Manuscript Click here to view linked References

Large-scale analysis of synonymous viral variants reveals global adaptation of the SARS-CoV-2 to the human codon usage

Daniele Ramazzotti[1]†, Fabrizio Angaroni[2]†, Davide Maspero[2,3], Mario Mauri[1], Deborah D'Aliberti[1], Marco Antoniotti[2,4], Alex Graudenzi[3,4], Rocco Piazza[1,4]*

Affiliations:

[1]Dept. of Medicine and Surgery, Università degli Studi di Milano-Bicocca; Monza, Italy
[2]Dept. of Informatics, Systems and Communication, Università degli Studi di Milano-Bicocca; Milan, Italy
[3]Inst. of Molecular Bioimaging and Physiology, Consiglio Nazionale delle Ricerche (IBFM-CNR); Segrate, Milan, Italy
[4]Bicocca Bioinformatics, Biostatistics and Bioimaging Centre – B4; Milan, Italy

*Corresponding author. Email: rocco.piazza@unimib.it.

†Equally contributing authors.

Abstract

Many large national and transnational studies were dedicated to the analysis of SARS-CoV-2 genome. Most studies are focused on missense and nonsense mutations, however approximately 30% of the SARS-CoV-2 variants are synonymous, therefore changing the target codon without affecting the corresponding protein sequence. Here, by performing a large-scale analysis of more than 40000 SARS-CoV-2 genome sequences, we show that, over time, the virus is adapting its codon usage to that of the human host through the accumulation of silent mutations, thus likely improving its effectiveness in using the human aminoacyl-tRNA set. The sole exception to this rule is represented by mutations bearing the signature of the defensive APOlipoprotein B Editing Complex (APOBEC) machinery, which show a negative profile. Taken globally, this study indicates that codon usage adaptation may play a relevant role in the evolution of SARS-CoV-2, which appears to be much more complex than previously anticipated.

One-Sentence Summary

Synonymous SARS-CoV-2 mutations positively impact viral evolution by increasing adaptation to the human codon usage.

Main Text

The COVID-19 pandemic caused by the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) has currently affected approximately 102.1 million people, with 2.2 million COVID-related deaths and half a million new cases per day (1) (updated 2nd February 2021), hence posing a major health threat on a global scale.

This led to a surge of active research aimed at studying SARS-CoV-2 biology: many large national and transnational efforts were dedicated to the analysis of SARS-CoV-2 genome, with more than 100000 raw sequencing datasets currently available. These analyses succeeded in reconstructing the phylogeny of the virus and, at least in part, to implement lineage tracking strategies (2, 3, 4, 5, 6, 7). However, virtually none of the variants detected so far, with the few relevant exceptions of the p.D614G Spike mutation (8), and the mutations associated with lineages B.1.1.7, B1.1.28 and P1 (9, 10, 11), were unequivocally shown to play a functional role. This evidence led to the conclusion that the vast majority of the reported SARS-CoV-2 sequence changes are neutral or purifying, therefore having limited or no impact on the aggressiveness of COVID-19 pandemic.

To the best of our knowledge, most of the studies performed so far focused on missense and nonsense mutations, i.e., mutations that either cause a change in the sequence of the target protein or generate a premature termination codon. However, a large fraction of the SARS-CoV-2 mutations (approximately ~30%) are synonymous, therefore changing the target codon without affecting the associated protein sequence on mRNA translation, as opposed to missense/nonsense variants.

This bulk set of mutations are usually discarded during variant analyses, as they are considered as functionally irrelevant. However, it is well known from decades (12, 13, 14, 15) that silent mutations may have a profound impact on protein expression, by changing the aminoacyl-tRNA

molecule responsible for translating the mutated codon into the corresponding amino acid, a well-known process coordinated by the ribosomal machinery. A silent substitution causing a switch to a more abundant aminoacyl-tRNA is prone to increase the overall protein translating efficiency, while the opposite may occur if the tRNA pairing with the mutated codon is less abundant than the original one (16, 17, 18, 19). Therefore, we hypothesized that synonymous SARS-CoV-2 variants may play a functional role by adapting the sequence of the virus to the codon usage (CU) of the human host, thus improving its effectiveness in using the human aminoacyl-tRNA set, ultimately making the translation of viral particles more efficient.

To investigate this hypothesis, we analyzed a total of 40743 viral genomes and 239546 silent viral mutations identified in 9 different studies (see Materials and Methods), showing:

1) that at the global level, the adaptation of mutated codons to the human CU is significantly greater than the wild-type viral counterpart, with the sole exception of mutations bearing the APOlipoprotein B Editing Complex (APOBEC) signature, showing a negative profile, and

2) that the viral adaptation to the human codon usage is increasing over time during the course of the pandemic.

To analyze the possible adaptation of the viral codons to the human CU during the course of COVID-19 pandemic, we initially analyzed high-quality samples from NCBI BioProject: PRJEB37886 (Dataset 1). This dataset comprises a total of 27485 COVID-19 cases, where viral SARS-CoV-2 RNA was sequenced using an Amplicon strategy with high coverage. After alignment of the raw viral data to the SARS-CoV-2-ANC reference genome (5) and variant calling (see Materials and Methods for further details), silent variants were further annotated in terms of reference and mutated codon, variant frequency (VF) and sampling time (Supplementary File 1). The relative codon usage (RCU) adaptation was calculated as the ratio between the human CU of the mutated codon and the wild-type counterpart.

As previously shown by our team and by others (6, 20, 21), a relevant subset of viral mutations,

linked to the C>T signature, are generated by the activation of the human APOBEC machinery. APOBEC represents a family of cytidine deaminases playing a critical role in the intrinsic responses of the host to infections operated by a large set of viruses, among which retroviruses, herpesviruses, papillomaviruses, parvoviruses, hepatitis B and retrotransposons. APOBEC cytidine deaminases target the viral genome, introducing large numbers of C>T mutations, eventually causing the functional inactivation of the pathogen. Therefore, APOBEC mutations globally represent a defense raised by the host cells to suppress viral infection (22).

To define the contribution of the APOBEC machinery, we isolated APOBEC C>T | G>A mutations from the entire set of synonymous variants identified in Dataset 1 and we analyzed the relative codon usage profile of mutated vs. wild-type codons in this subgroup. Interestingly, we noticed that the APOBEC variants are associated with a slight decrease in the log2-transformed RCU (see Figure 1A and Supplementary Figure 1A, median log2-transformed RCU≈-0.172.); as opposite, the non-APOBEC (named as OTHER in the following) group showed a marked increase in the log2-transformed RCU (see Figure 1A and Supplementary Figure 1A, median log2-transformed RCU≈0.148; two-sided t-student p-value<2^{-16}), supporting the hypothesis that synonymous mutations not driven by APOBEC can indeed increase the adaptation of the viral sequences to the human CU.

To validate our findings, we considered two additional datasets; the first one (Dataset 2, Figure 1B and Supplementary Figure 1B) comprising 7 NCBI BioProject (accession numbers: PRJEB39849, PRJNA614995, PRJNA625551, PRJNA633948, PRJNA636748, PRJNA639066 and PRJNA645906) and the second one (Dataset 3, Figure 1C and Supplementary Figure 1C) including data from NCBI BioProject: PRJNA613958. All the analyses corroborated our hypothesis, confirming the increase in the log2-transformed RCU for the synonymous mutations belonging to the OTHER group.

Importantly, this process is not restricted to individual viral genes, as it is widespread across the

entire SARS-CoV-2 genome, suggesting that the virus may benefit from a global adaptation to the human CU (Figure 1, Supplementary Figures 1 and 2). Of note, these results do not change significantly when the CoCoPUTs human CU reference dataset of the Hive Lab (https://hive.biochemistry.gwu.edu/cuts/about; 23, 24) is used (Supplementary Figure 3) instead of the classical Kazusa one (https://www.kazusa.or.jp/codon/cgi-bin/showcodon.cgi?species=9606; 25).

In principle, a RCU more fit to the human translational machinery could allow mutated SARS-CoV-2 genomes to overcome translational bottlenecks, therefore outcompeting wild-type viruses in protein synthesis, ultimately improving the efficiency of viral packaging. Although a direct demonstration of this hypothesis is challenging, a greater RCU should correspond to an increased viral fitness, therefore potentially translating into a greater VF at the global level (See Figure 2). To pinpoint this phenomenon, we reasoned that the process of synonymous variants selection over time could be possibly detected in very large time-series analyses. Accordingly, we binned the silent variants based on their collection date in two-weeks intervals and we analyzed the trend of the median per-bin log2-transformed VF-weighted RCU. In line with the original hypothesis, silent variants of the OTHER group showed a significant increase of the log2-transformed VF-weighted RCU over time (See Figure 3, one-sided Mann-Kendall test, p value= 2.9^{-5}, 0.0013, and 0.0011 for Dataset 1, Dataset 2, and Dataset 3 respectively). As expected, no increase could be detected for the APOBEC variants (see Supplementary Figure 4).

Taken globally, these results suggest that silent mutations may play a role in viral evolution, by increasing the adaptation of the viral genome to the human codon usage. It is tempting to speculate that this continuous adaptation may underlie an increase of the overall efficiency of viral protein production and packaging, with viral genomes with a better adaptation being able to generate more viral particles over time, therefore outperforming other, less adapted, viruses. It is widely accepted that improvements in codon usage adaptation, by removing translation

bottlenecks, may contribute to the overall protein synthesis efficiency (26). Indeed, proteins differing only at silent sites can have several orders of magnitude of difference in their expression level (27). This finding has been widely exploited in both research and industry to adapt coding sequences to specific organisms to optimize protein synthesis. Recently, codon usage optimization was also proposed as an effective strategy to drive high levels of viral protein expression in human cells to stimulate a very robust immune response (28). Recent studies pointed also to more complex scenarios, where the use of different codons may allow to fine-tune the translation time of individual gene regions, to allow proper folding of complex domains as well as to maximize the occurrence of post-translational modifications (29).

As the number of available viral sequences is quickly increasing over time, it will be interesting to isolate individual silent variants showing relevant changes in their VF over large time spans in order to study specific codon usage adaptation patterns in the context of individual viral genes/proteins.

In conclusion, global as well as time-series analyses of silent mutations indicate that codon usage adaptation may play a relevant role in the evolution of SARS-CoV-2, suggesting that the evolution of the viral genome has been more intense than previously anticipated. Further studies will be required to thoroughly dissect the overall impact and clinical relevance of these findings. In the next future it will be important to assess if specific viral genomes with a high codon usage adaptation also show an increased infectiousness and/or clinical aggressiveness, which, at the time being, is very difficult, owing to the limited availability of clinical data in publicly available SARS-CoV-2 sequencing studies. In this regard and given the profound impact of COVID-19 pandemic, every effort should be made to share clinical information together with sequencing data, which is unfortunately rarely done.

References and Notes

1. WHO. (2021, 2 5). WHO weekly epidemiological update. https://www.who.int/publications/m/item/weekly-epidemiological-update---2-february-2021.
2. Forster, P., Forster, L., Renfrew, C., Forster, M. (2020). "Phylogenetic network analysis of SARS-CoV-2 genomes". *Proceedings of the National Academy of Sciences*, ***117***(17), 9241-9243.
3. Zehender, G., Lai, A., Bergna, A., Meroni, L., Riva, A., Balotta, C., et. al (2020). "Genomic characterization and phylogenetic analysis of SARS-COV-2 in Italy". *Journal of medical virology*, ***92***(9), 1637-1640.
4. Seemann, T., Lane, C. R., Sherry, N. L., Duchene, S., da Silva, A. G., Caly, L., et. al. (2020). "Tracking the COVID-19 pandemic in Australia using genomics". *Nature communications*, ***11***(1), 1-9.
5. Ramazzotti, D., Angaroni, F., Maspero, D., Gambacorti-Passerini, C., Antoniotti, M., Graudenzi, A., Piazza, R. (2021). "VERSO: a comprehensive framework for the inference of robust phylogenies and the quantification of intra-host genomic diversity of viral samples". *Patterns*, 100212.
6. Graudenzi, A., Maspero, D., Angaroni, F., Piazza, R., Ramazzotti, D. (2021). "Mutational signatures and heterogeneous host response revealed via large-scale characterization of SARS-CoV-2 genomic diversity". *Iscience*, 102116.
7. Lemieux, J. E., Siddle, K. J., Shaw, B. M., Loreth, C., Schaffner, S. F., Gladden-Young, A., et. al (2021). "Phylogenetic analysis of SARS-CoV-2 in Boston highlights the impact of superspreading events". *Science*, eabe3261.
8. Korber, B., Fischer, W. M., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., et. al

(2020). “Tracking changes in SARS-CoV-2 Spike: evidence that D614G increases infectivity of the COVID-19 virus”. *Cell*, ***182***(4), 812-827.

9. Muik, A., Wallisch, A. K., Sänger, B., Swanson, K. A., Mühl, J., Chen, W., et. al (2021). “Neutralization of SARS-CoV-2 lineage B. 1.1. 7 pseudovirus by BNT162b2 vaccine–elicited human sera”. *Science*, eabg6105.
10. Williams, T. C., Burgers, W. A. (2021). !SARS-CoV-2 evolution and vaccines: cause for concern?”. *The Lancet Respiratory Medicine*. doi: 10.1016/S2213-2600(21)00075-8.
11. Grubaugh, N. D., Hodcroft, E. B., Fauver, J. R., Phelan, A. L., Cevik, M. (2021). Public health actions to control new SARS-CoV-2 variants. *Cell*. doi: 10.1016/j.cell.2021.01.044.
12. Ikemura, T. (1981). “Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system”. *Journal of molecular biology*, ***151***(3), 389-409.
13. Ikemura, T. (1982). “Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes: differences in synonymous codon choice patterns of yeast and Escherichia coli with reference to the abundance of isoaccepting transfer RNAs”. *Journal of molecular biology*, ***158***(4), 573-597.
14. Sharp, P. M., Tuohy, T. M., Mosurski, K. R. (1986). “Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes”. *Nucleic acids research*, ***14***(13), 5125-5143.
15. Yang, Z., Nielsen, R. (1998).”Synonymous and nonsynonymous rate variation in nuclear genes of mammals”. *Journal of molecular evolution*, ***46***(4), 409-418.
16. Lavner, Y., Kotlar, D. (2005). “Codon bias as a factor in regulating expression via

translation rate in the human genome". *Gene*, ***345***(1), 127-138.

17. Duret, L. (2000). "tRNA gene number and codon usage in the C. elegans genome are co-adapted for optimal translation of highly expressed genes". *Trends in Genetics*, **16**(7), 287-289.
18. Comeron, J. M. (2004). "Selective and mutational patterns associated with gene expression in humans: influences on synonymous composition and intron presence". *Genetics*, ***167***(3), 1293-1304.
19. Kanaya, S., Yamada, Y., Kinouchi, M., Kudo, Y., Ikemura, T. (2001). "Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis". *Journal of molecular evolution*, ***53***(4-5), 290-298.
20. Woo, P. C., Wong, B. H., Huang, Y., Lau, S. K., Yuen, K. Y. (2007). "Cytosine deamination and selection of CpG suppressed clones are the two major independent biological forces that shape codon usage bias in coronaviruses". *Virology*, ***369***(2), 431-442.
21. Milewska, A., Kindler, E., Vkovski, P., Zeglen, S., Ochman, M., Thiel, V., Pyrc, K. (2018). "APOBEC3-mediated restriction of RNA virus replication". *Scientific reports*, ***8***(1), 1-12.
22. Stavrou, S., Ross, S. R. (2015). "APOBEC3 proteins in viral immunity". *The Journal of Immunology*, ***195***(10), 4565-4570.
23. Alexaki, A., Kames, J., Holcomb, D. D., Athey, J., Santana-Quintero, L. V., Lam, P. V. N., et. al (2019). "Codon and Codon-Pair Usage Tables (CoCoPUTs): facilitating genetic variation analyses and recombinant gene design". *Journal of molecular biology*, ***431***(13), 2434-2441.

24. Athey, J., Alexaki, A., Osipova, E., Rostovtsev, A., Santana-Quintero, L. V., Katneni, U., ... & Kimchi-Sarfaty, C. (2017). "A new and updated resource for codon usage tables". *BMC bioinformatics*, ***18***(1), 1-10.
25. Nakamura, Y., Gojobori, T., Ikemura, T. (2000). "Codon usage tabulated from international DNA sequence databases: status for the year 2000". *Nucleic acids research*, ***28***(1), 292-292.
26. Plotkin, J. B., & Kudla, G. (2011). "Synonymous but not the same: the causes and consequences of codon bias". *Nature Reviews Genetics*, ***12***(1), 32-42.
27. Kudla, G., Murray, A. W., Tollervey, D., Plotkin, J. B. (2009). "Coding-sequence determinants of gene expression in Escherichia coli". *science*, ***324***(5924), 255-258.
28. Stachyra, A., Redkiewicz, P., Kosson, P., Protasiuk, A., Góra-Sochacka, A., Kudla, G., Sirko, A. (2016). "Codon optimization of antigen coding sequences improves the immune potential of DNA vaccines against avian influenza virus H5N1 in mice and chickens". *Virology journal*, ***13***(1), 1-11.
29. Fu, J., Murphy, K. A., Zhou, M., Li, Y. H., Lam, V. H., Tabuloc, C. A., Liu, Y. (2016). "Codon usage affects the structure and function of the Drosophila circadian clock protein PERIOD". *Genes & development*, ***30***(15), 1761-1775.

Acknowledgments

We acknowledge that this research was partially supported by the Italian Ministry of University and Research (MIUR) - Department of Excellence project PREMIA (PREcision MedIcine Approach: bringing biomarker research to clinic)

Funding

Italian Ministry of University and Research (MIUR) - Department of Excellence project PREMIA (PREcision MedIcine Approach: bringing biomarker research to clinic)

Author contributions

Conceptualization: RP

Methodology: RP, DR, AG

Investigation: DR, FA, DM

Visualization: FA, DM, RP

Funding acquisition: RP, AG

Supervision: RP, AG

Writing – original draft: DR, FA, AG, RP

Competing interests

Authors declare that they have no competing interests.

Data and materials availability

All data are available in the main text or the supplementary materials or can be downloaded from the original SARS-CoV-2 sequence repositories.

Supplementary Materials

Materials and Methods

Figs. S1 to S4

Table S1

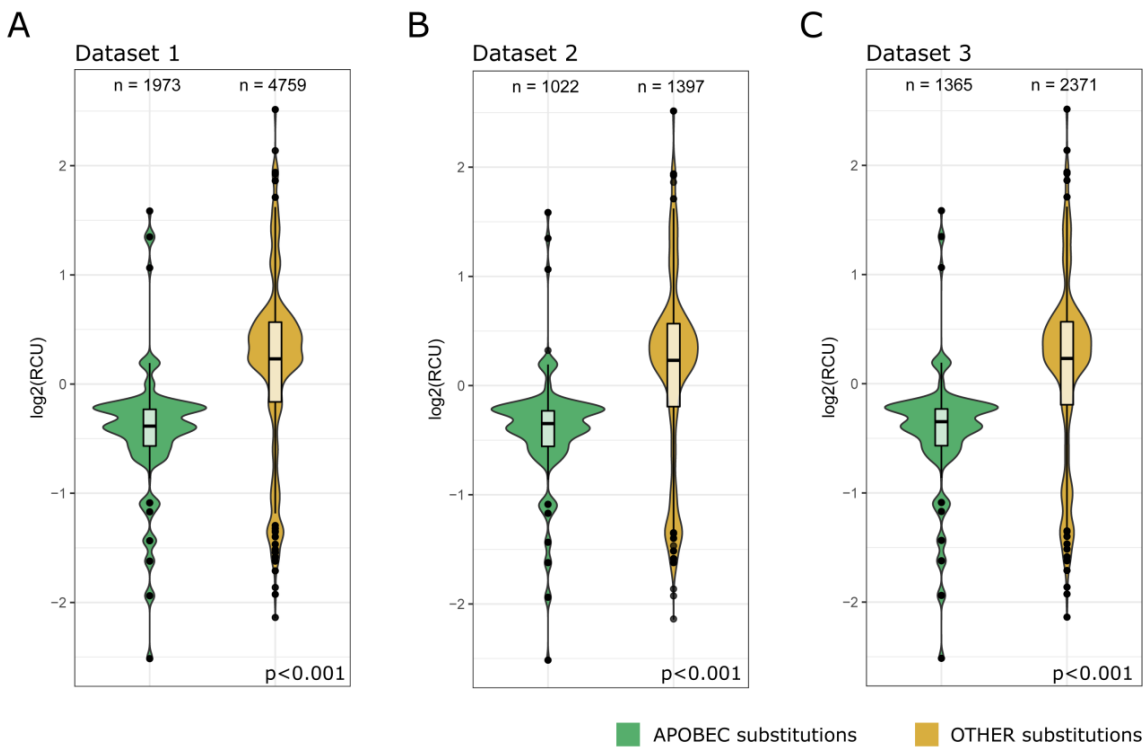

Figure 1 Empirical log2-transformed RCU distribution of distinct substitutions. Violin and box plots representing the empirical distribution of the log2-transformed RCU values for each mutation observed in (A) Dataset 1, (B) Dataset 2, and (C) Dataset 3. The substitutions were divided into two classes: APOBEC (i.e., G>A and T>C substitutions) and OTHER substitutions. The p-value of the two-sided t-student test is reported for every dataset.

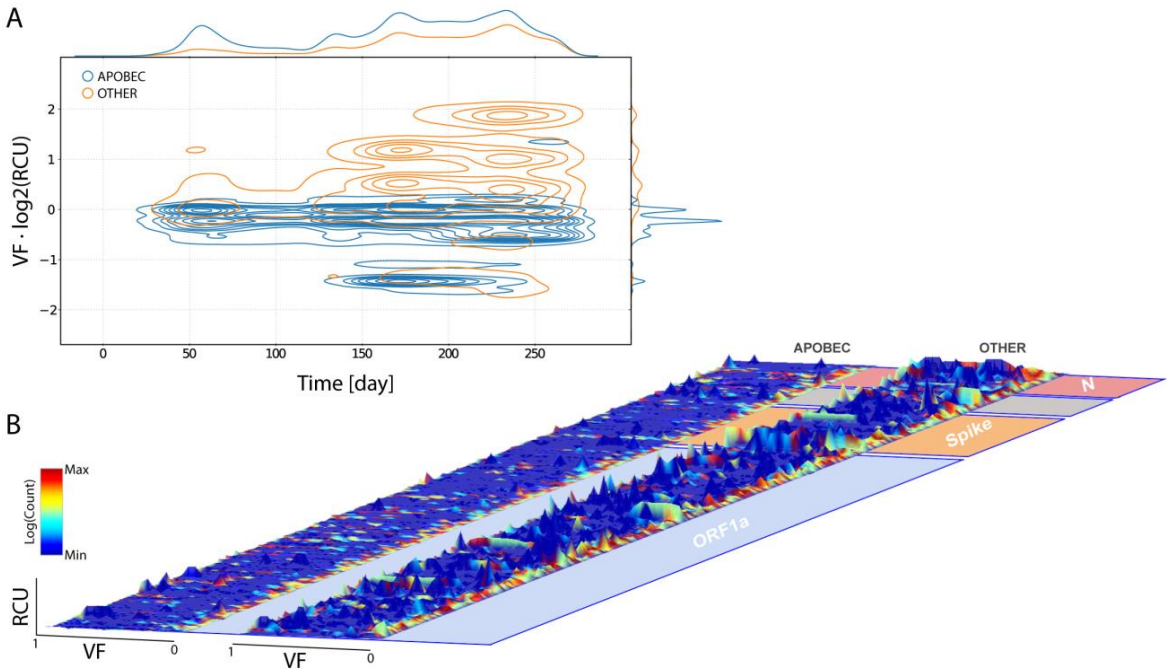

Figure 2 Density and surface RCU plots of synonymous SARS-CoV-2 variants divided into APOBEC and OTHER classes. A) Density plot of the VF-weighted log2-transformed RCU for the APOBEC (blue) and OTHER (orange) substitutions over the course of the disease. B) Surface plot representing the mean RCU of APOBEC and OTHER variants binned for VF and SARS-CoV-2 genome position. The log2 mean binned variant count is reported as a blue-to-red color gradient.

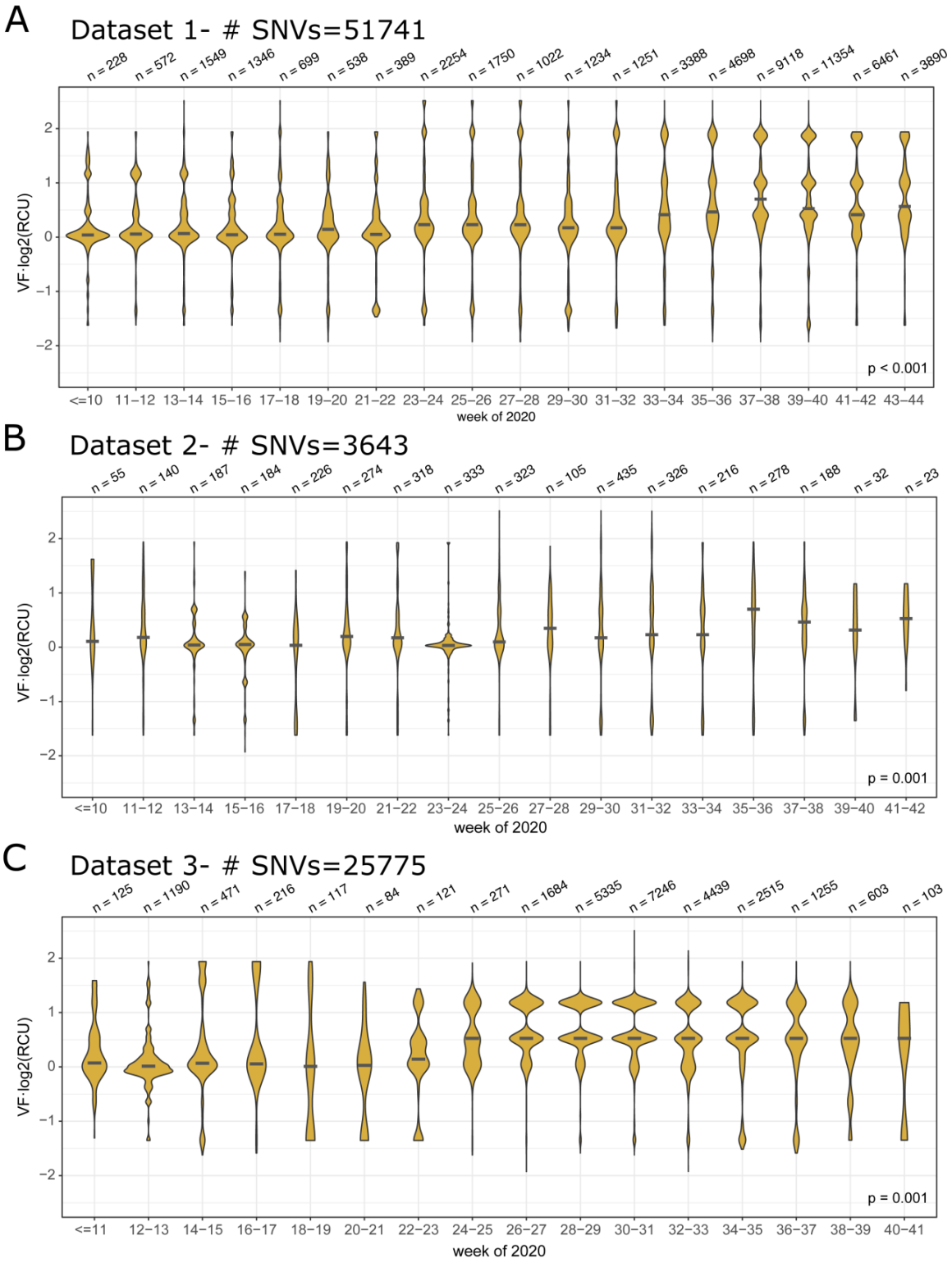

Figure 3 Temporal trend of the log2-transformed RCU of OTHER substitutions. The distribution of the log2-transformed RCU multiplied for the VF of every mutation not belonging

to the APOBEC group is presented with respect to collection date (binned every two weeks; grey lines represent the median values). The three datasets under study are presented in the different panels: (A) Dataset 1, (B) Dataset 2, and (C) Dataset 3. The p value of the one-sided Mann-Kendall trend test computed on median values is provided for each dataset.