Genome Composition and Divergence of the Novel Coronavirus (2019-nCoV) Originating in China

Aoping Wu,1,7,9 Yousong Peng,2,8 Baoying Huang,3,9 Xiao Ding,1,7,9 Xianyue Wang,1,7 Peihua Niu,3 Jing Meng,1,7 Zhaozhong Zhu,6 Zheng Zhang,2 Jianguang Wang,1,7 Jie Sheng,1,7 Lijun Quan,9 Zanxian Xia,5,8 Wenjie Tan,3,* Genhong Cheng,6,* and Taijiao Jiang1,7,*

1Center for Systems Medicine, Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100050, China
2College of Biology, Hunan Provincial Key Laboratory of Medical Virology, Hunan University, Changsha 410082, China
3Key Laboratory of Medical Virology, National Health and Family Planning Commission, National Institute for Viral Disease Control and Prevention, China CDC, Beijing 100206, China
4School of Computer Science and Technology, Soochow University, Suzhou, China
5Department of Cell Biology, School of Life Science, Central South University, Changsha 410013, China
6Suzhou Institute of Systems Medicine, Suzhou, Jiangsu 215123, China
7Department of Microbiology, Immunology and Molecular Genetics, University of California, Los Angeles, USA
8Hunan Key Laboratory of Animal Models for Human Diseases, Hunan Key Laboratory of Medical Genetics & Center for Medical Genetics, School of Life Science, Central South University, Changsha 410013, China
9These authors contributed equally
*Correspondence: tanwj@ivdc.chinacdc.cn (W.T.), gcheng@mednet.ucla.edu (G.C.), taijiao@ibms.pumc.edu.cn (T.J.)

https://doi.org/10.1016/j.chom.2020.02.001

An in-depth annotation of the newly discovered coronavirus (2019-nCoV) genome has revealed differences between 2019-nCoV and severe acute respiratory syndrome (SARS) or SARS-like coronaviruses. A systematic comparison identified 380 amino acid substitutions between these coronaviruses, which may have caused functional and pathogenic divergence of 2019-nCoV.
identical, with only five nucleotide differences in the genome of ~29.8 kb nucleotides (Figure S1). The 2019-nCoV genome was annotated to possess 14 ORFs encoding 27 proteins (Figure 1A and Tables S1A and S1B). The orf1ab and orf1a genes located at the 5'-terminus of the genome respectively encode the pp1ab and pp1a proteins, respectively. They together comprise 15 nsps including nsp1-nsp10 and nsp12-nsp16 (Figure 1A and Table S1B). The 3'-terminus of the genome contains four structural proteins (S, E, M, and N) and eight accessory proteins (3a, 3b, p6, 7a, 7b, 8b, 9b, and orf14). At the amino acid level, the 2019-nCoV is quite similar to that of SARS-CoV, but there are some notable differences. For example, the 8a protein is present in SARS-CoV and absent in 2019-nCoV; the 8b protein is 84 amino acids in SARS-CoV, but longer in 2019-nCoV, with 121 amino acids; the 3b protein is 154 amino acids in SARS-CoV, but shorter in 2019-nCoV, with only 22 amino acids (Table S1A). Further studies are needed to characterize how these differences affect the functionality and pathogenesis of 2019-nCoV.

Figure 1. Genome composition and phylogenetic tree for 2019-nCoV
(A) Schematic diagram of the genome organization and the encoded proteins of pp1ab and pp1a for the IVDC-HB-01/2019 (HB01) strain. The largest gene, namely the orf1ab, encodes the pp1ab protein that contains 15 nsps (nsp1-nsp10 and nsp12-nsp16). The pp1a protein encoded by the orf1a gene also contains 10 nsps (nsp1-nsp10). Structural proteins are encoded by the four structural genes, including spike (S), envelope (E), membrane (M), and nucleocapsid (N) genes. The accessory genes are distributed among the structural genes. The protein-encoding genes of the genome of 2019-nCoV were predicted by the online servers of GeneMarkS (http://exon.gatech.edu/GeneMark/genemarks.cgi) and ORFfinder (https://www.ncbi.nlm.nih.gov/orffinder/) with manual check.
(B) Phylogenetic relationship based on the whole genome for the HB01 strain and other coronaviruses. All viral strains were classified by the genus and the type, which are presented on the left and right schematic phylogenetic trees, respectively. The four genera of the coronaviruses, including Alphacoronavirus (red), Betacoronavirus (blue), Gammacoronavirus (green), and Deltacoronavirus (violet) are blocked in the left phylogenetic tree. The MERS coronavirus (brown), the SARS-like bat coronavirus (violet), human SARS coronavirus (light blue), and the HB01 strain (red) are highlighted by lines of different colors in the right phylogenetic tree.
(C) Schematic phylogenetic trees of individual genes for the HB01 strain. The coronavirus species were colored in the same way as (B). The amount of the strains in the phylogenetic clade is denoted by the area of the circles.

Analysis (MEGA) (version 7.0), the 2019-nCoV is in the same Betacoronavirus clade as MERS-CoV, SARS-like bat CoV, and SARS-CoV. The phylogenetic tree falls into two clades. The Betacoronavirus genus constitutes one clade, while the Alphacoronavirus, Gammacoronavirus, and Deltacoronavirus genera constitute the other clade. The 2019-nCoV is parallel to the SARS-like bat CoVs, while the SARS-CoVs are descended from the SARS-like bat CoVs, indicating that 2019-nCoV is closer to the SARS-like bat CoVs than the SARS-CoVs in terms of the whole genome sequence. Tables S1C and S1D also show that the genome
of 2019-nCoV has the highest similarity with that of a SARS-like bat CoV (MG772933). In comparison, 2019-nCoV is distant from and less related to the MERS-CoVs. In terms of the encoded proteins of pp1ab, pp1a, envelope, matrix, accessory protein 7a, and nucleocapsid genes, phylogenetic analyses showed that the 2019-nCoV is closest to the SARS-like bat CoVs (Figure 1C and Table S1D). Regarding the spike gene, the 2019-nCoV is closest to the bat CoVs, while the 3a and 8b accessory genes are both closest to the SARS-CoVs. Although phylogenetic analyses for the whole genome and individual genes clearly show that the 2019-nCoV is most closely related to SARS-like bat

![Figure 2. Amino Acid Substitutions of 2019-nCoV against SARS and SARS-like Viruses](image-url)
viruses (Figures 1B and 1C), we did not find a single strain of a SARS-like bat virus that harbors all proteins with the most similarity to counterparts of the 2019-nCoV (Figures 1B and 1C).

Given the close relationship between 2019-nCoV and SARS-CoVs or SARS-like bat CoVs (Figures 1B and 1C), an examination of the amino acid substitutions in different proteins could shed light into how 2019-nCoV differs structurally and functionally from SARS-CoVs. In total, there were 380 amino acid substitutions between the amino acid sequences of 2019-nCoV (HB01) and the corresponding consensus sequences of SARS and SARS-like viruses (Figure 2 and Tables S1E and S1F). No amino acid substitutions occurred in nonstructural protein 7 (nsp7), nsp13, envelope, matrix, or accessory proteins 6 and 8b (Table S1F). Respectively, 102 and 61 amino acid substitutions are located in nsp3 and nsp2. In addition, 27 amino acid substitutions were found in the spike protein with a length of 1,273 amino acids, including six substitutions in the RBD at amino acid region 357–528 and six substitutions in the underpinning subdomain (SD) at amino acid region 569–655. Moreover, four substitutions (Q560L, S570A, F572T, and S575A) in the C-terminal of the receptor-binding subunit S1 domain (Figure 2) are situated in two peptides previously reported to be antigens for SARS-CoV (Guo et al., 2004).

Due to very limited knowledge of this novel virus, we are unable to give reasonable explanations for the significant number of amino acid substitutions between the 2019-nCoV and SARS or SARS-like CoVs. For example, no amino acid substitutions were present in the receptor-binding motifs that directly interact with human receptor ACE2 protein in SARS-CoV (Ge et al., 2013), but six mutations occurred in the other region of the RBD. Whether these differences could affect the host tropism and transmission property of the 2019-nCoV compared to SARS-CoV is worthy of future investigation.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.chom.2020.02.001.

ACKNOWLEDGMENTS

This work was supported by the National Key Plan for Scientific Research and Development of China (2016YFD0500301 and 2016YFC1200200), CAMS Initiative for Innovative Medicine (CAMS-I2M and 2016-I2M-1-005), the National Natural Science Foundation of China (U1630316), the Central Public-Interest Scientific Institution Basal Research Fund (2016ZX10105, 2017PT31026, and 2018PT31018), and NIH R01AI069120 (United States).

REFERENCES


