Human SARS-CoV-2 has evolved to reduce CG dinucleotide in its open reading frames

CURRENT STATUS: UNDER REVIEW

natureresearch

Yong Wang
School of Food and Biological Engineering, Jiangsu University

ywang@ujs.edu.cn*Corresponding Author*

Jun-Ming Mao
Institute of Life Sciences, Jiangsu University

Guang-Dong Wang
Institute of Life Sciences, Jiangsu University

Ze Qiu
School of Food and Biological Engineering, Jiangsu University

Qin Yao
Institute of Life Sciences, Jiangsu University

Ke-Ping Chen
Institute of Life Sciences, Jiangsu University

DOI:

10.21203/rs.3.rs-21003/v1

SUBJECT AREAS

Evolutionary Genetics

KEYWORDS

Coronavirus, Odds ratio, Dinucleotide, Evolution, Free energy

Abstract

The causative agent of COVID-19 is a severe acute respiratory syndrome-related coronavirus which has been officially named SARS-CoV-2. Here we report the discovery of extremely low CG abundance in its open reading frames. We found that CG reduction in SARS-CoV-2 is achieved mainly through mutating C/G into A/T, and CG is the best target for mutation. In view of energy usage, a coronavirus with low CG abundance has higher efficiency in translating its RNA, because the secondary structure formed by viral genome is less stable. 5’-untranslated region of SARS-CoV-2 has much more CGs and is capable of recruiting host ribosomes to initiate translation. Notably, genomes of cellular organisms also have very low CG abundance, suggesting that mutating C/G into A/T occurs universally in all life forms. Moreover, CG is related to mutational hotspots and CpG islands in cellular organisms. The relationship between them is worthy of further investigations.

Introduction

The outbreak of COVID–19 (coronavirus disease 2019) was listed as a public health emergency of international concern on 30 January 2020 by the World Health Organization. As of 5 March 2020, it has caused 95,265 infection cases and 3,281 deaths worldwide[1]. Its causative virus (officially named SARS-CoV–2) has a genome of single-strand positive-sense RNA with approximately 30,000 nucleotides[2]. Based on its genome sequence, analyses have been conducted to characterize genomic features and to trace origin of the virus[3,4]. Meanwhile, many reports have been focused on developing efficient methods for detection[5,6] and screening effective drugs for treatment[7,8]. Here we report the discovery of extremely low abundance of CG dinucleotide in open reading frames (ORFs) of SARS-CoV–2 (named SCoV2 hereafter). In view of energy usage, a coronavirus with reduced CG content has higher efficiency in translating its RNA, because less energy is consumed in disrupting the stem-loops formed in its secondary structure.

Results And Discussion

DNA or RNA sequences are composed of four nucleotides. They can also be considered polymers of 16 dinucleotides. Odds ratio is a value defined to indicate relative abundance of a nucleotide, which is the ratio of observed to expected frequency of a dinucleotide[9]. The genome of SCoV2 (29,903 nucleotides[2], sequence number NC_045512) has 29.94% of A, 32.08% of T (T is used here instead of U for simplicity), 19.61% of G and 18.37% of C. Thus, the expected frequency of CG dinucleotide in viral genome is 3.60% (i.e. 19.61% x 18.37%). However, only 439 CGs are observed, which means the observed frequency is 1.47% (i.e. 439/29,902). Therefore, odds ratio of CG in SCoV2 is 0.41 (i.e. 1.47%/3.60%). Furthermore, odds ratio of CG in open reading frames (ORFs) of the virus is 0.39, being the lowest among 24 coronaviruses under survey (Fig. 1a and Extended Data Table 1). Because a codon is composed of three nucleotides, a dinucleotide (e.g. CG) has three possible locations. Herewith, they are designated as $(CG)_{12}$, $(CG)_{23}$ and $(CG)_{31}$ respectively. We found that the odds ratio of $(CG)_{23}$ in ORFs of SCoV2 is as low as 0.25, while that of $(CA)_{23}$ and $(CT)_{23}$ is as high as 1.54 and 1.92 respectively (Fig. 1c). Moreover, odds ratio of $(CG)_{31}$ in ORFs of SCoV2 is 0.50, while that of $(AG)_{31}$ and $(TG)_{31}$ is 1.52 and 2.64 respectively (Fig. 1d). These data strongly suggest that $(CG)_{23}$ has been mutated into $(CA)_{23}$ and $(CT)_{23}$, and $(CG)_{31}$ has been mutated into $(AG)_{31}$ and $(TG)_{31}$.

The above-stated mutations are possible because very few of these mutations lead to changes in amino acids. To be specific, there are four codons containing $(CG)_{23}$. They are TCG, CCG, ACG and GCG which code for serine, proline, threonine and alanine, respectively. Mutation of G at codon position 3 into T, C or A in all of them does not change the amino acid they encode. As for $(CG)_{31}$, there are 16 codons having C at position 3. If this C is mutated into T, all 16 codons have the same meanings. And if it is mutated into A, 9 out of 16 codons still have the same meanings. Therefore, it is concluded that SCoV2 has evolved to reduce CG in ORFs mainly through mutating its G of $(CG)_{23}$ and C of $(CG)_{31}$ into A and T. Among them, C-to-T (i.e. C-to-U in RNA) occurs at a very high frequency probably because it is the simplest way to change a nucleotide (C becomes U after deamination). Besides, odds ratio of $(CC)_{23}$ is much lower than that of $(CA)_{23}$ and $(CT)_{23}$. This does not mean that G of $(CG)_{23}$ has not been mutated into $(CC)_{23}$. In fact, low odds ratio of $(CC)_{23}$ is the result of high mutation frequency of $(CG)_{31}$ into $(TG)_{31}$ (Fig. 1c and 1d). The above views are also supported by codon usage bias in SCoV2 (Fig. 2), which shows that A/T-ended codons are much more frequently used than their synonymous G/C-ended codons. Besides, all four codons containing $(CG)_{23}$ have the lowest percentages of usage among synonymous codons.

Odds ratios of CG in ORFs of other coronaviruses are also very low (mean value = 0.50, Extended Data Figure 1 and Extended Data Table 1). This could have profound effect on viral replication, because ORFs of coronaviruses are immediately translated by host ribosomes after being released into the cytoplasm of host cells[10]. The translation of viral RNA is affected by two factors. One is that host ribosomes must be recruited to the 5’-UTR (untranslated region) of viral RNA for initiation of translation. The other is that stem-loops formed by ORFs of viral RNA must be disrupted during translation. In contrast to ORFs, 5’-UTR of coronaviruses have quite high odds ratios of CG (mean value = 0.84, Extended Data Table 2). This would facilitate formation of stable secondary structure that could serve as the internal ribosome entry site (IRES)[11–13] for host ribosome (Extended Data Figure 2). Meanwhile, the viral RNA beginning at the translation start site (TSS) forms relatively unstable secondary structure, because its stem-loops are maintained by less hydrogen bonds (an A-T base pair has one less hydrogen bond than a C-G base pair). Stability variations of viral genomes at 5’-UTR and TSS-to-end regions could probably determine virulence of different viruses, because high stability of IRES structure means high efficiency in initiating translation, and high stability of TSS-to-end region means high energy consumption during translation. For example, both 5’-UTR and TSS-to-end regions of human MCoV are highly stable (Table 1). High stability of 5’-UTR means that host ribosomes can be recruited to translate viral RNA at high rate. And, high stability of ORFs means that more energy is consumed to disrupt stem-loops in viral RNA during translation. Thus, normal translation of host cell mRNAs is greatly affected, suggesting that MCoV is highly virulent. 5’-UTRs of human SCoV and SCoV2 are less stable than MCoV, meaning that host ribosomes are not recruited to initiate translation of viral RNA at high rate. Yet, TSS-to-end region of SCoV2 is less stable than SCoV (Table 1), meaning that less energy is consumed by translation of viral RNA. Thus, SCoV2 is less virulent than SCoV. This conclusion is consistent with estimations on case fatality ratio of MCoV, SCoV and SCoV2, which is 35%, 9% and 2.4% respectively[14]. Three other human coronaviruses also have different stability in 5’-UTR and TSS-to-end regions (Table 1). Specifically, human CoV 229E has low stability in 5’-UTR and high stability in TSS-to-end region. Human CoV NL63 and HKU1 have medium and low stability in both regions, respectively. Such variations indicate that these coronaviruses could also have different virulence.

It seems that the strategy of “reducing CG content to increase gene expression efficiency” has also been adopted by cellular organisms. As we have observed, CG in both ORFs and inter-genic regions of bacteria, archaea, fungi, plants and animals has an average odds ratio of 0.81, and that in introns of fungi, plants and animals is as low as 0.69. At time of our previous report[15], we did not know why CG has such a low odds ratio in surveyed organisms. Now, after analysing cases in coronaviruses, we realize that low CG content in cellular organisms should also be the evolutionary consequence of increasing gene expression efficiency, because lowered CG content means reduced number of hydrogen bonds between DNA double strands (of the same length). Expression of a gene with low CG content saves energy not only in separating DNA double strands during transcription but also in disrupting stem-loops formed by mRNA during translation. Coincidently, CG is the very dinucleotide related to existence of mutational hotspots and CpG islands in DNA sequences of cellular organisms. A mutational hotspot is defined as CG with methylated C, in which the methylated C is frequently mutated into T through deamination[16–18]. A CpG island is defined as a region of DNA with less methylated C, and this region generally contains actively expressed genes[19–21]. The relationship between CG reduction and these two important features of cellular DNA sequences is worthy of further investigations.

If reducing hydrogen bonds is the goal of base mutation, why is CG but not GC, GG or CC taken as the target for mutation? An examination on number of silent mutations of each dinucleotide at various codon positions reveals that CG has the highest number (47) among these four dinucleotides (Table 1 and Extended Data Table 3). This explains why CG is the best target for mutation. Although CT has the same highest number like CG, it is not taken as the target for mutation because a T-to-C or T-to-G mutation would increase number of hydrogen bonds between potential base pairs, which is contradictory to the target of mutation. Our present study provides a novel insight into the evolution of human SCoV2. It is evident that this virus has evolved to reduce CG intensely in its ORFs. Such reduction is achieved mainly through mutating G of $(CG)_{23}$ and C of $(CG)_{31}$ into A or T (Fig. 1). Meanwhile, C or G not of CG may also be mutated. For example, TCA in SCoV2 of S-type has been mutated into TTA in that of L-type[22]. GTC and GGT in SCoV2 isolated from France have been mutated into TTC and GTT respectively in that from Wuhan (China)[23]. Although the mutated C or G is not of CG and not at codon position 3, they do reduce C or G in viral RNA. As such, it is speculated that G+C content may be used as an indicator of evolution degree for different SCoV2 isolates (i.e. the lower the G+C content, the higher the evolution degree). However, this speculation presumes that mutations aiming to reduce C or G occur predominantly in SCoV2. To test this presumption, further investigations are expected to identify and analyse detailed mutational events occurring in different SCoV2 isolates.

Methods

Genome sequences of coronaviruses were retrieved from GenBank (www.ncbi.nlm.nih.gov). Odds ratios of dinucleotides were calculated using self-compiled computer programs (C++ scripts are available upon request). Secondary structure and free energy of viral RNA is predicted using RNAstructure (version 5.7)[24]. Independent-sample *t*-test was run to compare difference in odds ratio of nucleotide between coronaviruses and cellular organisms using SPSS software (version 17.0).

Declarations

Acknowledgements

This study was supported by the National Natural Science Foundation of China (No. 31872425 and No. 31861143051).

Author contribution

Y. W., Q. Y. and K. P.C conceived the study. Y. W. and J. M.M compiled computer programs. Y. W., G. D. W. and Z. Q. performed surveys and analyses.

Competing interests

All authors declare no competing interests.

References

1. WHO Director-General’s opening remarks at the media briefing on COVID-19, 5 March 2020, https://www.who.int/dg/speeches/detail/
2. Wu, *et al*. A new coronavirus associated with human respiratory disease in China. *Nature*, 3 Feb 2020, doi: 10.1038/s41586-020-2008-3. [Epub ahead of print]
3. Zhou, *et al*. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, 3 Feb 2020, doi: 10.1038/s41586-020-2012-7. [Epub ahead of print]
4. Wassenaar, M. & Zou, Y. 2019_nCoV/SARS-CoV-2: rapid classification of betacoronaviruses and identification of Traditional Chinese Medicine as potential origin of zoonotic coronaviruses. *Lett Appl Microbiol*, 14 Feb 2020, doi: 10.1111/lam.13285. [Epub ahead of print]
5. Jin, H. *et al*. A rapid advice guideline for the diagnosis and treatment of 2019 novel coronavirus (2019-nCoV) infected pneumonia (standard version). *Mil Med Res* **7**, 4, (2020). doi: 10.1186/s40779-020-0233-6.
6. Li, Z. *et al*. Development and clinical application of a rapid IgM-IgG combined antibody test for SARS-CoV-2 infection diagnosis. *J Med Virol*, 27 Feb 2020, doi: 10.1002/jmv.25727. [Epub ahead of print]
7. Gao, J., Tian, & Yang, X. Breakthrough: Chloroquine phosphate has shown apparent efficacy in treatment of COVID-19 associated pneumonia in clinical studies. *Biosci Trends*, 19 Feb 2020, doi: 10.5582/bst.2020.01047. [Epub ahead of print]
8. Lim, J. *et al*. Case of the index patient who caused tertiary transmission of COVID-19 infection in Korea: the application of Lopinavir/Ritonavir for the treatment of COVID-19 infected pneumonia monitored by quantitative RT-PCR. *J Korean Med Sci* **35**, e79, (2020).
9. Karlin, S. & Mrázek, J. Compositional differences within and between eukaryotic

genomes. *Proc Natl Acad Sci USA* **94**, 10227-10232, (1997).

10. Fehr, R. & Perlman, S. Coronaviruses: an overview of their replication and pathogenesis. *Methods Mol Biol* **1282**, 1-23, (2015).

11. Sonenberg, N. & Pelletier, J. Poliovirus translation: a paradigm for a novel initiation mechanism. *Bioessays* **11**, 128-132, (1989).

12. Ren, Q. *et al*. Alternative reading frame selection mediated by a tRNA-like domain of an internal ribosome entry site. *Proc Natl Acad Sci USA* **109**, E630-639, (2012).

13. Renaud-Gabardos, E. *et al*. Internal ribosome entry site-based vectors for combined gene *World J Exp Med* **5**, 11-20, (2015).

14. Peeri, N. C. *et al*. The SARS, MERS and novel coronavirus (COVID-19) epidemics, the newest and biggest global health threats: what lessons have we learned? *Int J Epidemiol*, 22 Feb 2020, doi: 10.1093/ije/dyaa033. [Epub ahead of print]

15. Wang, *et al*. TA, GT and AC are significantly under-represented in open reading frames of prokaryotic and eukaryotic protein-coding genes. *Mol Genet Genomics* **294**, 637-647, (2019).

16. Shen, J. C., Rideout, M. & Jones, P. A. The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA. *Nucleic Acids Res* **22**, 972-976, (1994).

17. Krawczak, M., Ball, E. & Cooper, D. N. Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes. *Am J Hum Genet* **63**, 474-488, (1988).

18. Cooper, D. N., Mort, M., Stenson, D., Ball, E. V. & Chuzhanova, N. A. Methylation-mediated deamination of 5-methylcytosine appears to give rise to mutations causing human inherited disease in CpNpG trinucleotides, as well as in CpG dinucleotides. *Hum Genomics* **4**, 406-410, (2010).

19. Gardiner-Garden, M. & Frommer, M. CpG islands in vertebrate genomes. *J Mol Biol*

196, 261-282, (1987).

20. Deaton, A. M. & Bird, A. CpG islands and the regulation of transcription. *Genes Dev* **25**, 1010-1022, (2011).

21. Hartl, D. *et al*. CG dinucleotides enhance promoter activity independent of DNA methylation. *Genome Res* **29**, 554-563, (2019).

22. Tang, L. *et al*. On the origin and continuing evolution of SARS-CoV-2. *Natl Sci Rev*, 3 March 2020, https://doi.org/10.1093/nsr/nwaa036.

23. Cleemput, S. *et al*. Genome Detective Coronavirus Typing Tool for rapid identification and characterization of novel coronavirus genomes. *Bioinformatics*, 28 Feb 2020, doi: 10.1093/bioinformatics/btaa145. [Epub ahead of print]

24. Reuter, J. S. & Mathews, D. H. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinform* **11**,129, (2010).

Details Of Extended Data Available As Supplemental Files

Extended Data Figure 1 | Odds ratios of dinucleotides in open reading frames of coronaviruses and cellular organisms. a, odds ratios of dinucleotides at all codon positions. *b, c and d,* odds ratios of dinucleotides at codon positions 1 and 2, 2 and 3, 3 and 1, respectively. Data of coronaviruses are from Extended Data Table 1, which are shown in blue background. Those of cellular organisms are from our previous work[15]. or indicates that odds ratio of a dinucleotide in coronavirus is significantly higher or lower than that in cellular organisms at p = 0.05 level. or indicates that odds ratio of a dinucleotide in coronavirus is insignificantly higher or lower than that in cellular organisms.

Extended Data Figure 2 | Secondary structure formed by 5’-UTR of poliovirus (a) and SCoV2 (b). The secondary structure is based on 200 nucleotides immediately upstream of the translation start site. Sequence number of poliovirus is MG212486. That of SCoV2 is NC_045512. Free energy of the secondary structure is shown in centre of the structure.

Extended Data Table 1 | Odds ratios of dinucleotides in open reading frames of coronaviruses. This table lists the weighted average odds ratio of each dinucleotide based on ORF length. For example, the weighted average odds ratio of dinucelotide GT = [(odds ratio of GT in ORF1ab)*(length of ORF1ab) + (odds ratio of GT in ORF2)*(length of ORF2) + … + (odds ratio of GT in ORFn)*(length of ORFn)] / (length of ORF1ab + length of ORF2 + … + length of ORFn). Data of coronaviruses shown in Extended Data Figure 1 are mean values and standard deviations of each dinucleotide, which are shown in blue background in this table.

Extended Data Table 2 | Free energy of RNA secondary structure in 5'-UTR and TSS-to-end regions of coronaviruses. This table lists the free energy of RNA secondary structure formed by 200 nucleotides of 5'-UTR immediately upstream of ORF1ab and by TSS-to-end region. Accumulated free energy is the sum of free energy of TSS-to-end region separated into segments of 1,000 nucleotides. Normalized free energy is based on average TSS-to-end size (28,085 nt) of all surveyed coronaviruses. For example, normalized free energy of human SCoV2 = 28,085*(–8,295.0) / 29,638 = –7,860.4 kcal/ml. Data for Table 1 are shown in blue background in this table. TSS: translation start site.

Extended Data Table 3 | Number of silent mutations of each dinucleotide at various codon positions. Dinucleotide at codon positions 1 and 2, 2 and 3, or 3 and 1 is shown at upper-left corner of the genetic code table. Value beneath it indicates number of silent mutations of codons containing this dinucleotide. In the genetic code table, codons containing this dinucleotide are in yellow background, and those of correspondent silent mutations are in green background.

Tables

Table 1 Stability of secondary structure formed by genome of coronavirus

Genus	Virus	5’-UTR*		TSS-to-end	
		Free energy (kcal/mol)	Stability index	Free energy (kcal/mol)	Stabili index
Alphacoronavirus	Bat CoV CDPHE15	-66.8	92 (H)	-8,803.5	99 (H)
	Bat CoV HKU10	-61.3	84 (M)	-8,029.1	90 (H)
	Cat CoV1	-68.8	95 (H)	-7,963.0	89 (M
	Rat CoV	-59.5	82 (M)	-8,615.0	97 (H)
	Mink CoV1	-71.2	98 (H)	-7,790.4	88 (M
	Bat CoV1	-59.6	82 (M)	-8,153.4	92 (H)
	Bat CoV Sax2011	-66.7	92 (H)	-8,815.5	99 (H)
	Bat CoV SC2013	-57.3	79 (L)	-8,712.4	98 (H)
	PEDV	-62.4	86 (M)	-8,671.9	97 (H)
	Bat CoV HKU2	-64.3	88 (M)	-8,313.0	93 (H)
	Human CoV NL63	-58.9	81 (M)	-7,223.3	81 (M
	Human CoV 229E	-55.6	76 (L)	-7,982.5	90 (H)
Betacoronavirus	Human CoV HKU1	-43.1	59 (L)	-6,864.6	77 (L)
	Human MCoV	-72.8	100 (H)	-8,436.5	95 (H)
	Human SCoV	-63.2	87 (M)	-8,054.1	91 (H)
	Human SCoV2	-62.4	86 (M)	-7,860.4	88 (M
	Bat CoV ZJ2013	-58.9	81 (M)	-8,328.2	94 (H)
	Bat CoV HKU9	-55.2	76 (L)	-8,897.8	100 (
Deltacoronavirus	Wigeon CoV HKU20	-51.8	71 (L)	-8,273.2	93 (H)
	Bulbul CoV HKU11	-54.3	75 (L)	-8,387.6	94 (H)
	Heron CoV HKU19	-54.9	75 (L)	-7,687.2	86 (M
	Moorhen CoV HKU21	-51.2	70 (L)	-8,140.4	91 (H)
Gammacoronavirus	Whale CoV SW1	-62.8	86 (M)	-8,161.3	92 (H)
	Turkey CoV	-59.0	81 (M)	-8,195.4	92 (H)

* Fee energy of 5’-UTR (untranslated region) was obtained by using 200 nucleotides immediately upstream of TSS (translation start site) for secondary structure prediction. Free energy of TSS-to-end region is normalized using the average genome size (28,085 nt) of all surveyed coronaviruses based on actual accumulated free energy of a specific genome (Extended Data Table 2). 5’-UTR region of human MCoV and TSS-to-end region of bat CoV HKU9 have the lowest free energy respectively, which

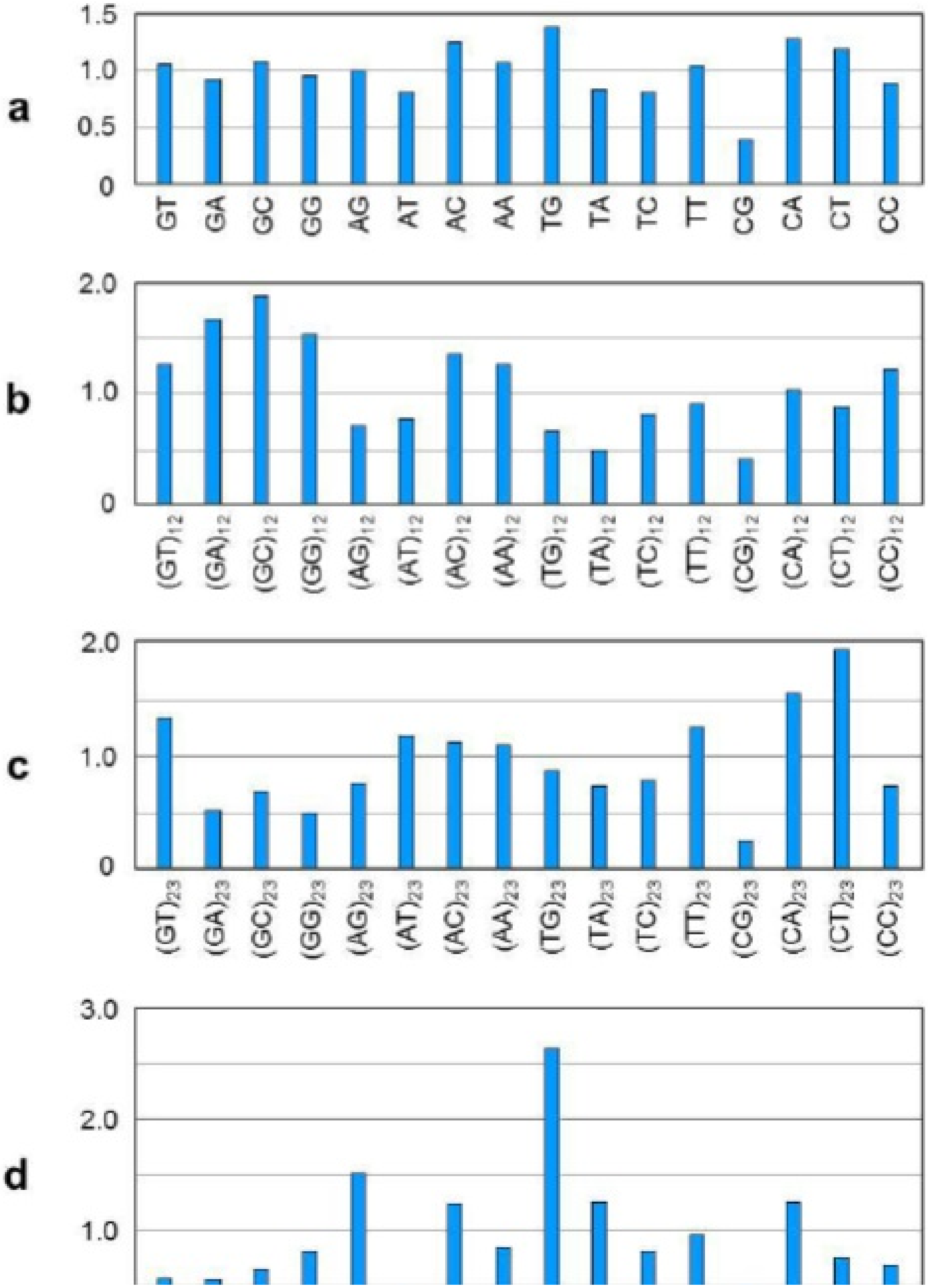
are thus given the highest stability index (100). H (high), M (medium) and L (low) indicate stability index of ≥90, 80 to 89, and <79, respectively. MCoV: Middle East respiratory syndrome-related coronavirus. SCoV: Severe acute respiratory syndrome-related coronavirus. PEDV: Porcine epidemic diarrhea virus. The viruses listed in the table were selected to represent different subgenera of coronaviruses.

Table 2 Number of silent mutations of each dinucleotide at various codon positions

Dinucleotide	Codon positions			Total
	1 and 2	2 and 3	3 and 1	
GT	0	8	30	38
GA	0	8	30	38
GC	0	8	30	38
GG	0	7	30	37
AG	4	4	32	40
AT	0	4	32	36
AC	0	4	32	36
AA	0	5	32	37
TG	1	7	33	41
TA	1	9	33	43
TC	2	9	33	44
TT	2	9	33	44
CG	2	12	33	47
CA	0	12	33	45
CT	2	12	33	47
CC	0	12	33	45

When a dinucleotide is located at codon positions 1 and 2 or at codon positions 2 and 3, there are four codons that contain this dinucleotide. Theoretically, they can be mutated into any of the rest 60 codons. When a dinucleotide is located at codon positions 3 and 1, only the nucleotide at position 3 is considered to mutate. There are 16 codons containing this nucleotide. Theoretically, they can be mutated into any of the rest 48 codons. Therefore, values in the table are number of silent mutations out of 60, 60 and 48 mutations for a dinucleotide at codon positions 1 and 2, 2 and 3, or 3 and 1, respectively.

Figures



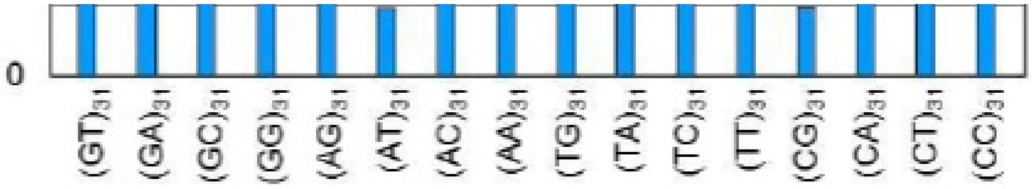
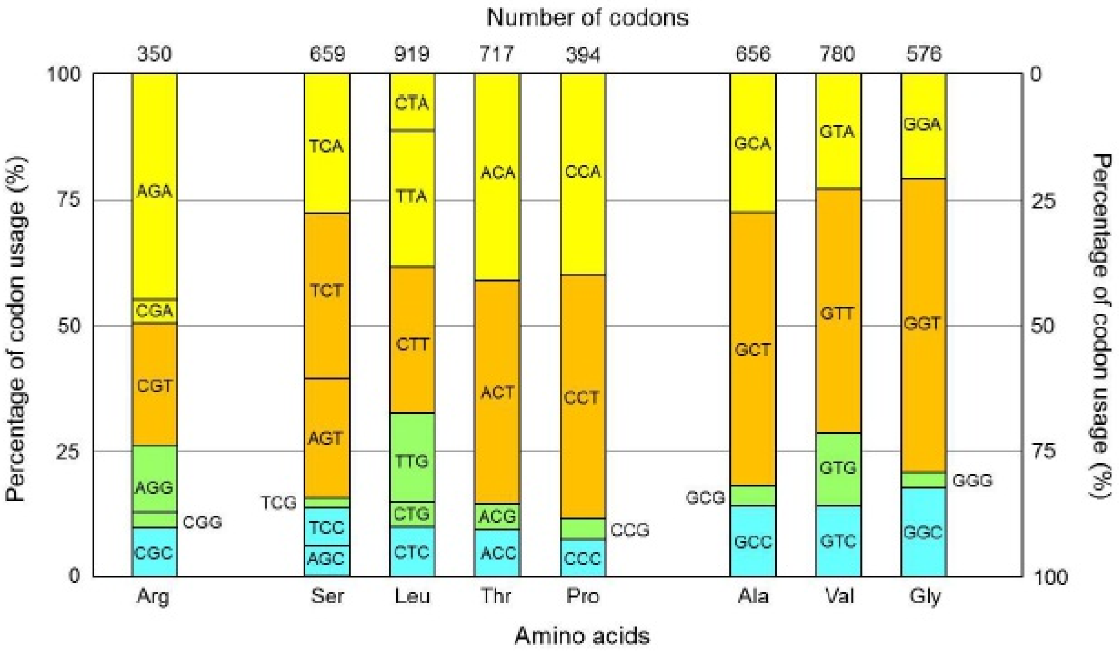
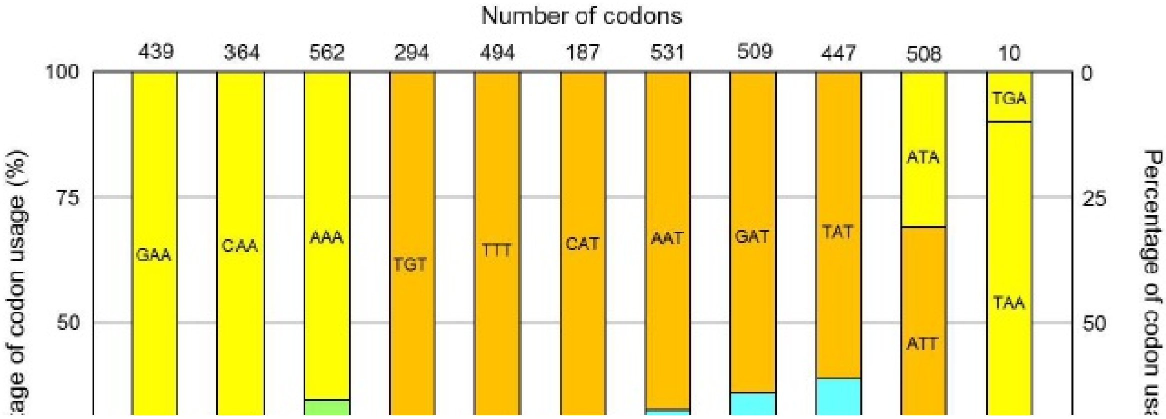


Figure 1

Odds ratios of dinucleotides in open reading frames of SCoV2. a, odds ratios of dinucleotides at all codon positions. b, c and d, odds ratios of dinucleotides at codon positions 1 and 2, 2 and 3, 3 and 1, respectively. Value shown in the figure is weighted average odds ratio of each dinucleotide. Odds ratio of each dinucleotide in ten ORFs (i.e. ORF1ab and ORF 2 to 10) of SCoV2 is calculated respectively first. Then, a weighted average odds ratio is obtained based on length of each ORF.





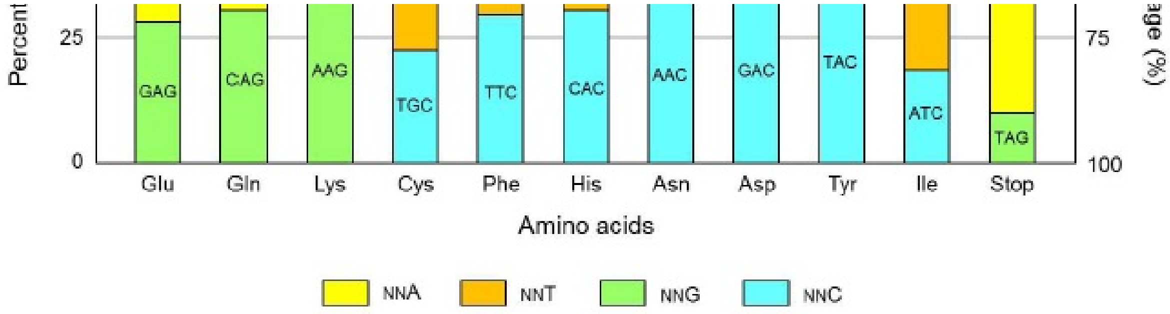


Figure 2

Percentages of codon usage in open reading frames of SCoV2. Usage of synonymous codons for eighteen amino acids (except methionine and tryptophan) and three stop codons are shown in the figure. Percentages of codons with A, T, G and C at codon position 3 are in yellow, brown, green and aqua blue background, respectively. Total number of codons for each amino acid is indicated at top of the percentage bar.

Supplementary Files

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

Extended Data Figure 1.png
Extended Data Figure 2.png
Extended Data Table 1.xlsx
Extended Data Table 2.xlsx
Extended Data Table 3.xlsx