Characterization of Codon Usage Pattern in Novel Coronavirus SARS-CoV-2

Wei Hou (houweicn@163.com)

Short report

Keywords: COVID-19, coronaviruses, SARS-CoV-2, codon usage pattern

Posted Date: April 17th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-21553/v1

Version of Record: A version of this preprint was published on September 14th, 2020. See the published version at https://doi.org/10.1186/s12985-020-01395-x.

Abstract

The outbreak of COVID-19 due to a novel coronavirus SARS-CoV-2 has posed significant threats to international health. In this study we perform bioinformatic analysis to take a snapshot of the codon usage pattern of SARS-CoV-2 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 SARS-CoV-2, 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 coronavirus disease 2019 (COVID-19) due to a novel coronavirus SARS-CoV-2 (tentatively named 2019-nCoV) has posed 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 SARS-CoV-2 to get a better picture of the codon architecture of this novel coronavirus, genomic sequences of the SARS-CoV-2 and other six representative coronaviruses were analyzed via bioinformatic approaches.

Materials And Methods

Genomic sequences retrieval

Genomic sequences of the SARS-CoV-2 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 (Fig. 1A) revealed that the newly identified coronavirus SARS-CoV-2 Wuhan-Hu-1 sequence was closer to SARS-CoV Tor-2 as well as SARS-CoV Urbani, and more distant from two alphacoronaviruses (HCoV-229E, HCoV-NL63).

Nucleotide composition analysis revealed that SARS-CoV-2 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% (SARS-CoV-2 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 SARS-CoV-2 Wuhan-Hu-1 as well as other representative coronaviruses in this study are all AU rich.

RSCU analysis of the complete coding sequences of SARS-CoV-2 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 (Fig. 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 SARS-CoV-2 and SARS-CoVs.

The profiles of codon usage patterns among different genes of coronaviruses were further analyzed (Fig. 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 SARS-CoV-2. As for envelop (E) gene, two codons (GCG, UAC) were over-represented only in SARS-CoV-2 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 SARS-CoV-2 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 SARS-CoV-2. 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 SARS-CoV-2, 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
SARS-CoV-2 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 SARS-CoV-2. 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 SARS-CoV-2, 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.

Abbreviations

COVID-19
coronavirus disease 2019
CoVs
coronaviruses
SARS-CoV
severe acute respiratory syndrome-related coronavirus
MERS-CoV
Middle East respiratory syndrome-related coronavirus
SARS-CoV-2
severe acute respiratory syndrome-related coronavirus 2
RSCU
relative synonymous codon usage

ICDI
codon bias index
CBI
codon bias index
ENC
effective number of codons

Declarations

Ethics approval and consent to participate

Not applicable.

Consent to publication

Not applicable.

Availability of data and material

This work was posted online in a preprint platform Research Square (https://www.researchsquare.com/article/rs-15071/v1; DOI:10.21203/rs.2.24512/v1) on 26 Feb, 2020.

Competing interests

The authors declare that they have no competing interests.

Funding

Not applicable.

Authors' contributions

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

Acknowledgements

Not applicable.

References

1. Cui J, Li F, Shi ZL. Origin and evolution of pathogenic coronaviruses. Nat Rev Microbiol. 2019;17:181–92.
2. de Wit E, van Doremalen N, Falzarano D, Munster VJ. SARS and MERS: recent insights into emerging coronaviruses. Nat Rev Microbiol. 2016;14:523–34.

3. Graham RL, Donaldson EF, Baric RS. A decade after SARS: strategies for controlling emerging coronaviruses. Nat Rev Microbiol. 2013;11:836–48.
4. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, Zhao X, Huang B, Shi W, Lu R, Niu P, Zhan F, Ma X, Wang D, Xu W, Wu G, Gao GF, Tan W. China Novel Coronavirus Investigating and Research Team. A Novel Coronavirus from Patients with Pneumonia in China, 2019. N Engl J Med. 2020;382:727–33.
5. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, Zhang L, Fan G, Xu J, Gu X, Cheng Z, Yu T, Xia J, Wei Y, Wu W, Xie X, Yin W, Li H, Liu M, Xiao Y, Gao H, Guo L, Xie J, Wang G, Jiang R, Gao Z, Jin Q, Wang J, Cao B. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet. 2020;395:497–506.
6. Chan JF, Yuan S, Kok KH, To KK, Chu H, Yang J, Xing F, Liu J, Yip CC, Poon RW, Tsoi HW, Lo SK, Chan KH, Poon VK, Chan WM, Ip JD, Cai JP, Cheng VC, Chen H, Hui CK, Yuen KY. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. Lancet. 2020;395:514–23.
7. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, Hu Y, Tao ZW, Tian JH, Pei YY, Yuan ML, Zhang YL, Dai FH, Liu Y, Wang QM, Zheng JJ, Xu L, Holmes EC, Zhang YZ. A new coronavirus associated with human respiratory disease in China. Nature. 2020;579:265–9.
8. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, Si HR, Zhu Y, Li B, Huang CL, Chen HD, Chen J, Luo Y, Guo H, Jiang RD, Liu MQ, Chen Y, Shen XR, Wang X, Zheng XS, Zhao K, Chen QJ, Deng F, Liu LL, Yan B, Zhan FX, Wang YY, Xiao GF, Shi ZL. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020;579:270–3.
9. Ji W, Wang W, Zhao X, Zai J, Li X. Cross-species transmission of the newly identified coronavirus 2019-nCoV. J Med Virol. 2020;92:433–40.
10. Puigbò P, Bravo IG, Garcia-Vallve S. CAIcal: a combined set of tools to assess codon usage adaptation. Biol Direct. 2008;3:38.

Figures

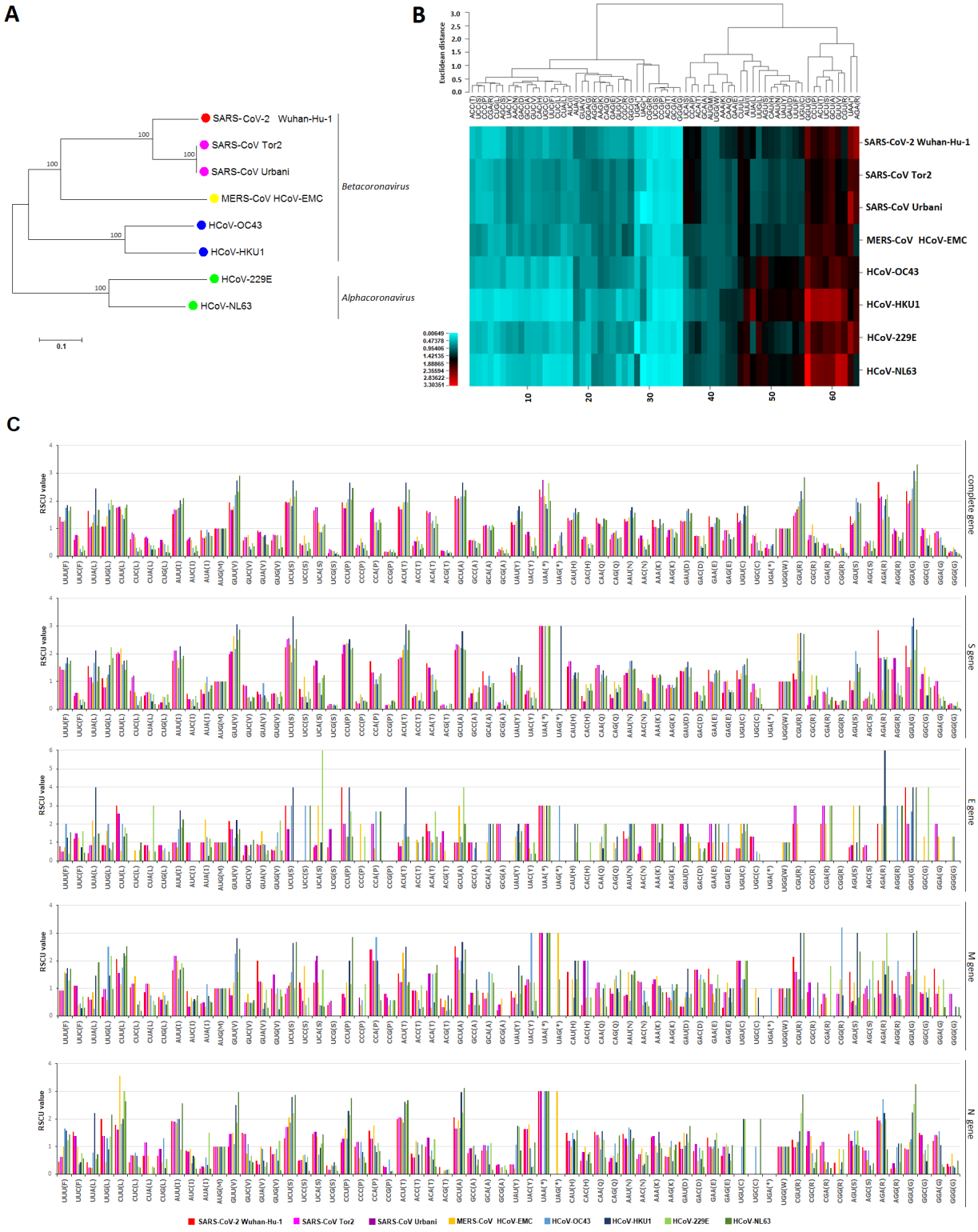

Figure 1

Bioinformatic analyses of SARS-CoV-2. (A) Maximum likelihood phylogenetic tree of the whole genome sequences of SARS-CoV-2 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 SARS-CoV-

2 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 SARS-CoV-2 and related coronaviruses. RSCU values were shown as the vertical bar graph. S:spike; E:envelop; M: membrane; N: nucleocapsid.