K. Kuma, T. Miyata, MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).
9. Lan, J. Ge, J. Yu, S. Shan, H. Zhou, S. Fan, Q. Zhang, X. Shi, Q. Wang, L. Zhang, X. Wang, Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature* doi: 10.1038/s41586-020-2180-5 (2020).
10. Ou, Y. Liu, X. Lei, P. Li, D. Mi, L. Ren, L. Guo, R. Guo, T. Chen, J. Hu, Z. Xiang, Z. Mu, Chen, J. Chen, K. Hu, Q. Jin, J. Wang, Qian, Characterization of spike glycoprotein of SARS-CoV-2 on virus entry and its immune cross-reactivity with SARS-CoV. *Nat. Commun.* **11**, 1620 (2020).
11. Zhang, Q. Wu, Z. Zhang, Probable Pangolin Origin of SARS-CoV-2 Associated with the COVID-19 Outbreak. *Curr. Biol.* **30**, 1578 (2020).
12. X. Lescure, L. Bouadma, D. Nguyen, M. Parisey, P. H. Wicky, S. Behillil, A. Gaymard, M. Bouscambert-Duchamp, F. Donati, Q. Le Hingrat, V. Enouf, N. Houhou-Fidouh, M. Valette, Mailles, J. C. Lucet, F. Mentre, X. Duval, D. Descamps, D. Malvy, J. F. Timsit, B. Lina, S. van-der-Werf, Y. Yazdanpanah, Clinical and virological data of the first cases of COVID-19 in Europe: a case series. *Lancet Infect Dis.* p: S1473-3099(20)30200-0 (2020).
13. S. Jairajpuri, S. Rana, M. J. Hassan, F. Nabi, S. Jetley, An analysis of hematological parameters as a diagnostic test for malaria in patients with acute febrile illness: an institutional experience. *Oman Med. J.* **29**, 12-17 (2014).
14. S. Koltas, H. Demirhindi, S. Hazar, K. Ozcan, Supportive presumptive diagnosis of Plasmodium vivax malaria. Thrombocytopenia and red cell distribution width. *Saudi Med. J.* **28**, 535-539 (2007).
15. H. Foy *et al.*, <https://www.medrxiv.org/content/10.1101/2020.05.05.20091702v1.full.pdf> (2020).
16. Lodigiani, G. Iapichino, L. Carenzo, M. Cecconi, P. Ferrazzi, T. Sebastian, N. Kucher, J. D. Studt, C. Sacco, B. Alexia, M. T. Sandri, S. Barco; Humanitas COVID-19 task force, Venous and arterial thromboembolic complications in COVID-19 patients admitted to an academic hospital in Milan, Italy. *Thromb. Res.* **191**, 9-14 (2020).
17. Rismanbaf, S. Zarei .Liver and kidney injuries in COVID-19 and their effects on drug therapy; a letter to editor. *Arch. Acad. Emerg. Med.* **8**, e17 (2020).
18. J. Guzik, S. A. Mohiddin, A. Dimarco, V. Patel, K. Savvatis, F. M. Marelli-Berg, M. S. Madhur, M. Tomaszewski, P. Maffia, F. D'Acquisto, S. A. Nicklin, A. J. Marian, R. Nosalski, C. Murray, B. Guzik, C. Berry, R. M. Touyz, R. Kreutz, D. W. Wang, D. Bhella, O. Sagliocco, F. Crea, E. C. Thomson, I. B. McInnes, COVID-19 and the cardiovascular system: implications for risk assessment, diagnosis, and treatment options. *Cardiovasc Res.* pii: cvaa 106. DOI: 10.1093/cvr/cvaa106 (2020).
19. K. Singh, A. Singh, A. Shaikh, R. Singh, A. Misra, Chloroquine and hydroxychloroquine in the treatment of COVID-19 with or without diabetes: A systematic search and a narrative review with a special reference to India and other developing countries. *Diabetes Metab. Syndr.* **14**, 241-246 (2020).

20. L. M. Arnold, F. Buckner, Hydroxychloroquine for treatment of SARS-CoV-2 infection? Improving our confidence in a model-based approach to dose selection. *Clin. Transl. Sci.*(2020). DOI: 10.1111/cts.12797.
21. Hussain, N. Jabeen, F. Raza, S. Shabbir, A. A. Baig, A. Amanullah, B. Aziz, Structural variations in human ACE2 may influence its binding with SARS-CoV-2 spike protein. *J. Med. Virol.* (2020). DOI: 10.1002/jmv.25832.

Declarations

Acknowledgments

Funding: This work was supported by the funding from JINIS BDRD Research Institute, JINIS Biopharmaceuticals Inc. (Wanju, South Korea).

Author contributions: S.T.H. conceived and designed the study and wrote the paper. M.M.H. and S.S. performed the analysis of data. J.H. wrote the paper. H.J.K. and H.S.L discussed the result and further validated the data thoroughly. All authors participated in interpretation of data.

Competing interests: The authors declare no competing interesting.

Data availability: All data are available in the main manuscript and supplementary material.

Table

Table 1. Amino acid sequence similarities among the 14 different ORFs of SARS-CoV-2 with Bat-CoV, Bat-SL-CoV, Pan-CoV and SARS-CoVs.

Gene	Bat-CoV (%)	Bat-SL-CoV (%)	Pan-CoV (%)	SARS-CoV (%)
Whole genome	98.67	94.35	94.51	82.93
ORF1a	98.62	96.94	92.90	80.44
ORF1ab	99	96.73	94.87	86.19
Spike (S)	98.04	83.52	93.87	75.96
3a	98.91	95.27	95.64	72.36
3b	77.27	81.82	90.91	61.11
Envelope (E)	100	100	100	94.74
Membrane (M)	99.52	99.05	98.57	91.43
6	100	93.44	96.72	68.85
7a	99.17	88.52	94.21	85.25
7b	97.67	93.02	90.70	85.37
8	95.87	29.73	95.87	8a: 31.71; 8b: 40.48
Nucleocapsid (N)	99.28	95.70	97.85	90.52
9b	92.78	76.29	93.81	72.45
14	90.41	92.98	96.49	84.21

Figures

Figure 1

Comparison of SARS-CoV-2 and Bat-CoVs. (A) The phylogeny of SARS-CoV-2 based on the 14 ORF and receptor-binding domain (RBD) sequences of Bat-CoV, Pan-CoV, Bat-SL-CoV, and SARS-CoV. Phylogenies were estimated by the neighbor-joining method using Unipro UGENE bioinformatics toolkits. (B) Organization of genes in SARS-CoV-2 and Bat-CoVs. The distribution of mutated amino acids in SARS-CoV-2 compare to Bat-CoV are represented with red lines. (C) Homotrimer 3D model of Bat-CoV and SARS-CoV-2 spike protein constructed using the SWISS-MODEL protein homology-modeling server. Circle showing the altered surface structure in the spike region of SARS-CoV-2 for antigen binding.

Figure 2

The culprit origin of the spike protein of SARS-CoV-2 by an unusual genome reconstitution. The 3D model for the NNLDSKV motif of SARS-CoV-2 and *Plasmodium malariae* surface protein was built by the SWISS-MODEL protein homology-modeling server and the PDB sequence was analyzed by UCSF CHIMERA software. The orientation of 7-amino acids in the NNLDSKV motif is shown in the box with different colors.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [CoV2Supplemental.pdf](#)