Characterization of Codon Usage Pattern in SARS-CoV-2

Wei Hou (✉ houweicn@163.com)

Short report

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

Posted Date: August 6th, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-21553/v2>

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

[Read Full License](#)

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 coronavirus disease 2019 (COVID-19) due to severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has posed significant threats to international health. 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 get a better picture of the codon architecture of this newly emerging coronavirus, in this study we perform bioinformatic analysis, based on publicly available nucleotide sequences of SARS-CoV-2 along with those of other members of human coronaviruses as well as non-human coronaviruses in different hosts, to take a snapshot of the genome-wide codon usage pattern of SARS-CoV-2 and uncover that all over-represented codons end with A/U and this newly emerging coronavirus has a relatively low codon usage bias, which is shaped by both mutation pressure and natural selection. Additionally, there is slight variation in the codon usage pattern among the SARS-CoV-2 isolates from different geo-locations. Furthermore, the overall codon usage pattern of SARS-CoV-2 is generally similar to that of its phylogenetic relatives among non-human betacoronaviruses such as RaTG13. Taken together, we comprehensively analyze the characteristics of codon usage pattern in SARS-CoV-2 via bioinformatic approaches. 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 *Coronaviridae* comprises large, single, positive-sense single-stranded RNA viruses including four genera of CoVs, namely, *Alphacoronavirus*, *Betacoronavirus*, *Deltacoronavirus*, and *Gammacoronavirus* [1]. Several coronavirus species were extensively 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 severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has posed significant threats to international health [4-9]. However, the genetic traits as well as evolutionary processes in this newly emerging 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 representative coronaviruses were analyzed via bioinformatic approaches.

Materials And Methods

Genomic sequences acquisition

Genomic sequences of the SARS-CoV-2 Wuhan-Hu-1 (MN908947.3), other representative coronaviruses including human coronaviruses such as HCoV-229E (AF304460.1), HCoV-NL63 (AY567487.2), HCoV-OC43(AY585228.1), HCoV-HKU1MH940245.1, and SARS-CoV (strain: Urbani, AY278741.1; strain: Tor2,

AY274119.3), MERS-CoV (strain: HCoV-EMC, JX869059.2) and nonhuman coronaviruses (Supplementary Table 1) were all 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 Calcal [10] and COUSIN programs (<http://cousin.ird.fr/index.php>).

Cluster analysis

Cluster analysis (Heat map) was performed using CIMminer (<https://discover.nci.nih.gov/cimminer/>).

Results And Discussion

Phylogenetic analysis of human coronavirus genomes (Figure 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 relatively distant to two alphacoronaviruses (HCoV-229E, HCoV-NL63).

Nucleotide composition analysis (Supplementary Figure 1A) 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). At the third position, We observed that nucleotide U also occurred most frequently. Thus, most codons of SARS-CoV-2 Wuhan-Hu-1 tended to be U ending. Moreover, the mean GC and AU compositions (Supplementary Figure 1B) 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 human coronaviruses in this study were all AU rich, which was consistent with recent reports [11-15].

RSCU analysis of the complete coding sequences of SARS-CoV-2 Wuhan-Hu-1 revealed that all the over-represented codons (RSCU value >1.6) ended with A/U whereas most of the under-represented codons (RSCU value <0.6) ended with C/G (Supplementary Table 2). The highest RSCU value for the codon was AGA for R (2.67) amino acid and lowest was in UCG for S (0.11). The heatmap analysis (Figure 1B) further revealed that all human coronaviruses analyzed in this study share the over-represented codons

(UAA, GGU, GCU, UCU, GUU, CCU, ACU) and the average RSCU value >2.0, whereas UCA were over-represented only in SARS-CoV-2 and SARS CoV.

The profiles of codon usage patterns among different genes of human coronaviruses were further analyzed (Figure 1C and Figure 2). As for spike (S) gene, all human coronaviruses analyzed in this study shared the over-represented codons (UCU, GCU, CUU, GUU, ACU) and all ended with U, whereas two codons (CCA, ACA) were over-represented only in SARS-CoV-2. In addition, SARS-CoV-2 did not use CGA for arginine nor CCG for proline. As for envelop (E) gene, two codons (UAC, GCG) were over-represented only in SARS-CoV-2 and SARS-CoV. All human coronaviruses analyzed in this study did not use two synonymous codons (CGC, CGG) for arginine as well as CCG for proline and UGA for stop codon at all. SARS-CoV-2 and SARS-CoV did not use CAA for glutamine nor UAU for tyrosine, whereas they use GCG for alanine, AUC for isoleucine, UCG and AGC for serine . As for membrane (M) gene, three codons (GUA, GAA, GGA) were over-represented only in SARS-CoV-2. As for nucleocapsid (N) gene, all human coronaviruses analyzed in this study share the over-represented codons (GCU, ACU, CUU) 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 human coronaviruses analyzed in this study were 2.22, 2.12, 1.79, 2.13, 2.16, respectively. GCU for alanine was identified as the highly preferred codon among the human coronaviruses.

Amino acids are degenerate and each amino acid has different number of synonymous codons except for methionine (Met, M) and tryptophan (Trp, W). The overall amino acid usage of the human coronaviruses was shown in Supplementary Figure 2. Leucine and valine were the two most frequently used amino acids in all human coronaviruses analyzed in this study, CUU and GUU were preferred codons for leucine and valine, respectively (Figure 3), whereas tryptophan, histidine and methionine were the three least used ones, which was consistent with recent report [14].

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.

We next attempted to determine the forces influencing the codon usage bias. Accumulating evidence suggests that the formation of codon usage bias is affected by many factors, and two generally accepted major forces are mutation pressure and natural selection[16]. Other influential factors include gene expression level, gene length, GC content, GC contents at the third base of one codon (GC3), RNA stability, hydrophilicity, and hydrophobicity, etc. When G or C is in high or low proportion at the third position of the codon, mutational pressure is involved [17]. From Supplementary Figure1, it clearly showed that both G3 and C3 were lower than A3 and U3, suggesting the contribution of mutational force acting on codon usage pattern. Moreover, all preferred codons were A/U ending (Figure 1B, 1C and Figure 2), which further suggested that mutational force contributed to shape codon usage in this virus. Furthermore, to better

understand the relations between gene composition and codon usage bias, an ENC–GC3 scatter diagram of ENC versus GC3S (ENC plotted against G+C content at the third codon position) was constructed. When codon usage pattern is only affected by GC3 resulting from mutation pressure, the expected ENC values should be just on the solid curved line. As shown in Supplementary Figure 3, all points lie together under the expected ENC curve, indicating that some independent factors, such as natural selection might also play a role in codon usage bias of human coronaviruses.

Apart from human, many animal species can also be infected by different types of coronaviruses. Previous studies have shown that some animals such as bats are believed to represent the original reservoir of several human-infecting coronaviruses (1). In order to provide additional information to better understand the evolution of SARS-CoV-2, we further compared the codon usage pattern of SARS-CoV-2 and non-human coronaviruses (Supplementary Table 1).

Phylogenetic analysis (Figure 4A) showed that SARS-CoV-2 was most closely related to recently reported Bat coronavirus RaTG13[18]. Nucleotide composition analysis (Supplementary Figure 4) revealed that similar to SARS-CoV-2 Wuhan-Hu-1, all the non-human coronaviruses analyzed in this study had the highest compositional value of U% and nucleotide U occurred most frequently at the third position. The heatmap analysis (Figure 4B) revealed that SARS-CoV-2 and all the non-human coronaviruses analyzed in this study shared the over-represented codons (GGU, UCU, CCU) and all ended with U, meanwhile they shared the under-represented codons (UCG, GGG, GCG, CCG, CGG, ACG, CGA) and most ended with G except for CGA. Codon usage pattern of SARS-CoV-2 was generally found a high similarity to that of betacoronaviruses except for Bat coronavirus HKU4-1, Bat coronavirus HKU5-1(Figure 4C, Supplementary Figure 5-8). Moreover, the profiles of codon usage patterns among different genes of SARS-CoV-2 and non-human coronaviruses were further analyzed, as shown in Figure 5 and Supplementary Figure 9-12. We found similar codon usage pattern among SARS-CoV-2 and its phylogenetic relatives such as RaTG13, Bat-SL-CoVZC45, Bat-SL-CoVZXC21, PCoV_GX-P1E, PCoV_GX-P4L, which may reflect the evolutionary relationship between SARS-CoV-2 and these non-human coronaviruses. These results are in accordance with the full-genome phylogenetic analysis (Figure 4A). The overall amino acid usage of the non-human coronaviruses was shown in Supplementary Figure 13 . Similar to SARS-CoV-2 , leucine and valine were the two most frequently used amino acids in all non-human coronaviruses analyzed in this study, CUU and GUU were preferred codons for leucine and valine, respectively.

Furthermore, similar to SARS-CoV-2, all the non-human coronaviruses analyzed in this study exhibited relatively low codon usage bias according to the intrinsic codon bias index (ICDI), codon bias index (CBI) and effective number of codons (ENC) values, as shown in Supplementary Figure 14. Nucleotide composition analysis (Supplementary Figure 4) and ENC-GC3S plot (Supplementary Figure 15) revealed that both mutational force and natural selection contribute to shape codon usage in non-human coronaviruses.

Overall, in the present study we attempted to take a snapshot of the characteristics of codon usage pattern in novel coronavirus SARS-CoV-2. As a result, we found all over-represented codons ended with

A/U and this novel coronavirus had a relatively low codon usage bias. Both mutation pressure and natural selection were contributors to the bias. Additionally, the overall codon usage pattern of SARS-CoV-2 was generally similar to that of its phylogenetic relatives among non-human coronaviruses such as RaTG13. Our findings are consistent with the recent observations [11-15] and provide new insights into the characteristics of codon usage pattern in coronaviruses. These results also have important implications for future work.

Firstly, the information of genome-wide codon usage pattern of SARS-CoV-2 may be helpful to get new insights into the evolution of this new virus. With the increase of SARS-CoV-2 genome data available, we could reevaluate the codon usage pattern of SARS-CoV-2 more comprehensively to track the evolutionary changes between them. In this regard, genome-wide codon usage patterns in 100 complete genome sequences of SARS-CoV-2 isolates including SARS-CoV-2 Wuhan-Hu-1 from different geo-locations were analyzed herein. All information about the used isolates can be found in Supplementary table 3. The heatmap analysis (Supplementary Figure 16) revealed 12 preferred codons (GGU, GCU, UAA, GUU, UCU, CCU, ACU, UAA, GGU, GCU, UCU, GUU, CCU, ACU) ending with A/U among all the 100 isolates, and the average RSCU value of these over-represented codons vary from 1.63 to 2.67 (Supplementary Figure 17). The highest RSCU value was for the codon AGA for R (2.67) amino acid and lowest was in UCG for S (0.11). We noted that the overall codon usage pattern appeared to be similar among the tested 100 SARS-CoV-2 isolates from different geo-locations, reflecting minimal evolutionary changes among them.

Additionally, compared to other members of human coronaviruses as well as non-human coronaviruses in different hosts, we found that the overall codon usage pattern of SARS-CoV-2 is generally similar to that of its phylogenetic relatives among non-human betacoronaviruses such as RaTG13 (Figure 5), which may reflect the evolutionary relationship between SARS-CoV-2 and these non-human coronaviruses.

Secondly, the information of genome-wide codon usage pattern of SARS-CoV-2 may have potential value for developing coronavirus vaccines to combat this pandemic disease. The information on codon usage by SARS-CoV-2 may pave the way to design strategies such as codon deoptimization[19-21], the use of the least preferred codons to modify the SARS-CoV-2 genome to reduce virulence for the development of a safe and effective vaccine. This strategy has several advantages. Deoptimized viruses could express an identical antigenic repertoire of T- and B-cell epitopes because they contain the intact wide type amino acid sequence. Moreover, deoptimized viruses can efficiently replicate in vitro while being highly attenuated in vivo, which is important for vaccine production and their safe implementation.

Conclusions

Taking all these results together, our studies reveal that SARS-CoV-2 has a relatively low codon usage bias, which is shaped by both mutation pressure and natural selection. Additionally, there is slight variation in the codon usage pattern among the SARS-CoV-2 isolates from different geo-locations. Furthermore, the overall codon usage pattern of SARS-CoV-2 is generally similar to that of its phylogenetic relatives among non-human betacoronaviruses such as RaTG13. The information from this

research may not only be helpful to get new insights into the evolution of human coronaviruses, but also have potential value for developing coronavirus vaccines.

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

We thank researchers for making viral sequences available for public access.

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

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 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. CAlcal: a combined set of tools to assess codon usage adaptation. *Biol Direct.* 2008;3:38.

11. Anwar AM, Khodary SM. Insights into The Codon Usage Bias of 13 Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) Isolates from Different Geo-locations bioRxiv 2020.04.01.019463; doi: <https://doi.org/10.1101/2020.04.01.019463>
12. Dilucca M, Forcelloni S, Georgakilas AG, Giansanti A, Pavlopoulou A. Codon Usage and Phenotypic Divergences of SARS-CoV-2 Genes. *Viruses*. 2020 Apr 30;12(5). pii: E498. doi: 10.3390/v12050498. PubMed PMID: 32366025.
13. Kandeel M, Ibrahim A, Fayez M, Al-Nazawi M. From SARS and MERS CoVs to SARS-CoV-2: Moving toward more biased codon usage in viral structural and nonstructural genes. *J Med Virol*. 2020 Mar 11. doi: 10.1002/jmv.25754. [Epub ahead of print] PubMed PMID: 32159237.
14. Gu H, Chu DKW, Peiris M, Poon LLM. Multivariate analyses of codon usage of SARS-CoV-2 and other betacoronaviruses. *Virus Evol*. 2020 May 14;6(1):veaa032. doi: 10.1093/ve/veaa032. eCollection 2020 Jan. PMID: 32431949
15. Tort FL, Castells M, Cristina J. A comprehensive analysis of genome composition and codon usage patterns of emerging coronaviruses. *Virus Res*. 2020 Apr 12;283:197976. doi: 10.1016/j.virusres.2020.197976. [Epub ahead of print] PubMed PMID: 32294518; PubMed Central PMCID: PMC7152894.
16. Plotkin JB, Kudla G. Synonymous but not the same: the causes and consequences of codon bias. *Nat Rev Genet*. 2011 Jan;12(1):32-42. doi: 10.1038/nrg2899. Epub 2010 Nov 23. PMID: 21102527
17. Chaney JL, Clark PL. Roles for Synonymous Codon Usage in Protein Biogenesis. *Annu Rev Biophys*. 2015;44:143-66. doi: 10.1146/annurev-biophys-060414-034333. Epub 2015 Feb 26. PMID: 25747594.
18. 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 Mar;579(7798):270-273. doi: 10.1038/s41586-020-2012-7. Epub 2020 Feb 3. PMID: 32015507
19. Baker SF, Nogales A, Martínez-Sobrido L. Downregulating viral gene expression: codon usage bias manipulation for the generation of novel influenza A virus vaccines. *Future Virol*. 2015 Jun;10(6):715-730. PubMed PMID: 26213563; PubMed Central PMCID: PMC4508661.
20. Fan RL, Valkenburg SA, Wong CK, Li OT, Nicholls JM, Rabadan R, Peiris JS, Poon LL. Generation of Live Attenuated Influenza Virus by Using Codon Usage Bias. *J Virol*. 2015 Nov;89(21):10762-73. doi: 10.1128/JVI.01443-15. Epub 2015 Aug 12. PMID: 26269186
21. Blanco-Lobo P, Nogales A, Rodríguez L, Martínez-Sobrido L. Novel Approaches for The Development of Live Attenuated Influenza Vaccines. *Viruses*. 2019 Feb 22;11(2). pii: E190. doi: 10.3390/v11020190. Review. PubMed PMID: 30813325; PubMed Central PMCID: PMC6409754.

Tables

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

Coronaviruses	GenBank ACCESSION	Genome Length(nt)	ICDI	CBI	ENC
SARS-CoV-2 Wuhan-Hu-1	MN908947.3	29903	0.144	0.306	45.38
SARS-CoV Tor2	AY274119.3	29751	0.075	0.223	49.746
SARS-CoV Urbani	AY278741.1	29727	0.08	0.228	48.965
MERS-CoV HCoV-EMC	JX869059.2	30119	0.082	0.248	50.033
HCoV-OC43	AY585228.1	30741	0.213	0.367	43.794
HCoV-HKU1	MH940245.1	29811	0.372	0.532	35.617
HCoV-229E	AF304460.1	27317	0.172	0.358	43.45
HCoV-NL63	AY567487.2	27553	0.307	0.476	37.275

Figures

Figure 1

Bioinformatic analyses of SARS-CoV-2 and other human coronaviruses. (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 other human 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 other human. RSCU values were shown as the vertical bar graph. S:spike; E:envelop; M: membrane; N: nucleocapsid.

Figure 2

Over-represented codons (RSCU value >1.6) and under-represented codons (RSCU value <0.6) were compared among SARS-CoV-2 and other human coronaviruses. (A) Complete gene; (B) S gene; (C) E gene; (D) M gene; (E) N gene.

Figure 3

Graphical representation of synonymous codon usage pattern of each amino acid among SARS-CoV-2 and other human coronaviruses using Weblogo.

Figure 4

Comparative analysis of SARS-CoV-2 and non-human coronaviruses. (A) Maximum likelihood phylogenetic tree of the whole genome sequences of SARS-CoV-2 Wuhan-Hu-1 (MN908947.3) and non-human coronaviruses including Scotophilus bat coronavirus 512(NC_009657.1), Swine enteric alphacoronavirus strain SeACoV-p10(MK977618.1), Bat coronavirus HKU4-1(NC_009019.1), Bat coronavirus HKU5-1(NC_009020.1), Bat coronavirus RaTG13(MN996532.1), Bat SARS-like coronavirus

isolate bat-SL-CoVZC45(MG772933.1), Bat SARS-like coronavirus isolate bat-SL-CoVZXC21(MG772934.1), Pangolin coronavirus isolate PCoV_GX-P1E(MT040334.1), Pangolin coronavirus isolate PCoV_GX-P4L(MT040333.1), Avian coronavirus strain H120(MK071267.1), Avian coronavirus strain Ma5(KY626045.1), Bulbul coronavirus HKU11-796(FJ376620.1), Thrush coronavirus HKU12-600(NC_011549.1), Munia coronavirus HKU13-3514(NC_011550.1). (B) Heat map of RSCU values for the complete coding sequences of SARS-CoV-2 and non-human 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 complete gene of SARS-CoV-2 and non-human coronaviruses. Over-represented codons (RSCU value >1.6) and under-represented codons (RSCU value <0.6) were shown as line graph.

Figure 5

Over-represented codons (RSCU value >1.6) and under-represented codons (RSCU value <0.6) were compared among SARS-CoV-2 and its phylogenetic relatives including RaTG13, Bat-SL-CoVZC45, Bat-SL-CoVZXC21, PCoV_GX-P1E, PCoV_GX-P4L. (A) Complete gene; (B) S gene; (C) E gene; (D) M gene; (E) N gene.

Supplementary Files

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

- [SupplementaryTable1.docx](#)
- [SupplementaryFigure1.tif](#)
- [SupplementaryFigure13.png](#)
- [SupplementaryFigure15.png](#)
- [SupplementaryFigure4.tif](#)
- [SupplementaryFigure5.tif](#)
- [SupplementaryFigure2.png](#)
- [SupplementaryFigure7.tif](#)
- [SupplementaryFigure6.tif](#)
- [SupplementaryFigure8.tif](#)
- [SupplementaryFigure17.tif](#)
- [SupplementaryFigure9.tif](#)
- [SupplementaryFigure10.tif](#)
- [SupplementaryFigure14.tif](#)
- [SupplementaryFigure12.tif](#)
- [SupplementaryFigure11.tif](#)
- [SupplementaryFigure16.tif](#)
- [SupplementaryTable2.docx](#)
- [SupplementaryTable3.docx](#)
- [SupplementaryFigure3.png](#)