Mutation patterns of human SARS-COV-2 and bat RaTG13 coronaviruses genomes are strongly biased towards C>U indicating rapid evolution in their hosts

Roman Matyasek
Czech Academy of Science
Ales Kovarik (kovarik@ibp.cz)
Czech Academy of Science https://orcid.org/0000-0003-2896-0698

Research Article

Keywords: Genetic variation, cytosine deamination, CpG depletion, coronavirus, SARS-CoV-2

Posted Date: April 7th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-21377/v1

Version of Record: A version of this preprint was published at Genes on April 7th, 2020. See the published version at https://doi.org/10.3390/genes11070761.

Title: Mutation patterns of human SARS-COV-2 and bat RaTG13 coronaviruses genomes are strongly biased towards C>U indicating rapid evolution in their hosts.

Author: Roman Matyášek[1] and Aleš Kovařík[1]

Address: [1]Laboratory of Molecular Epigenetics, Institute of Biophysics, Academy of Sciences of the Czech Republic, Královopolská 135, Brno 61265, Czech Republic,

Address for correspondence: Aleš Kovařík, email: kovarik@ibp.cz, phone: +420 541517178, ORCID: http://orcid.org/0000-0003-2896-0698

Key words: Genetic variation, cytosine deamination, CpG depletion, coronavirus, SARS-CoV-2

Abstract

Background: The world pandemy caused by SARS-CoV-2 spreading has raised considerable interest about its evolutionary origin and genome structure. Here we analysed mutation patterns in 13 human SARS-COV-2 isolates and a closely related RaTG13 isolated from *Rhinolophus affinis* bat. We also evaluated the CpG dinucleotide contents in SARS-COV-2 and other human and animal coronavirus genomes.

Results: Out of 1107 single nucleotide differences (c. 4% divergence) between human SARS-COV-2 and bat RaTG13, 672 (61%) can be attributed to C>U and U>T substitutions significantly ($P<0.001$) exceeding other types of SNPs. A similar trend was observed among the 13 sequenced SARS-COV-2 genomes. Accumulation of C>U mutations was also observed in a highly variable subregion encoding the ACE2 receptor contact domain. Contrast to most other coronaviruses both SARS-COV-2 and RaTG13 exhibited CpG depletion in their genomes.

Conclusion: The data support that the C-to-U conversion played a significant role in the evolution of pathogenic RNA coronaviruses including SARS-COV-2. These mutations apparently also influenced amino acid composition of the SARS-Cov-2 spike protein domain receptor implicated in virus pathogenicity. We propose that SARS-COV-2 was evolving relatively long in humans following the transfer from animals before spreading world-wide.

Introduction

The world pandemy caused by SARS-CoV-2 spreading has raised considerable interest about its evolutionary origin and genome structure. Its c. 30 kb-long single stranded plus RNA genome is AT rich (62%) and encodes 15 proteins (Zhou et al. 2020) preferring pyrimidine rich codons to purines (Kandeel et al. 2020). The spike protein of SARS-CoV-2 contains a domain important for the contact with the surface angiotensin converting enzyme 2 (ACE2) in human cells (Wan et al. 2020; Wrapp et al. 2020). Phylogenetically, SARS-COV-2 is closely related to *Rhinolophus affinis* (bat) virus, strain RaTG13 (96% identity)(Zhou et al. 2020) and the *Malaysian pangolin* coronavirus (91% identity)(Zhang et al. 2020). Most epidemiology and sequence data suggest that primary transfer occurred from bats to humans while timing and place of the transfer remains a topic of intensive debates. The GenBank contained about 110 entries of completely sequence SARS-CoV-2 genomes to the date (27th March 2020).

Cytosine appears to be the least stable base in nucleic acids due to deamination to uracil. In single stranded molecules, the half-life of any specific cytosine is estimated to about 200 years(Poole et al. 2001). Consequently, many genomes including those of viruses exhibit relatively low GC content particularly in areas under low selection constrains such as various repeats and pseudogenes. The CpG dinucleotides have long been observed to occur with a much lower frequency in the sequence of vertebrate genomes than would be expected due to random chance (Jabbari and Bernardi 2004; Stevens et al. 2013). In mammals, the depletion is explained by cytosine methylation which appears to be concentrated to CpG dinucleotides (Bird 2002). However, some pathogenic viruses (both RNA and DNA) including, flu, papilloma, polyoma and HIV also show reduction CpG dinucleotides in their genomes (Kypr et al. 1989) (Cheng et al. 2013; Upadhyay and Vivekanandan 2015; Alinejad-Rokny et al. 2016) suggesting that reduction of CpG content may not be bound to nuclear genomes

In this work we addressed the following questions: (i) What are the mutation patterns in SARS-CoV-2 and closely related RaTG13. (ii) What mutation types contribute to amino acid divergence in the critical ACE2 binding domain. (iii) Does CpG depletion occur in SARS-CoV-2 and other human and animal coronaviruses? We used bioinformatic analyses carried out on publicly available sequences in the GenBank. Evidence was obtained that mutation trends

are similar in both SARS-CoV-2 and RaTG13 biased towards C>U potentially driving CpG depletion in these genomes.

Material and Methods

Sequences were retrieved from the Genbank. Strains, Genbank accessions and other details are listed in Supplementary **Table S1**.

CLC Genomics Workbench 11.1 (Qiagen) was used to estimate intragenomic variation between coronaviruses. Briefly, a whole SARS-COV-2 sequence (GenBank accession number MN908947 (previously called NC_045512) was mapped to the RaTG13 accession as the reference. SNPs were called using the following parameters: genome coverage_1; counts_1; frequency_1. Accuracy of mapping was confirmed by a pairwise comparison and divergence computations. In a population level study of genetic variation different SARS-CoV-2 genomes were mapped the reference genome (MN908947).

Nucleotide composition of the genomes were calculated using a „Seqtk_comp " tool on a Galaxy server (toolshed.g2.bx.psu.edu/repos/iuc/seqtk/seqtk_comp/1.3.1) developed by Heng Li at the Broad Institute. The data were downloaded and exported to MsExcel sheet. The CpG depletion was expressed as the number of observed CpG/number of expected CpG dinucleotides calculated according to formula: $CpG_{(obs/exp)} = n(CpG)*L/(nC*nG)$ where n is the number of nucleotides, L is sequence length.

On line tools were used to calculate chi square statistics (https://www.socscistatistics.com/) and construct box plot graphs (http://shiny.chemgrid.org/boxplotr/).

Results

Sequence comparison and SNP analyses of related SARS-COV-2 and RaTG13 genomes

We compared frequency of mutations in SARS-CoV-2 (accession #MN908947) using bat RATG13 (MN996532.1) genome as a reference. The SARS-CoV-2 and RATG13 were respectively 29903 and 29855 nt long. Compared to RATG13, there were 7 insertions and 2 deletions in the SARS-CoV-2 genome, the longest insertion spanning 12 nt long GC rich

region between 23582-23583 of a bat sequence. There were 1136 SNPs between SARS-COV-2 and RaTG13 equalling to about 4% divergence consistent with the previous study (Zhou et al. 2020). Summary of the mutation analysis is given in Supplementary **Table S2**. Characteristics of single nucleotide variation (SNV) are shown in **Fig. 1a**. It is evident that the C>U and reverse U>C transitions were far most frequent mutations totally accounting for 86% of SNV variation. Comparison of other isolates of human coronavirus from 2019/2020 yielded essential the same results since there are no or little variations between the genomes (as further below).

Spike protein harbours a domain binding the ACE2 receptor on the cell surface believed to be important for host-virus interactions. Within a stretch of 20 amino acids (positions 486-501 in the protein sequence) differences in five amino acids were previously identified between SARS-COV-2 and RATG13 peptides (Andersen et al. 2020). Alignment of the corresponding 60 nt-long RNA sequence revealed 19 differences (31.7% divergence**)(Fig. 2**). Out of these, 9 (50%) showed C>U (U>C) patterns. Out of five, three polymorphic amino acids (bold) contained the C>U mutations in their codons.

Mutation patterns in different SARS-COV-2 isolates

We analysed 13 sequenced coronavirus isolates from different populations (**Table 1**). We selected sequences representing world-wide virus diversity. The number of SNPs per genome ranged from two (three isolates) to six (#MT093571 isolate from Sweden). Out of the 35 polymorphic sites, 14 were shared between at least two accessions, 21 were unique variants occurring in a single accession. The U>C SNP at 28144 position was found in three genomes (#MN985325 and # MT020880 (both USA) and #MT066175 from Taiwan).The G>U SNP at 26144 was found in #MT126808 (Brazil), #MT093571 (Sweden), #MT007544 (Australia) and Italy (#MT066156). Variants involving C>U and U>C mutations were abundantly represented in the data sets (**Fig. 1b**).

CpG depletion analysis in coronaviruses

The CpG depletion is a characteristic feature of eukaryotic genomes and some viruses. We determined the frequency of CpG dinucleotide in 17 human (including 8 SARS-CoV-2 accessions), 9 bat and 8 other animal betacoronaviruses (**Fig. 3a**). The number of CpGs

ranged 664-1766 between the genomes. The lowest level (652) was found in the human NL63 strain (#MK334043.1); the highest in the Japanese bat *Pipistrellus abramus* (1766). The *Rhinolophus affinis* (bat) RATG13 coronavirus had 882 CpGs in its genome, 4 sites more than SARS-CoV-2 (accession). Both ARS-CoV-2 and NL63 strain exhibited the lowest $CpG_{obs/exp}$ levels. Statistical evaluation of data is presented by box plots (**Fig. 3b**).

Discussion

Although C>U (U>C) conversions (transitions) are the most frequent SNPs in eukaryotic genomes a strong mutation bias towards this types of mutation in SARS-COVID-2 and bat RATG13 genomes, accounting for 61% of all differences between both genomes is unprecedented. The G>A and A>G transitions were the second most abundant while their frequency was significantly lower (23%; chi square, $P<0.001$) than that of C>U (U<C). The remaining 8 mutation types represented only 16% of all mutations. The adaptive character of C>U is demonstrated in the ACE2 receptor domain where three out of five polymorphic amino acids contained C>U signatures in their codons. One amino acid substitution contained a G>A mutation suggesting that coding regions are not excluded from the overall mutation bias. It is intriguing that a similar mutation trend was observed among human isolates in this and a recent report (Koyama et al. 2020). There is only partial overlap between both datasets. Though the number of mutations was low (overall 0-0.02% from genome to genome variation) it is likely that the world pandemy generated variants already.

In both SARS-CoV-2 and bat RATG13 genomes the ratio between CpG observed/expected was significantly lower than in most other coronavirus genomes. CpG depletions were reported in other human viruses including flu, HIV, papillomavirus and polyomaviruses. In viruses which integrate into the genome such as HIV, this may be explained by methylcytosine demethylation of CpG motifs in integrated copies (Alinejad-Rokny et al. 2016). However, in some non-integrating small dsDNA viruses the $CpG_{OBS/EXP}$ values were as low as 0.2 (Upadhyay and Vivekanandan 2015) suggesting other mechanisms. Perhaps, some kind of cytoplasmatic cytosine deaminase activity (Roberts et al. 2013) acting on single or double stranded intermediates might be responsible for C>U conversions. Human papilomaviruses and polyomaviruses showed more severe depletion of CpG dinucleotides

compared to their avian counterparts (Upadhyay and Vivekanandan 2015) and it has been proposed that host-specificity may play a role in shaping CpG content (Upadhyay et al. 2013). However, in our study such relationship is not so obvious since we observed large variation in CpG depletion between both human and animal coronaviures (**Fig. 3**). It is more likely that phylogenetic distance rather than host specificity plays a more decisive role. Currently, we can only speculate about the reasons why human SARS-COVID-2 and NL63 strains exhibit the highest level of CpG depletion out of all coronaviruses analysed. Interestingly, both strains cause severe pathological effects in their hosts. The CpG rich DNA has been shown to be immunogenic in humans (Klinman et al. 1997). Perhaps, both viruses are under the selection constrains to avoid immune systems. Another possibility is that CpG depletion reflects increased genome mutability perhaps associated rapidly spreading new variants. In this context HIV retrovirus shows far more stronger CpG depletion than the HTLV1 retrovirus causing leukemia in endemic areas. Only HIV has spread world-wide (Kypr et al. 1989).

Several hypotheses have been proposed to explain the origin of SARS-COVID-2 (Andersen et al. 2020): (i) natural selection in animal host before zoonotic transfer. (ii) Natural selection in humans following zoonotic transfer. (iii) Virus is a product of artificial manipulation. Our data makes the third possibility unlikely, since the SARS-COVID-2 genome follows similar evolutionary trajectories as the closely related bat RATG13 virus. We also propose that the C>U (and reverse U>C) are the most frequent mutations in pathogenic coronaviruses rapidly modifying their critical regions such as the ACE2 receptor binding domain important for host-virus interactions.

Declarations

Funding. The work was supported by the Czech Science Foundation (20-14133J and 20-28029S)).

Conflicts of interest/Competing interests. There are no conflicts of interests

Availability of data and material. All data are presented in the ms.

Code availability. Not applicable.

Authors' contributions. RM retrieved the data and carried out bioinformatic analysis. AK carried out bioinformatic analysis and wrote a paper.

Ethics approval. Not applicable.

Consent to participace. Not applicable.

Consent for publication. Not applicable

Figure legends

Fig. 1. Frequencies of twelve types of nucleotide substitutions in coronaviruses. **(a**) SARS-CoV-2 and RaTG13 coronaviruses. The data are from Supplementary **Table S1**. (**b**) 12 human SARS-CoV-2 isolates. The data are from **Table 1**.

Fig. 2. Coronarvirus genome with highlighted spike gene. A subregion encoding an ACE receptor contact domain (part) is magnified below. Differences between SARS-CoV-2 and bat RaTG13 sequences are highlighted. Changes in amino acid sequence are shown below the alignment. Numbers are according to the spike protein reading frame in the #MN908947 accession. Codons involving the C>U mutations are in bold. The direction of mutations is from the human to bat sequence.

Fig. 3. CpG depletion analysis in coronavirus genomes. (**a**) Bar charts showing CpG(OBS/EXP) in individual genomes. (**b**) Statistical representation of CpG depletions in individual groups. Differences between human and animal strains were significant (chi square, $P<0.001$, a single SARS-CoV-2 (#MN908947) plus all other human coronaviruses formed one group; bat and other animals the other group). Note, except of RaTG13 (MN996532.1) bat coronaviruses show no or little CpG depletion.

Supplementary files

Table S1. List of genomes used in the study

Table S2. Spreadsheet. Mutation analysis of SARS-CoV-2 and RaTG13 genomes.

Table 1. SNP analysis of 12 SARS-CoV-2 genomes

Accession	Country	Position[1]	Type of variation	Reference[2]	Allele
MT019529	China	3778	SNV	A	G
		8388	SNV	A	G
		8987	SNV	U	A
MT066175	Taiwan	**8782**	SNV	C	U
		28144	SNV	U	C
MT012098	India	2277	SNV	U	C
		6695	SNV	C	U
		14657	SNV	C	U
		17373	SNV	C	U
		22785	SNV	G	U
MT066156	Italy	2269	SNV	A	U
		26144	SNV	G	U
MT093571	Sweden	2717	SNV	G	A
		9274	SNV	A	G
		13225	SNV	C	G
		13226	SNV	U	C
		17376	SNV	A	G
		23952	SNV	U	G
		26144	SNV	G	U
MT007544	Australia	19065	SNV	U	C
		22303	SNV	U	G
		26144	SNV	G	U
MT126808	Brazil	**11083**	SNV	G	U
		14805	SNV	C	U
		17247	SNV	U	C
		26144	SNV	G	U
MN985325	USA	**8782**	SNV	C	U
		18060	SNV	C	U
		28144	SNV	U	C
MT184913	USA	**11083**	SNV	G	U
		28916	SNV	G	A
MT106053	USA	24325	SNV	A	G
MT020880	USA	**8782**	SNV	C	U
		18060	SNV	C	U
		28144	SNV	U	C

[1] Variants occuring in more than one genome are in bold

[2]Ref. genome: MN908947

References

Alinejad-Rokny H, Anwar F, Waters SA, Davenport MP, Ebrahimi D (2016) Source of CpG Depletion in the HIV-1 Genome. Mol Biol Evol 33:3205-3212

Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF (2020) The proximal origin of SARS-CoV-2. Nature Medicine. Published 17 March. https://doi.org/10.1038/s41591-020-0820-9

Bird A (2002) DNA methylation patterns and epigenetic memory. Genes Dev 16:6-21

Cheng X, Virk N, Chen W, Ji S, Ji S, Sun Y, Wu X (2013) CpG usage in RNA viruses: data and hypotheses. PLoS One 8:e74109

Jabbari K, Bernardi G (2004) Cytosine methylation and CpG, TpG (CpA) and TpA frequencies. Gene 333:143-149

Kandeel M, Ibrahim A, Fayez M, Al-Nazawi M (2020) From SARS and MERS CoVs to SARS-CoV-2: Moving toward more biased codon usage in viral structural and nonstructural genes. J Med Virol. https://doi.org/10.1002/jmv.25754

Klinman DM, Yamshchikov G, Ishigatsubo Y (1997) Contribution of CpG motifs to the immunogenicity of DNA vaccines. J Immunol 158:3635-3639

Koyama T, Platt D, Parida L (2020) Variant analysis of COVID-19 genomes. Bull World Health Organ. http://dx.doi.org/10.2471/BLT.20.25359

Kypr J, Mrazek J, Reich J (1989) Nucleotide composition bias and CpG dinucleotide content in the genomes of HIV and HTLV 1/2. Biochim Biophys Acta 1009:280-282

Poole A, Penny D, Sjoberg BM (2001) Confounded cytosine! Tinkering and the evolution of DNA. Nat Rev Mol Cell Biol 2:147-151

Roberts SA, Lawrence MS, Klimczak LJ, Grimm SA, Fargo D, Stojanov P, Kiezun A, Kryukov GV, Carter SL, Saksena G, Harris S, Shah RR, Resnick MA, Getz G, Gordenin DA (2013) An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. Nat Genet 45:970-976

Stevens M, Cheng JB, Li D, Xie M, Hong C, Maire CL, Ligon KL, Hirst M, Marra MA, Costello JF, Wang T (2013) Estimating absolute methylation levels at single-CpG resolution from methylation enrichment and restriction enzyme sequencing methods. Genome Res 23:1541-1553

Upadhyay M, Samal J, Kandpal M, Vasaikar S, Biswas B, Gomes J, Vivekanandan P (2013) CpG dinucleotide frequencies reveal the role of host methylation capabilities in parvovirus evolution. J Virol 87:13816-13824

Upadhyay M, Vivekanandan P (2015) Depletion of CpG Dinucleotides in Papillomaviruses and Polyomaviruses: A Role for Divergent Evolutionary Pressures. PLoS One 10:e0142368

Wan Y, Shang J, Graham R, Baric RS, Li F (2020) Receptor Recognition by the Novel Coronavirus from Wuhan: an Analysis Based on Decade-Long Structural Studies of SARS Coronavirus. J Virol 94. https://doi.org/10.1128/JVI.00127-20

Wrapp D, Wang N, Corbett KS, Goldsmith JA, Hsieh CL, Abiona O, Graham BS, McLellan JS (2020) Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science 367:1260-1263

Zhang T, Wu Q, Zhang Z (2020) Probable Pangolin Origin of SARS-CoV-2 Associated with the COVID-19 Outbreak. Curr Biol. https://doi.org/10.1016/j.cub.2020.03.022

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 (2020) A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579:270-273

Figures

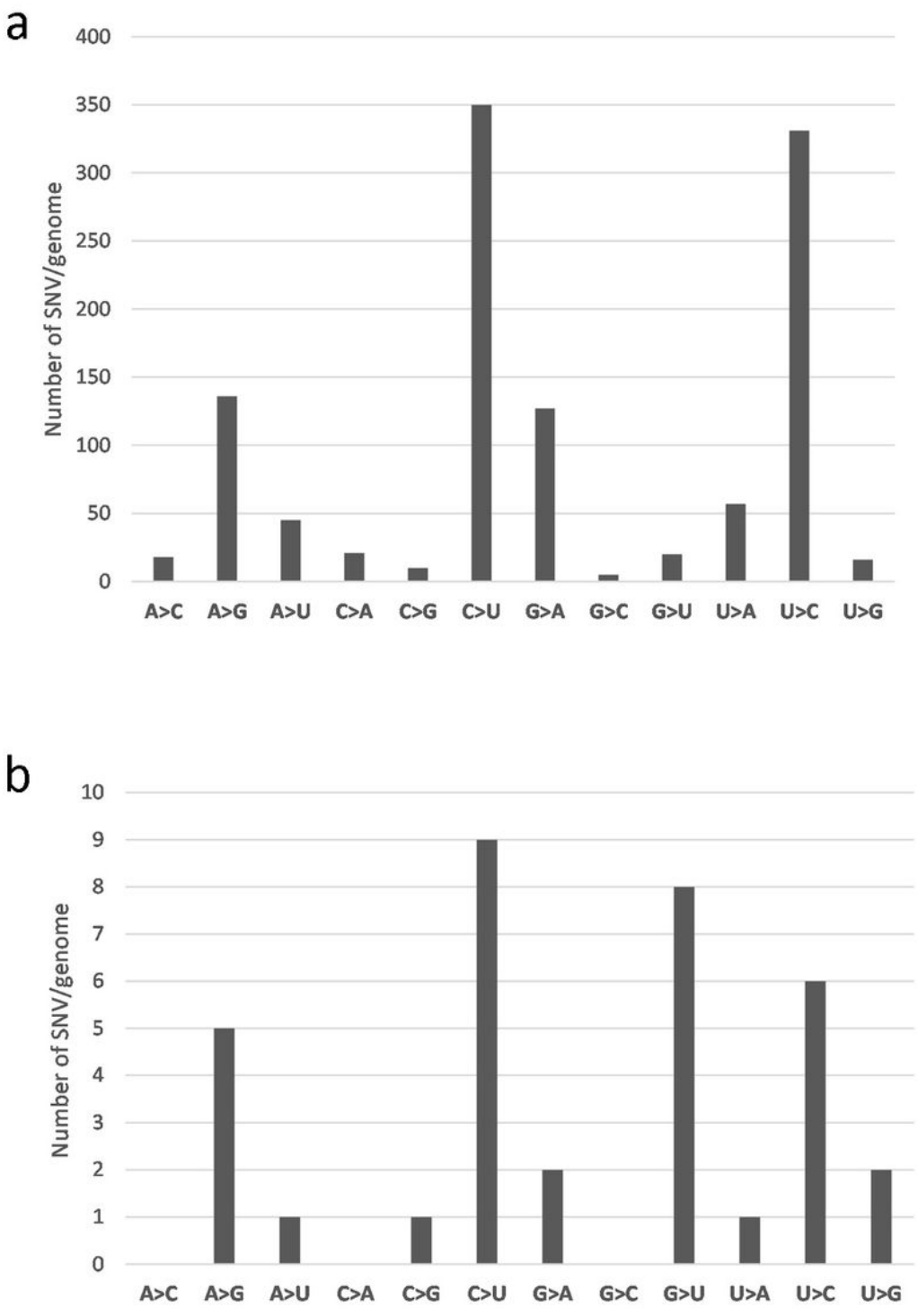

Figure 1

Frequencies of twelve types of nucleotide substitutions in coronaviruses. (a) SARS- CoV-2 and RaTG13 coronaviruses. The data are from Supplementary Table S1. (b) 12 human SARS-CoV-2 isolates. The data are from Table 1.

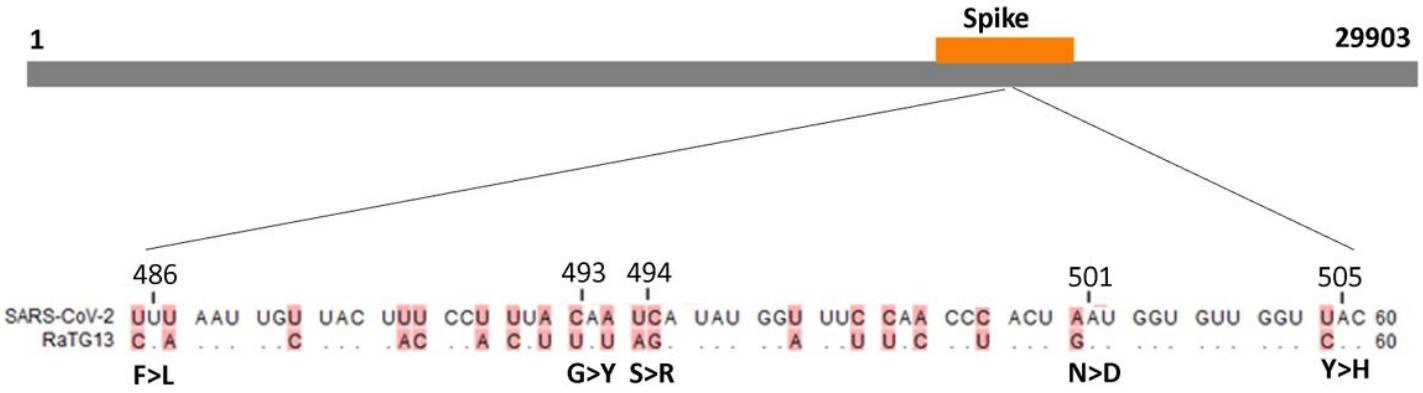

Figure 2

Coronarvirus genome with highlighted spike gene. A subregion encoding an ACE receptor contact domain (part) is magnified below. Differences between SARS-CoV-2 and bat RaTG13 sequences are highlighted. Changes in amino acid sequence are shown below the alignment. Numbers are according to the spike protein reading frame in the #MN908947 accession. Codons involving the C>U mutations are in bold. The direction of mutations is from the human to bat sequence.

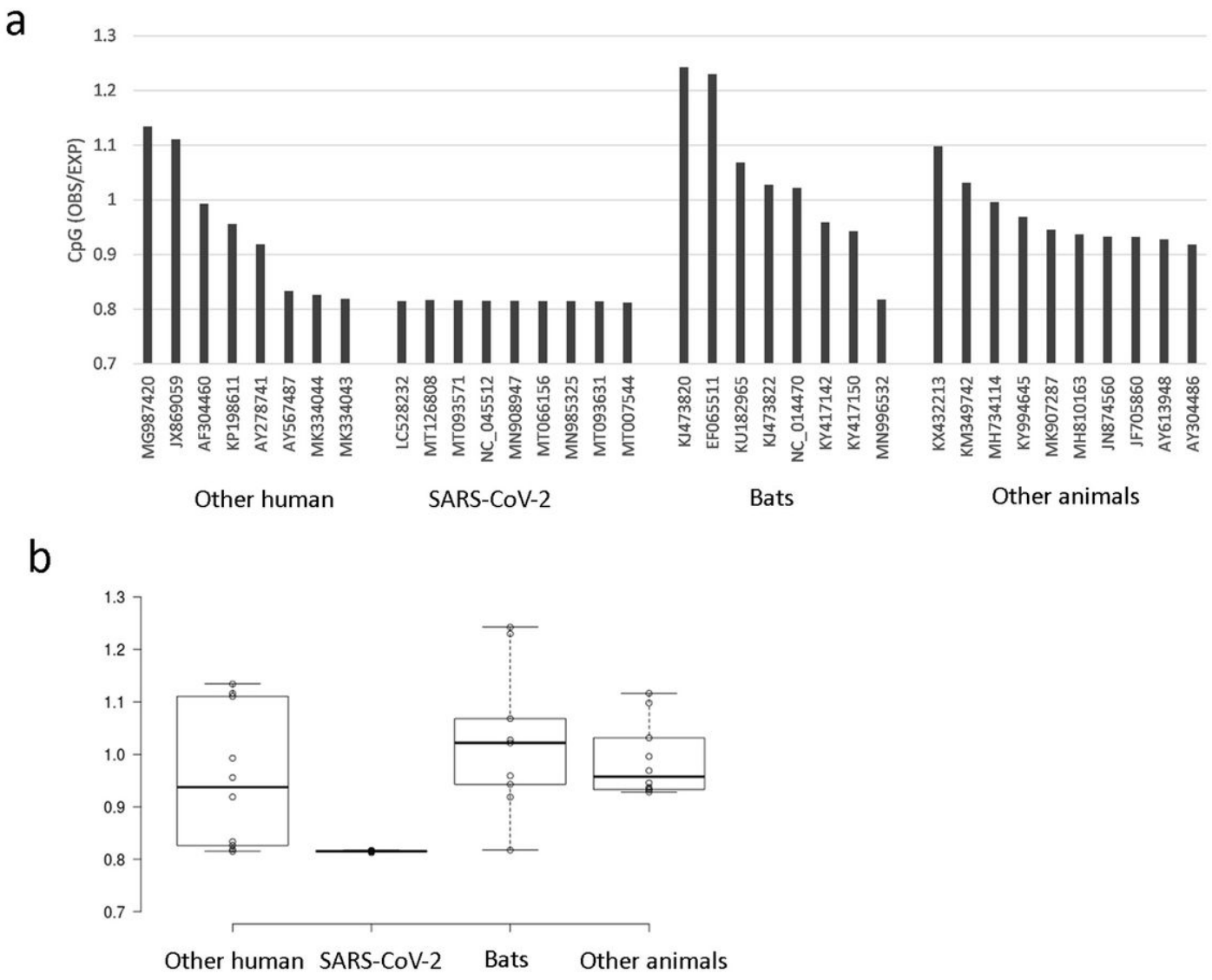

Figure 3

CpG depletion analysis in coronavirus genomes. (a) Bar charts showing CpG(OBS/EXP) in individual genomes. (b) Statistical representation of CpG depletions in individual groups. Differences between human and animal strains were significant (chi square, P<0.001, a single SARS-CoV-2 (#MN908947) plus all other human coronaviruses formed one group; bat and other animals the other group). Note, except of RaTG13 (MN996532.1) bat coronaviruses show no or little CpG depletion.

Supplementary Files

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

- TableS1.pdf
- TableS2b.xlsx