Check for updates

Received: 19 April 2021 | Revised: 26 April 2021 | Accepted: 27 April 2021

DOI: 10.1002/jmv.27056

SHORT COMMUNICATION

JOURNAL OF MEDICAL VIROLOGY WILEY

Betacoronaviruses genome analysis reveals evolution toward specific codons usage: Implications for SARS-CoV-2 mitigation strategies

Elisson N. Lopes[1] | Vagner Fonseca[2,3] | Diego Frias[4] | Stephane Tosta[1] | Álvaro Salgado[1] | Ricardo Assunção Vialle[5] | Toscano S. Paulo Eduardo[6] | Fernanda K. Barreto[7] | Vasco Ariston de Azevedo[1] | Michele Guarino[8] | Silvia Angeletti[9] | Massimo Ciccozzi[10] | Luiz C. Junior Alcantara[1,11] | Marta Giovanetti[1,11]

[1]Laboratório de Genética Celular e Molecular, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil

[2]KwaZulu-Natal Research Innovation and Sequencing Platform (KRISP), College of Health Sciences, University of KwaZuluNatal, Durban, South Africa

[3]Coordenação Geral dos Laboratórios de Saúde Pública/Secretaria de Vigilância em Saúde, Ministério da Saúde, Brasília, Distrito Federal, Brazil

[4]Departamento de Ciências Exatas e da Terra, Universidade do Estado da Bahia, Salvador, Bahia, Brazil

[5]Nash Family Department of Neuroscience & Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, New York, USA

[6]Laboratório de Biologia Molecular Aplicada, Departamento de Bioquímica, Universidade Federal do Rio Grande do Norte, Natal, Brazil

[7]Universidade Federal da Bahia, Vitória da Conquista, Bahia, Brazil

[8]Department of Gastrointestinal Diseases, Campus Bio-Medico University, Rome, Italy

[9]Unit of Clinical Laboratory Science, University Campus Bio-Medico of Rome, Rome, Italy

[10]Medical Statistic and Molecular Epidemiology Unit, University of Biomedical Campus, Rome, Italy

[11]Laboratório de Flavivírus, Instituto Oswaldo Cruz, Fundação Oswaldo Cruz, Rio de Janeiro, Brazil

Correspondence
Massimo Ciccozzi, Medical Statistic and Molecular Epidemiology Unit, University of Biomedical Campus, Rome 00128, Italy.
Email: M.ciccozzi@unicampus.it

Luiz Carlos Junior Alcantara, Laboratório de Genética Celular e Molecular, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte 21040-360, Minas Gerais, Brazil.
Email: alcantaraluiz42@gmail.com

Marta Giovanetti, Laboratório de Genética Celular e Molecular, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte 21040-360, Minas Gerais, Brazil.
Email: giovanetti.marta@gmail.com

Abstract

Since the start of the coronavirus disease 2019 (COVID-19) pandemic, the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has rapidly widespread worldwide becoming one of the major global public health issues of the last centuries. Currently, COVID-19 vaccine rollouts are finally upon us carrying the hope of herd immunity once a sufficient proportion of the population has been vaccinated or infected, as a new horizon. However, the emergence of SARS-CoV-2 variants brought concerns since, as the virus is exposed to environmental selection pressures, it can mutate and evolve, generating variants that may possess enhanced virulence. Codon usage analysis is a strategy to elucidate the evolutionary pressure of the viral genome suffered by different hosts, as possible cause of the emergence of new variants. Therefore, to get a better picture of the SARS-CoV-2 codon bias, we first identified the relative codon usage rate of all

Elisson Nogueira Lopes, Vagner Fonseca, Diego Frias, Massimo Ciccozzi, Luiz Carlos Junior Alcantara, Marta Giovanetti contributed equally to this study.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.
© 2021 The Authors. *Journal of Medical Virology* Published by Wiley Periodicals LLC

Betacoronaviruses lineages. Subsequently, we correlated putative cognate transfer ribonucleic acid (tRNAs) to reveal how those viruses adapt to hosts in relation to their preferred codon usage. Our analysis revealed seven preferred codons located in three different open reading frame which appear preferentially used by SARS-CoV-2. In addition, the tRNA adaptation analysis indicates a wide strategy of competition between the virus and mammalian as principal hosts highlighting the importance to reinforce the genomic monitoring to prompt identify any potential adaptation of the virus into new potential hosts which appear to be crucial to prevent and mitigate the pandemic.

KEYWORDS
codon deoptimization, codon usage, coronaviruses, COVID-19, SARS-CoV-2

1 | INTRODUCTION

Viral species members of the Coronaviridae family are enveloped by single-stranded positive-sense RNA viruses. Their genome encodes different nonstructural or accessory proteins that may differ according to the species and four structural proteins: envelope (E), nucleocapsid (N), membrane (M), and spike (S). The organization of spike protein across viral envelopes is responsible for the crown-shape that names the family.[1]

The Coronaviruses (CoVs) are organized into four genera: *Alphacoronavirus* and *Betacoronavirus* which have as natural hosts bats and rodents, and *Deltacoronavirus* along with *Gammacoronavirus* that are more frequently found in avian species.[2]

After the emergence of severe acute respiratory syndrome (SARS) and middle east respiratory syndrome (MERS),[1,3,4] the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which is the etiological agent of the coronavirus disease 2019 (COVID-19), is the third major coronavirus outbreak in the last 20 years.

COVID-19 may cause symptoms such as fever, cough, fatigue, and other severe complications leading to death.[4] According to the World Health Organization (WHO) updated in April 2021, more than 137 million people have been infected, causing more than 2.9 million deaths worldwide.[4]

SARS-CoV-2 has a natural host, bats, and a secondary one, probably a mammalian host, who was the key to originating the jumping species mutation needed for human infection.[3] Viruses with multiple host species such as the Coronaviruses evolve to successfully thrive under different hosts environments and available resources. Therefore, the virus may suit better with codons matching their hosts' codon usage.[3] This selective pressure may cause the emergence of mutations on SARS-CoV-2 to their hosts, one example is the variant identified in Minks Farm[5] (MT396266). Since then, due to the advanced whole genome sequencing technologies, an unprecedented number of genomes have been generated, providing invaluable insights into the ongoing evolution and epidemiology of the virus allowing the identification of hundreds of circulating genetic variants during the pandemic.

Currently, three variants (B.1.1.7 or VOC202012/01, B.1.351 or 20H/501Y.V2 and P.1) carrying several mutations in the receptor-binding domain (RBD) of the spike (S) protein, raise concerns about their potential to shift the dynamics and public health impact of the pandemic.[6-9]

Those variants of concern (VOCs) appear to share a common aspect: the viral adaptation to the human host, resulting in changeable effects on COVID-19 and complicating attempts to control the pandemic.[10,11] In addition, it should be noted that effective adaptation of CoVs to a new host needs not only such mutations affecting receptor binding but also a complete set of positive gene mutations that improve the reproduction and transmission of viruses in the new host. On this respect, here, using a codon usage bias (CUB) as an unequal frequency in the usage of synonymous codons we shield light on how SARS-CoV-2 acquired its adaptation to human host and provide insight regarding how other possible mammalian hosts might be ideal environments to promote viral infection.

2 | MATERIALS AND METHODS

2.1 | Data collection

We collected all fourteen (14) reference Betacoronavirus sequences from the National Center for Biotechnology Information (NCBI) Genbank. An *in house* R script was used to split the sequences in open reading frames (ORFs) (Table S1) and check for different read frames. Finally, we build a data set to represent all ORFs of Betacoronaviruses. After that, we collected the frequency of eight mammalian hosts available from the Codon Usage Database (http://www.kazusa.or.jp/codon/) and the transfer ribonucleic acid (tRNA) counts available from tRNA database (http://gtrnadb.ucsc.edu/), which were *Homo sapiens*, *Bostaurus*, *Canis familiaris*, *Equus caballus*, *Felis catus*, *Mus musculus*, *Mustela putorius furo*, *Rattus sp*. Using the relative synonymous codon usage (RSCU) formula we then calculated the RSCU of the hosts.

2.2 | RSCU

RSCU is a measure of nonuniform usage of synonymous codons in a sequence and it has been found to have causes and implications in RNA viruses.[12] Higher RSCU values indicate a higher bias toward a codon in detriment of its synonymous codon using codon metrics. To calculate RSCU, the observed codon value is divided by the expected codon value.

$$\mathrm{RSCU}_{i,j} = \frac{n_i x_{i,j}}{\sum_{k=1}^{n_i} x_{ik}}.$$

Hence, the maximum possible RSCU values are proportional to the number of synonymous codons. To compare the codon bias of all Betacoronaviruses and their hosts, we did a normalization of data with the RSCU values between 0 and 1. Where 1 is the higher value and 0 is the smaller value to RSCU. In addition to identifying codons more representative in viruses and less in humans (codon targets), we compared hosts and virus RSCU; (i) codons with a RSCU value close to 1 for coronavirus and (ii) lower than expected value to human host.

2.3 | Euclidean distance

We used Euclidean distance algorithms to identify putative relationships between the virus sequences and hosts. The construction of the Euclidean distance matrix was based on RSCU values calculated in previous section of each host and viruses, the analysis was performed using the following equation:

$$d_{ik} = \sum_{j=1}^{p} (x_{ij} - x_{kj})^2.$$

2.4 | Translational adaptation estimation method

The availability of tRNA was inferred from the hosts' genomes, counting the number of genes that encode each type of tRNA and taking into account the mechanism of tRNA sharing between synonymous codons ending with pyrimidines.[14]

We compare the hosts tRNA distribution with Betacoronavirus RSCU values; then we calculated the ratio from each host and virus using the following formula:

$$r_i = \frac{\mathrm{RSCU}_{i,\mathrm{virus}}}{f_{i,\mathrm{host}}},\ i = 1, 2, \ldots, 61.$$

Each i codon RSCU value is divided to f tRNA frequency, resulting in a group of codons with more disponibility to each host's tRNA pool.

After that, we calculated the relative distance to each host and virus based on the frequency of codons and tRNA:

$$D_{\mathrm{host}} = 100\frac{F_{\mathrm{abundant,host}} - F_{\mathrm{abundant,virus}}}{F_{\mathrm{abundant,host}}}.$$

Finally, we calculated a translational adaptation index (TAI) varying between 0% and 100% was measured as:

$$\mathrm{TAI} = 100 - D_{\mathrm{host}}.$$

3 | RESULTS

3.1 | Betacoronavirus codon usage

To investigate Betacoronaviruses' codon usage biases, we calculated RSCU values for each of the 14 genomes considering each ORFs individually. We noted that all viral genomes presented a similar rate with codons ending with A or T having higher RSCU values, whereas G and C end had the lower rates, as previously reported to RNA viruses.[13] Additionally, the top five codons with higher RSCU values on average represented ≥50% of the amino acids used in the process of translation and this feature was shared across almost all Betacoronaviruses, suggesting a possible coevolution and permanence of specific codons groups. This could indicate a clade bias, probably connected to a successful survival strategy (Table S2).

In addition, we created a matrix of RSCU values and calculated the Euclidean Distance across all Betacoronavirus, then we noticed a correlation between SARS-CoV-2 and SARS coronavirus (Table S3), which was expected based on their similarity as already described.[2]

After performing Betacoronavirus codon analysis, we focused on SARS-CoV-2 and we found that codons with the highest RSCU values (optimal codons) [TGT (Cys), GAA (Glu), TTT (Phe), CAA (Gln), and AAT (Asn)] were all AT-rich and that the human codon usage presented as the five most used codons [CAG (Gln), CAC (His), GAG (Glu), AAG (Lys), and TAC (Tyr)] were, in the majority of case, GC-rich (for more details see Table S2).

Soon after, we compared SARS-CoV-2 RSCU values to mammalian hosts (Figure 1), we search for codons targets which appear to be more important for the virus and less to human, to elucidate the adaptive mechanism. Figure 1 represents SARS-CoV-2 RSCU values close to 1, in red; and human RSCU codons values close to 0, in blue. We found seven codons, which appear to be more preferentially used by the virus than the human host, which are located in three distinct ORFs: CCG, ACG, CTC located in ORF10; GGC, TAC located in E; and GAC in M.

3.2 | Codons identification from relative frequency point of view

We searched in hosts' tRNA pools for tRNA corresponding to Betacoronavirus codons with higher RSCU codons, and our goals were found codons which appear to be more crucial to virus translation than for the hosts. Thus, correlating RSCU results for humans and SARS-CoV-2, we found four of the target codons as tRNA abundant and three as tRNA scarce (Table S4). After that, we used the TAI index to measure the adaptation scenario which will be able to explain how this newly emergent virus was able to adapt to the hosts in relation to their preferred codon usage (Table 1). These data compared the tRNA distribution and codon frequencies (for all virus TAI see Table S5).

Our results point out to an high SARS-CoV-2 adaptation, with TAIs values over 70% compared with all mammalian hosts tested suggesting that also other mammal host might be ideal environments

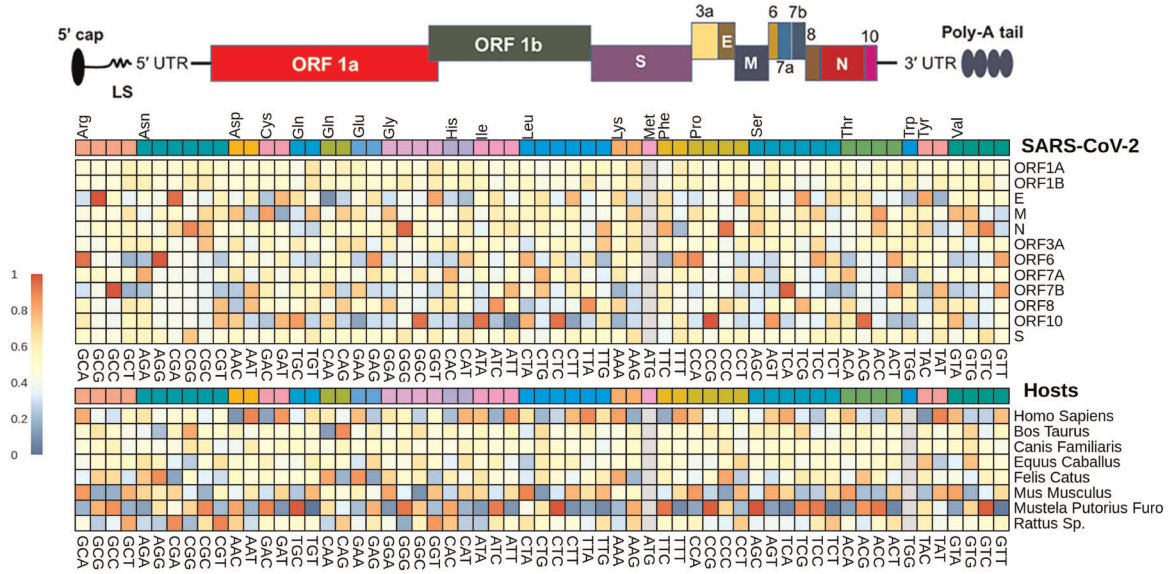

FIGURE 1 Graphical representation of synonymous codon usage pattern of each amino acid among SARS-CoV-2 and mammalian hosts. Open read frames of SARS-CoV-2 genome representation; Heatmap of observed RSCU values representing codon more used in red, and less used in blue. Rows are SARS-CoV-2 ORFS and hosts, columns are codons organized by amino acids. The values are normalized between 0 and 1 for comparison purposes. RSCU, SARS-CoV-2, severe acute respiratory syndrome coronavirus 2

to promote viral infection, highlighting the importance of strength the active monitoring to further elucidate adaptation of the virus to newly potential hosts.

4 | DISCUSSION

The global death toll from COVID-19 topped 2.9 million as April 2021,[4] crossing the threshold amid a vaccine rollout so immense but so uneven that in some countries there is real hope of vanquishing the outbreak, while in other, less-developed parts of the world, it seems a far-off dream. In this view, the identification of viral adaptation as well as the unbridled spread of this virus in a new host leading to the accumulation of mutations appear to be challenging to prevent the emergence of new SARS-CoV-2 variants of international concern. In this context, we analyzed the codon usage of diverse endemic and epidemic CoVs sequences to investigate the selective pressure that may cause the emergence of mutations on SARS-CoV-2 to their hosts. Codon usage carries a strategy for comparing sequences in a different way reflecting the viral evolutionary pressure suffered by the virus, the genetic drift and the natural selection for translational optimization.[12,15] In our analysis, we analyzed the codon pattern relating to Betacoronaviruses and their hosts, focusing on humans and SARS-CoV-2. We found a group of codons and classified them as targets, representing SARS-CoV-2 codons used more than expected compared with their human host. Our work brought to light seven codons and three ORFs as preferred in the SARS-CoV-2 selection. These codons present a nucleotide preference: A and T ending codons, which is in line with previous findings.[13] Using translational adaptation models, we further inferred SARS-CoV-2 capability to survive in mammalian hosts highlighting the importance to reinforce the genomic monitoring to prompt identify any potential adaptation of the virus into new potential hosts which appear to be crucial to prevent and mitigate the pandemic.

TABLE 1 Translational metrics for all hosts to SARS-CoV-2

Hosts	SARS-COV-2 TAI, %
Equus caballus	80.24
Bostaurus	77.00
Canis familiaris	76.83
Felis Catus	74.71
Mus musculus	74.71
*Homo Sapiens**	74.40
Mustela putorius furo	73.87
Rattus sp	73.00

Note: These data present the Euclidean distance between observed values and ideal values for viral and each host.

Abbreviations: SARS-COV-2, severe acute respiratory syndrome coronavirus 2; TAI, translational adaptation index.

ACKNOWLEDGEMENTS

This study was supported by the CNPq (421598/2018-2); MG is supported by Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro FAPERJ, Vagner Fonseca is supported by CAPES (88882.349290/2019-01) and a Flagship grant from the South African Medical Research Council (MRC-RFA-UFSP-01-2013/UKZN HIVEPI).

CONFLICT OF INTERESTS

The authors declare that there are no conflict of interests.

AUTHOR CONTRIBUTIONS

Conception and design: Elisson Nogueira Lopes, Vagner Fonseca, Diego Frias, and MG. *Performed the experiments*: Elisson Nogueira Lopes, Vagner Fonseca, Diego Frias, Stephane Tosta, Álvaro Salgado, Ricardo Assunção Vialle, and Paulo Eduardo Toscano Soares. *Data analysis*: Elisson Nogueira Lopes, Vagner Fonseca, Diego Frias, Stephane Tosta, Álvaro Salgado, Ricardo Assunção Vialle, Paulo Eduardo Toscano Soares, and MG. *Writing and revision*: Elisson Nogueira Lopes, Vagner Fonseca, Diego Frias, Fernanda Khouri Barreto, Vasco Ariston de Azevedo, Michele Guarino, Silvia Angeletti, Massimo Ciccozzi, Luiz Carlos Junior Alcantara, and Marta Giovanetti.

ORCID

Elisson N. Lopes http://orcid.org/0000-0003-1661-6304
Vagner Fonseca http://orcid.org/0000-0001-5521-6448
Ricardo Assunção Vialle http://orcid.org/0000-0003-3311-4197
Silvia Angeletti http://orcid.org/0000-0002-7393-8732
Massimo Ciccozzi http://orcid.org/0000-0003-3866-9239
Marta Giovanetti http://orcid.org/0000-0002-5849-7326

REFERENCES

1. Khailany RA, Safdar M, Ozaslan M. Genomic characterization of a novel SARS-CoV-2. *Gene Reports*. 2020;16:100682.
2. Chan JF, Kok K-H, Zhu Z, et al. Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. *Emerg Microbes Infect*. 2020;9(1):221-236.
3. Chan JFW, Lau SKP, To KKW, Cheng VCC, Woo PCY, Yuen K-Y. Middle east respiratory syndrome coronavirus: another zoonotic betacoronavirus causing SARS-like disease. *Clin Microbiol Rev*. 2015;28(2):465-522.
4. World Health Organization. *Coronavirus Disease (COVID-19) Pandemic*. WHO: Geneva, Switzerland; 2020. Accessed March 16, 2021.
5. Oreshkova N, Molenaar RJ, Vreman S, et al. SARS-CoV-2 infection in farmed minks, the Netherlands, April and May 2020. *Euro Surveill*. 2020;25(23):2001005. https://doi.org/10.2807/1560-7917.ES.2020.25.23.2001005
6. Rambaut A, Loman N, Pybus O, et al. Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations. https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563. Accessed December 21, 2020.
7. Tegally H, Wilkinson E, Lessells RJ, et al. Sixteen novel lineages of SARS-CoV-2 in South Africa [published online ahead of print February, 2 2021]. *Nat Med*. 2021. Preprint. 27(3):440–446. https://doi.org/10.1038/s41591-021-01255-3
8. Faria NR, Mellan TA, Whittaker C, et al. Genomic and epidemiology of the P.1SARS-CoV-2lineage in Manaus, *Science* 2021: eabh2644 https://doi.org/10.1126/science.abh2644. Accessed April 18, 2021.
9. Naveca F, Nascimento V, Souza V, et al. Phylogenetic relationship of SARS-CoV-2 sequences from amazonas with emerging Brazilian variants harboring mutations E484K and N501Y in the spike protein. Virological. https://virological.org/t/phylogenetic-relationship-of-sars-cov-2-sequences-from-amazonas-with-emerging-brazilian-variants-harboring-mutations-e484k-and-n501y-in-the-spike-protein/585. Accessed April 18, 2021.
10. Candido DS, Claro IM, de Jesus JG, et al. Evolution and epidemic spread of SARS-CoV-2 in Brazil. *Science*. 2020;369(6508):1255-1260. https://doi.org/10.1126/science.abd2161
11. Ramanathan M, Ferguson ID, Miao W, Khavari PA. SARS-CoV-2 B.1.1.7 and B.1.351 spike variants bind human ACE2 with increased affinity. *bioRxiv [Preprint]*. 2021. https://doi.org/10.1101/2021.02.22.432359
12. Sharp PM, Tuohy TMF, Mosurski KR. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res*. 1986;14(13):5125-5143.
13. Chen Z, Boon SS, Wang MH, Chan RWY, Chan PKS. Genomic and evolutionary comparison between SARS-CoV-2 and other human coronaviruses. *J Virol Methods*. 2021;289:114032. https://doi.org/10.1016/j.jviromet.2020.114032
14. Frias D, Monteiro-Cunha JP, Mota-Miranda AC, et al. Human retrovirus codon usage from tRNA point of view: Therapeutic insights. *Bioinform Biol Insights*. 2013;7:BBI.S12093-45.
15. Miller JB. Human viruses have codon usage biases that match highly expressed proteins in the tissues they infect. *Biomed Genet Genomics*. 2017;2:2. https://doi.org/10.15761/BGG.1000134

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Lopes EN, Fonseca V, Frias D, et al. et al. Betacoronaviruses genome analysis reveals evolution toward specific codons usage: implications for SARS-CoV-2 mitigation strategies. *J Med Virol*. 2021;1-5. https://doi.org/10.1002/jmv.27056