Characterization of codon usage pattern in novel coronavirus 2019-nCoV

Wei Hou (houweicn@163.com)

Tianjin Second People's Hospital and Tianjin Institute of Hepatology https://orcid.org/0000-0001-9223-424X

Short Report

Keywords: coronaviruses, 2019-nCoV, codon usage pattern

Posted Date: February 26th, 2020

DOI: https://doi.org/10.21203/rs.2.24512/v1

Abstract

The outbreak of viral pneumonia in China due to a novel coronavirus 2019-nCoV poses significant threats to international health. In this study we perform bioinformatic analysis to take a snapshot of the codon usage pattern of 2019-nCoV and uncover that this novel coronavirus has a relatively low codon usage bias. The information from this research may not only be helpful to get new insights into the evolution of 2019-nCoV, but also have potential value for developing coronavirus vaccines.

Introduction

Coronaviruses (CoVs) belong to the family *Coronavirdiae* comprises large, single, plus-stranded RNA viruses including four genera of CoVs, namely, *Alphacoronavirus*, *Betacoronavirus*, *Deltacoronavirus*, and *Gammacoronavirus* [1]. There are six coronavirus species known to cause human disease[1-3], including two alphacoronaviruses (HCoV-229E, HCoV-NL63) and four betacoronaviruses (HCoV-OC43, HCoV-HKU, Severe acute respiratory syndrome-related coronavirus SARS-CoV and Middle East respiratory syndrome-related coronavirus MERS-CoV). Very recently, the outbreak of viral pneumonia in China due to a novel coronavirus 2019-nCoV poses significant threats to international health [4-9]. However, the genetic traits as well as evolutionary processes in this novel coronavirus are not fully characterized, and their roles in viral pathogenesis are yet largely unknown. To further explore the codon usage pattern of 2019-nCoV to get a better picture of the codon architecture of this novel coronavirus, genomic sequences of the 2019-nCoV and other six representative coronaviruses were analyzed via bioinformatic approaches.

Materials And Methods

Genomic sequences retrieval

Genomic sequences of the 2019-nCoV Wuhan-Hu-1 (MN908947.3) and other representative coronaviruses including human coronavirus HCoV-229E (AF304460.1), HCoV-NL63 (AY567487.2), HCoV-OC43(AY585228.1), HCoV-HKU1（MH940245.1）, and SARS-CoV (strain: Urbani, AY278741.1; strain: Tor2, AY274119.3), MERS-CoV (strain: HCoV-EMC, JX869059.2) were retrieved from GenBank.

Phylogenetic analysis

Phylogenetic tree of the whole genome sequences of coronaviruses were constructed by using MEGA software version 6.0 (http://www.megasoftware.net) with the Maximum likelihood algorithm and Kimura 2-parameter model with 1000 bootstrap replicates.

Codon usage pattern analysis

The basic nucleotide composition (A%, U%, C%, and G%), AU and GC contents, relative synonymous codon usage (RSCU) were analyzed using MEGA software. The parameters of codon usage bias

including intrinsic codon bias index (ICDI), codon bias index (CBI), effective number of codons (ENC) were analyzed using CAIcal [10] and COUSIN programs (http://cousin.ird.fr/index.php). Cluster analysis (Heat map) was performed using CIMminer (https://discover.nci.nih.gov/cimminer/).

Results And Discussion

Phylogenetic analysis of coronavirus genomes (Figure 1A) revealed that the newly identified coronavirus 2019-nCoV Wuhan-Hu-1 sequence was closer to SARS-CoVs and more distant from two alphacoronaviruses (HCoV-229E, HCoV-NL63).

Nucleotide composition analysis revealed that 2019-nCoV Wuhan-Hu-1 had the highest compositional value of U% (32.2) which was followed by A% (29.9), and similar composition of G% (19.6) and C% (18.3). Moreover, the mean GC and AU compositions were 37.9% and 62.1% (2019-nCoV Wuhan-Hu-1), 41.0% and 59.0% (SARS-CoV Tor2), 40.8% and 59.2% (SARS-CoV Urbani), 41.5% and 58.5% (MERS-CoV HCoV-EMC), 36.8% and 63.2% (HCoV-OC43), 32.0% and 68.0% (HCoV-HKU1), 38.0% and 62.0% (HCoV-229E), 34.4% and 65.6% (HCoV-NL63), respectively indicating that 2019-nCoV Wuhan-Hu-1 as well as other representative coronaviruses in this study are all AU rich.

RSCU analysis of the complete coding sequences of 2019-nCoV Wuhan-Hu-1 revealed that the following codons (AGA, UAA, GGU, GCU, UCU, GUU, CCU, ACU, CUU, UCA, ACA, UUA) were over-represented (RSCU value >1.6) and all ended with A/U. The highest RSCU value for the codon AGA for R (2.67) amino acid and lowest in UCG for S (0.11), which was consistent with recent report by Codon W1.4.2 analysis [9]. The heatmap analysis (Figure 1B) further revealed that all the coronaviruses analyzed in this study share the over-represented codons (GGU, GCU, UAA, GUU, UCU, CCU, ACU) and the average RSCU value >2.0, whereas two codons (UCA, ACA) were over-represented only in 2019-nCoV and SARS-CoVs.

The profiles of codon usage patterns among different genes of coronaviruses were further analyzed (Figure 1C). As for spike (S) gene, all the coronaviruses analyzed in this study share the over-represented codons (UCU, GUU, GCU, CCU, ACU, AUU) and all ended with U, whereas two codons (CCA, ACA) were over-represented only in 2019-nCoV. As for envelop (E) gene, two codons (GCG, UAC) were over-represented only in 2019-nCoV and SARS-CoVs. All the coronaviruses analyzed in this study did not use two synonymous codons (CGC, CGG) for arginine as well as CCG for proline at all. Only 2019-nCoV and SARS-CoVs did not use CAA for glutamine whereas they use AUC for isoleucine and UCG for serine. As for membrane (M) gene, two codons (GUA, GAA) were over-represented only in 2019-nCoV. As for nucleocapsid (N) gene, all the coronaviruses analyzed in this study share the over-represented codons (CUU, ACU, GCU) and all ended with U. The average RSCU values of GCU in complete gene, S gene, E gene, M gene and N gene in all the coronaviruses were 2.22, 2.30, 1.79, 2.13, 2.16, respectively. GCU for alanine was identified as the highly preferred codon.

To further estimate the degree of codon usage bias, intrinsic codon bias index (ICDI), codon bias index (CBI) and effective number of codons (ENC) values were calculated (Table 1). ICDI value (0.144), CBI value (0.306) and ENC value (45.38) all exhibited relatively low codon usage bias of 2019-nCoV, similar to

SARS-CoV Tor2, SARS-CoV Urbani, MERS-CoV HCoV-EMC, HCoV-OC43, HCoV-229E whereas different from HCoV-HKU1 (ICDI 0.372; CBI 0.532; ENC 35.617) and HCoV-NL63 (ICDI 0.307; CBI 0.476; ENC 37.275), which exhibited moderate codon usage bias.

Table 1 The parameters of codon usage bias among the coronaviruses analyzed in this study.

Coronaviruses	ICDI	CBI	ENC
2019-nCoV Wuhan-Hu-1	0.144	0.306	45.38
SARS-CoV Tor2	0.075	0.223	49.746
SARS-CoV Urbani	0.08	0.228	48.965
MERS-CoV HCoV-EMC	0.082	0.248	50.033
HCoV-OC43	0.213	0.367	43.794
HCoV-HKU1	0.372	0.532	35.617
HCoV-229E	0.172	0.358	43.45
HCoV-NL63	0.307	0.476	37.275

Overall, this study has taken a snapshot of the codon usage pattern of 2019-nCoV. This novel coronavirus has a relatively low codon usage bias, similar to most of the representative coronaviruses, which might help to adapt to the host or the varied environment. Influence factors account for the low codon usage bias of 2019-nCoV, e.g. natural selection and mutational pressure, warrant further investigation. The information from this research may not only be helpful to get new insights into the evolution of human coronavirus, but also have potential value for developing coronavirus vaccines.

Declarations

Author contributions WH conceived and designed the study, analyzed data, drafted the manuscript, and agreed to be accountable for all aspects of the work.

Conflict of interest The authors declare that they have no conflict of interest.

Research involving Human Participants and/or Animals Not applicable.

Informed consent Not applicable.

References

1. Cui J, Li F, Shi ZL. Origin and evolution of pathogenic coronaviruses. Nat Rev Microbiol. 2019 Mar;17(3):181-192. doi: 10.1038/s41579-018-0118-9. Review. PubMed PMID: 30531947.
2. de Wit E, van Doremalen N, Falzarano D, Munster VJ. SARS and MERS: recent insights into emerging coronaviruses. Nat Rev Microbiol. 2016 Aug;14(8):523-34. doi: 10.1038/nrmicro.2016.81. Epub 2016 Jun 27. Review. PubMed PMID: 27344959.
3. Graham RL, Donaldson EF, Baric RS. A decade after SARS: strategies for controlling emerging coronaviruses. Nat Rev Microbiol. 2013 Dec;11(12):836-48. doi: 10.1038/nrmicro3143. Epub 2013

Nov 11. Review. PubMed PMID: 24217413; PubMed Central PMCID: PMC5147543.

4. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. N Engl J Med. 2020 Jan 24. doi: 10.1056/NEJMoa2001017. [Epub ahead of print] PubMed PMID: 31978945.
5. Chaolin Huang, Yeming Wang, Xingwang Li, Lili Ren, Jianping Zhao, Yi Hu, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. 2020 Jan 24. doi:10.1016/S0140-6736(20)30183-5.
6. Jasper Fuk-Woo Chan, Shuofeng Yuan, Kin-Hang Kok, Kelvin Kai-Wang To, Hin Chu, Jin Yang, et al. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. 2020 Jan 24. doi: 10.1016/S0140-6736(20)30154-9.
7. Fan Wu, Su Zhao, Bin Yu, Yan-Mei Chen, Wen Wang, Zhi-Gang Song, et al. A new coronavirus associated with human respiratory disease in China. Nature (2020). https://doi.org/10.1038/s41586-020-2008-3.
8. Peng Zhou, Xing-Lou Yang, Xian-Guang Wang, Ben Hu, Lei Zhang, Wei Zhang, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature (2020). https://doi.org/10.1038/s41586-020-2012-7
9. Ji W, Wang W, Zhao X, Zai J, Li X. Homologous recombination within the spike glycoprotein of the newly identified coronavirus may boost cross-species transmission from snake to human. J Med Virol. 2020 Jan 22. doi:10.1002/jmv.25682. [Epub ahead of print] PubMed PMID: 31967321.
10. Puigbò P, Bravo IG, Garcia-Vallve S. CAIcal: a combined set of tools to assess codon usage adaptation. Biol Direct. 2008 Sep 16;3:38. doi:10.1186/1745-6150-3-38. PubMed PMID: 18796141; PubMed Central PMCID: PMC2553769.

Figures

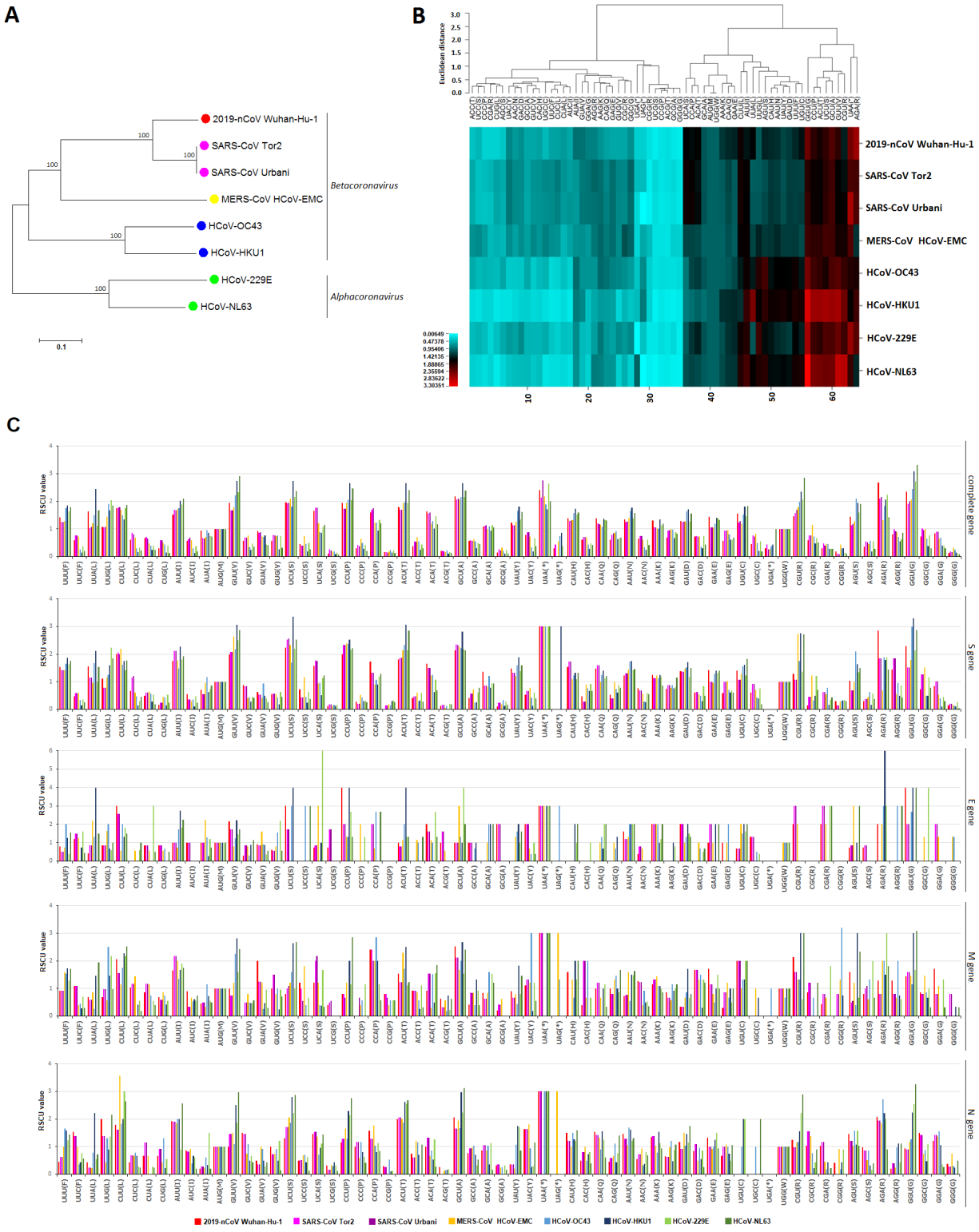

Figure 1

Bioinformatic analyses of 2019-nCoV. (A) Maximum likelihood phylogenetic tree of the whole genome sequences of 2019-nCoV Wuhan-Hu-1 (MN908947.3) and related coronaviruses including human coronavirus HCoV-229E (AF304460.1), HCoV-NL63 (AY567487.2), HCoV-OC43(AY585228.1), HCoV-HKU1 (MH940245.1), and SARS-CoV (strain: Urbani, AY278741.1; strain: Tor2, AY274119.3), MERS-CoV (strain: HCoV-EMC, JX869059.2). (B) Heat map of RSCU values for the complete coding sequences of 2019-nCoV

and related coronaviruses. The heatmap analysis was performed using CIMminer. Each row represents a codon. Codons with higher RSCU values are highlighted with a red background.(C) The profiles of the relative synonymous codon usage for different genes of 2019-nCoV and related coronaviruses. RSCU values were shown as the vertical bar graph. S:spike; E:envelop; M: membrane; N: nucleocapsid.