viruses

MDPI

Article

The Mutation Profile of SARS-CoV-2 Is Primarily Shaped by the Host Antiviral Defense

Cem Azgari, Zeynep Kilinc, Berk Turhan, Defne Circi and Ogun Adebali *

Faculty of Engineering and Natural Sciences, Sabanci University, Istanbul 34956, Turkey; cemazgari@gmail.com (C.A.); zeynepkilinc@sabanciuniv.edu (Z.K.); berkturhan@sabanciuniv.edu (B.T.); defnecirci@sabanciuniv.edu (D.C.)

* Correspondence: oadebali@sabanciuniv.edu

Abstract: Understanding SARS-CoV-2 evolution is a fundamental effort in coping with the COVID-19 pandemic. The virus genomes have been broadly evolving due to the high number of infected hosts world-wide. Mutagenesis and selection are two inter-dependent mechanisms of virus diversification. However, which mechanisms contribute to the mutation profiles of SARS-CoV-2 remain under-explored. Here, we delineate the contribution of mutagenesis and selection to the genome diversity of SARS-CoV-2 isolates. We generated a comprehensive phylogenetic tree with representative genomes. Instead of counting mutations relative to the reference genome, we identified each mutation event at the nodes of the phylogenetic tree. With this approach, we obtained the mutation events that are independent of each other and generated the mutation profile of SARS-CoV-2 genomes. The results suggest that the heterogeneous mutation patterns are mainly reflections of host (i) antiviral mechanisms that are achieved through APOBEC, ADAR, and ZAP proteins, and (ii) probable adaptation against reactive oxygen species.

Keywords: SARS-CoV-2; COVID-19; evolution; mutation; phylogenetics; APOBEC; ROS; ZAP; ADAR

Citation: Azgari, C.; Kilinc, Z.; Turhan, B.; Circi, D.; Adebali, O. The Mutation Profile of SARS-CoV-2 Is Primarily Shaped by the Host Antiviral Defense. *Viruses* **2021**, *13*, 394. https://doi.org/10.3390/v13030394

Academic Editors: Concetta Castilletti, Luisa Barzon and Francesca Colavita

Received: 3 February 2021
Accepted: 24 February 2021
Published: 2 March 2021

1. Introduction

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has spread world-wide since its emergence in December 2019 [1], reportedly infecting more than 83 million people, with a death toll of 2,455,131 as of 22 February 2021, according to World Health Organization (WHO) (https://covid19.who.int/). Studies have been focused on effective treatment of the disease, mostly by the drug re-purposing approach due to the urgency [2] and by finding a vaccine that will stop the spread of the virus. Though there are dozens of vaccine candidates in clinical development, the evolutionary potential of the virus might affect the efficacy of the immunizations and treatments. Therefore, understanding the genomic features and mutation dynamics of the virus is crucial to interpret its evolutionary patterns and its response to the available treatments and potential vaccines.

Analyzing virus sequence context and mutations has revealed essential characteristics of SARS-CoV-2. For example, the origin of SARS-CoV-2 was linked to bats and pangolins using phylogenetic analyses [3–6]. Through mutational analyses, some genomic variants of the virus were associated with increased transmissibility [7,8]. In addition, we and others studied the spread of the virus in a variety of countries by tracking the mutation events of sequences over time [8–10].

Mutation profile analysis of SARS-CoV-2 can lead to the identification of mechanisms that drive the SARS-CoV-2 evolution; however, care should be taken when counting mutations to create a mutation profile. Considering that virus genomes are evolutionarily linked to each other, counting all the mutations in the sequences with respect to a reference genome creates a mutation bias towards the most abundant or frequently sequenced isolates. In other words, if a mutation occurs in an ancestral genome, it will also be seen in all of its descendants unless it reverts. When the mutations are called relative to a reference

genome, variants of a common origin will be counted multiple times, even though they are linked to a single mutation event. To overcome this issue, we created a phylogenetic tree and assigned only nucleotides that differ from the parent node as a mutation.

In this study, we retrieved SARS-CoV-2 genome sequences from the GISAID (Global Initiative on Sharing All Influenza Data) database [11] and analyzed the mutation profiles and sequence diversity of SARS-CoV-2.

2. Methods

2.1. Data Retrieval and Mutation Assignment

495,159 SARS-CoV-2 genomes, their pre-computed multiple sequence alignment, and metadata in the GISAID database, which was dated until 9 February 2021, were retrieved [11]. Initially, the pre-computed multiple sequence alignment was used for filtering undesired genomes. The low-quality sequences (5% NNNNs) and duplicates were removed by providers of pre-computed multiple sequence alignment; we filtered out the genomes with more than (i) 30 single point substitutions; or (ii) 200 inserted nucleotides; or (iii) 200 deleted nucleotides (relative to the reference genome). Next, the remaining genomes (381,048) were obtained from the unaligned genome sequences for further analyses. Because alignment and tree construction with more than 380,000 genomes was computationally intense, the genomes were randomly subsampled to 30,000 with a custom bash script and all the sequences with incomplete information (proper date or location of sample collection) in the metadata file were filtered out with a custom python script. Then, we used cd-hit to cluster sequences and choose representatives (-c 0.9999 -M 0 -T 80) [12]. 18,050 clusters were created, of which 16,122 contained only a single sequence. Then, the first sequence of each cluster was assigned as the representative of that cluster. Representative sequences were aligned with the MAFFT algorithm using Augur toolkit [13,14]. Wuhan-Hu-1 genome (GenBank: NC_045512.2) was chosen as the reference genome for the alignment. Then, a phylogenetic tree was constructed using IQ-TREE (-fast -n AUTO -m GTR).

The tree was then reconstructed into a time-resolved tree using the treetime option of Augur [13]. The sample with the earliest collection date among the representative sequences was chosen as the root, and marginal maximum likelihood estimation was used for date inference. The clock rate was applied across the genome to estimate the evolution rate and set to 0.0008, with a standard deviation of 0.0004 and using the date confidence flag to take the uncertainty of divergence time estimates into account. A constant coalescent model was chosen, and the "covariance-aware" mode of Augur was turned off with no covariance flag.

To assign the mutations to the nodes of the time-resolved tree, the ancestral option of Augur, which infers the ancestral sequences, was used by giving the time-resolved tree and the multiple sequence alignment of representative sequences as input (inference joint).

2.2. Mutation Profile Analysis

Mutation list was obtained from the phylogenetic tree and includes the mutations observed in each step of the tree. Then, this mutation list was divided into 192 groups based on their 12 mutation types (i.e., A > U, G > C) and 16 different trinucleotide contents where the mutating position is centered (i.e., A > U:UAA, G > C:AGU). Each of the 192 mutation groups were normalized with their corresponding trinucleotide count in the reference genome. Finally, these normalized mutation count values were plotted within the ggplot2 package [15] using R language, colored by their corresponding mutation type and trinucleotide content.

Observed mutations are first grouped by their position, mutation type, and trinucleotide, and frequently observed mutations (more than 8 times) at the same position in the same trinucleotide content are recorded. The observed mutations are grouped by their mutation type and trinucleotide, which resulted in 192 groups as indicated in the mutation

profile. The contribution of each mutation to its profile is calculated, and the ones which contributed more than 10% are reported.

2.3. Measuring Codon Changes and Codon Usage

By using ancestral mutations from the time-resolved tree, the mutated codons (labelled as deformed) were counted, while the number of the forming codons were referred to as formed codons. For each codon type, the ratio between formed and deformed count was taken and plotted in log2 scale by using the ggplot2 package [15].

Human codon usage table was retrieved [16], SARS-CoV-2 codon usage table was calculated with a custom R script. Number of occurrences in the reference genome was retrieved for each codon, then, they were grouped by their corresponding amino acids. The ratio of use per codon was calculated by dividing the occurrence of that codon to the sum of itself and its synonymous codons occurrence. Afterwards, the relative ratio of codon usage between *Homo sapiens* and SARS-CoV-2 was calculated by dividing the ratio of a codon in one genome to the sum of ratios in genomes.

2.4. Dinucleotide Changes

Observed mutations in the time-resolved tree were used to calculate the number of deformations observed for each dinucleotide. Dinucleotides were formed by these mutations, which were also calculated and recorded as the number of formations. Then, observed counts in the reference genome per codon were retrieved. The deformation counts were normalized by their division with their observation counts in the reference genome and plotted with their formation counts by ggplot2 in R studio [15].

3. Results

We reconstructed a phylogenetic tree and inferred the ancestral sequences of the viral genomes [17]. After constructing the tree, mutations were assigned based on the differences between the sequences and their parent node (Figure 1). This method enabled us to capture all the mutation events without recounting ancestral mutations. Moreover, we could also identify mutations that occurred repeatedly in different lineages, which would not be possible if the mutations were assigned relative to a reference genome.

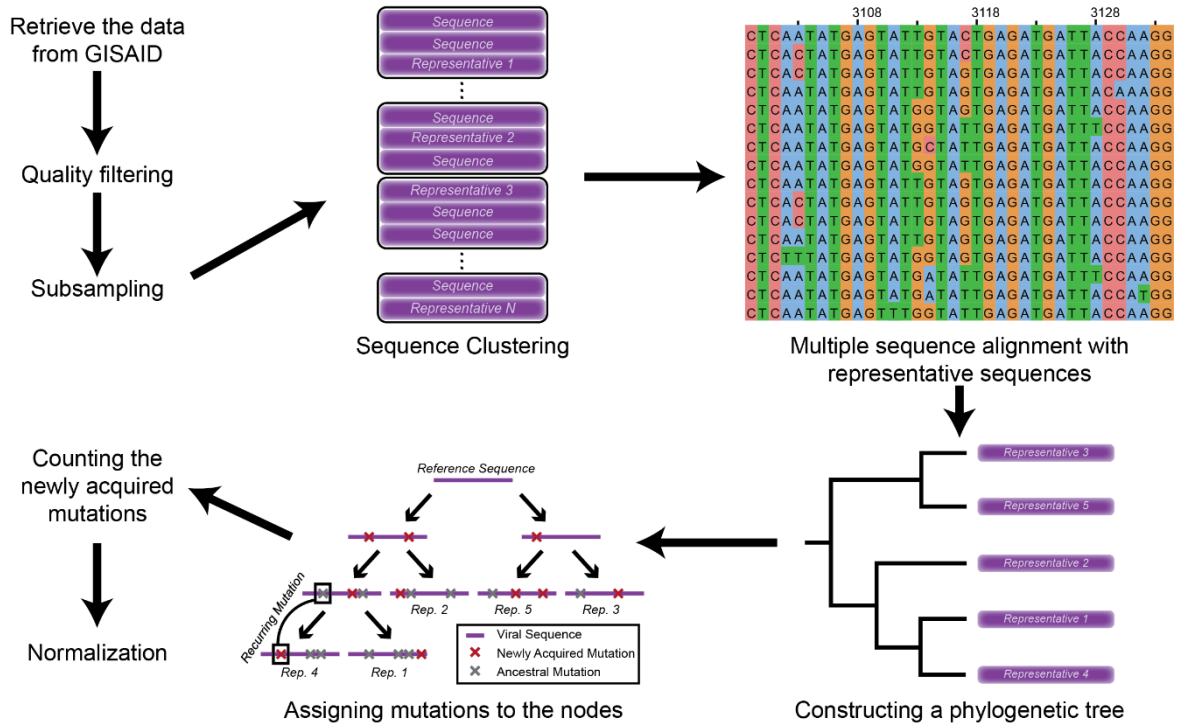

Figure 1. Schematic representation of the methodology.

3.1. Mutation Profile of SARS-CoV-2 and Potentially Related Mechanisms

To generate the mutation profile of SARS-CoV-2, we performed mutational signature analysis for all 192 trinucleotide changes using 54,353 mutations from the 33,540 representative sequences and nodes (Figure 2A). We normalized all the trinucleotide changes by the occurrence of the corresponding trinucleotide in the reference genome to eliminate any sequence context bias. In general, the most abundant mutational patterns are C > U, G > U, U > C, and A > G substitutions, that are 46%, 18.2%, 9.4%, and 8.8% of total substitutions, respectively (Figure 2A).

An enzyme family known for causing C > U substitution is called apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like (APOBEC) family. Enzymes of APOBEC family have an antiviral activity against some RNA viruses including coronaviruses [18–20]. Briefly, they can deaminate cytosine to thymine (uracil in the RNA genome), which can either result in C > U substitution on single-stranded viral RNA (plus strand) or G > A reflection if the C > U substitution occurs on a complementary strand (minus strand). In agreement with previous studies, the impact of APOBEC is highly visible at C > U substitutions, while it is relatively low at G > A substitutions (7.2% of total substitutions) [21–23]. This result suggests an asymmetric activity for APOBEC enzymes in favor of single-stranded viral RNA. Because virus RNA is frequently present as the plus strand, we see the effect of the APOBEC activity majorly in the form of C > U substitution relative to G > A, which reflects APOBEC activity on the negative strand during RNA replication. Moreover, APOBEC proteins show target inclination towards 5′-[T/U]C-3′ and 5′–CC–3′ motifs while deamination of cytosine [24]. Target sequence preferences of APOBEC proteins are observed in our mutational profile, where 5 out of 7 highest normalized mutational counts on C > U distributes along 5′-UC-3′ and 5′–CC–3′ motifs (Figure 2A). It is also experimentally found that A1CF RNA editing cofactor, which is APOBEC1 complementation factor, is among the SARS-CoV-2 RNA binders [25] that strengthens the hypothesis of APOBEC proteins' activity on the C > U substitutions.

The second most prevalent substitution is G > U, which might be associated with reactive oxygen species (ROS) in APOBEC-related manner. A recent study revealed that DNA damage response mediated by APOBEC3A (a member of APOBEC family) results in ROS production [26]. ROS can induce oxidative DNA damage, usually transforming guanine into 7,8-dihydro-8-oxo-20-deoxyguanine (oxoguanine), which can pair with adenine and lead to G > U substitution [27,28]. However, to date, there is no direct evidence of ROS-caused damage in the SARS-CoV-2 genome.

Another mechanism that can mutate the viral genome is adenosine deaminase acting on RNA (ADAR), which is an enzyme that mediates deamination of adenine to inosine (A > I) and later changes to guanine (A > G) [21]. A > G (plus strand) and U > C (minus strand) substitutions are observed at similar levels (8.8%, and 9.4% of total substitutions, respectively) (Figure 2A). ADAR targets dsRNA, and therefore, equivalent levels of ADAR activity are expected to be present at both strands [21]. The symmetric mutation profile for this pattern strongly suggests that ADAR working on replication RNA is effective in A > G and U > C substitutions.

In the context of trinucleotides, mutations dominantly occurred in U(C > U)G, C(C > U)G, A(C > U)G, U(C > U)U, and A(C > U)U (Figure 2A). Notably, 3 out of the 5 most frequently changed trinucleotides contain CG at their second and third positions. To examine whether these mutations were predominantly located at a single position in the viral genome or are distributed throughout the genome, we identified dynamic positions, where more than 8 recurring mutations were observed. Afterwards, we investigated the contribution of these trinucleotide positions to the mutation profile (Figure 2B,C). With some exceptions, most mutations in dynamic positions do not dominate the overall mutation profile. One of the exceptions is G(G > C)G mutations occurred at position 28,883, which correspond to the 66.6% of all mutations occurring on GGG. Although the percentage is high, the number of mutations occurring on GGG trinucleotides is only 9. Similarly G(A > U)C mutations at position 29,869 correspond to the 29.4% of all mutations

occurring on GAC trinucleotide, but the number of mutations of GAC is as low as 17. However, when trinucleotides with a total mutation number exceeding 500 are considered, the position with the highest mutation becomes position 11,083 with U(G > U)U mutations, composing 16.5% of total mutations of UGU. In conclusion, the mutation profile is not dominated by the switching positions; a position bias on mutation distribution is only observable when the total number of mutations of a trinucleotide is low. Several mutations labeled as impactful on signatures (Figure 2B) have been investigated for their possible effect on the severity and transmissibility of the virus by various studies. The highest mutational position, 11,083, has been associated with the severity of the virus [29]. Together with 11,083, mutations at 23,403, 21,575, 28,881, and 28,883 positions [30–33] have been associated with significant indication towards selection. In particular, the D614G mutation on Spike protein is associated with the fitness of the virus by both computational and clinical studies [32–34].

Figure 2. Mutation profile of SARS-CoV-2 genomes. Mutation counts are normalized by the trinucleotide content for each trinucleotide generated from 33,540 (representative sequences and nodes) sequences (**A**). From unstable positions where more than 8 mutations of the same type at the same position, highly occurred mutations are retrieved. The occurrences of these mutations are divided by the total number of the mutations of the same type and observed in the same trinucleotide. The calculated ratio is used to visualize the impact of highly occurring mutations on signature profile, as percentages (**B**). The bars are labelled with their positions in the genome. Number of mutations observed in the phylogenetic tree, per position (**C**). Mutations which have a significant contribution to their signature visualized in part B are marked according to the mutation count they represent at the position. Marked mutations are colored and reshaped with respect to their mutation type and impact range on their signatures.

3.2. Codon Usage of SARS-CoV-2 Differentiates in Favor of A and U Containing Codons

We investigated the impact of a potential contribution of codon bias selection on the mutation profile. First, we counted all the formed and altered codons, which we referred to as "form" and "deform", respectively. We calculated the relative difference between

form and deform values for each codon to test potential convergence of virus genome to host codon usage through mutations (Figure 3A). While UUU, AUA, AUU, and UAU are the intensively formed codons, CCA, UGG, GCU, and ACA are the most diminished ones. These results indicate a dominant forming of A and U containing trinucleotides, whereas G and C containing trinucleotides tend to reduce in number. In addition, all the codons that are translated into alanine (A) and proline (P) tend to diminish, resulting in lower translation of these amino acids in viral proteins. Considering that all the codons of A and P contain GC and CC in the first and second position, respectively, the reduction in these amino acids is probably related to selection against G and C presence (Figure 3A).

Human coronaviruses are known to have low GC content (GC%), and SARS-CoV-2 is not an exception with ~38 GC% [35,36]. Moreover, it was suggested that the reduction in GC% is an adaptation strategy of SARS-CoV-2, particularly towards the codon usage of the genes expressed in the human lung [37]. To determine whether the mutations of SARS-CoV-2 is an adaptation strategy to increase its viability inside the host or just the byproduct of host immune response to the viral RNA, we obtained the human codon usage values [16] and calculated the codon usage of SARS-CoV-2 (see methods). We calculated the relative ratio of these values and grouped codons that are translating the same amino acid (Figure 3B). If the viral genome is to adapt to the host genome, one can hypothesize that the codons that are used dominantly in the host relative to the virus should be formed in the viral genome to increase the similarity, while the percentage of codons that are used dominantly in the virus should decrease. GCU, GAA, GGU, and CGU codons that are used relatively high in SARS-CoV-2, have the tendency to deform, in agreement with the hypothesis. However, UGU, AUA, UUA, and GUU codons that are also used relatively high in SARS-CoV-2 have the tendency to be formed. A similar contradiction is also observed in the codons that are highly used in the human genome. In general, adaptation to the host codon usage does not explain the formation tendency of the codons. The main driver of the formation tendency is likely to be selection pressure against GC%, and thus, A and U increase.

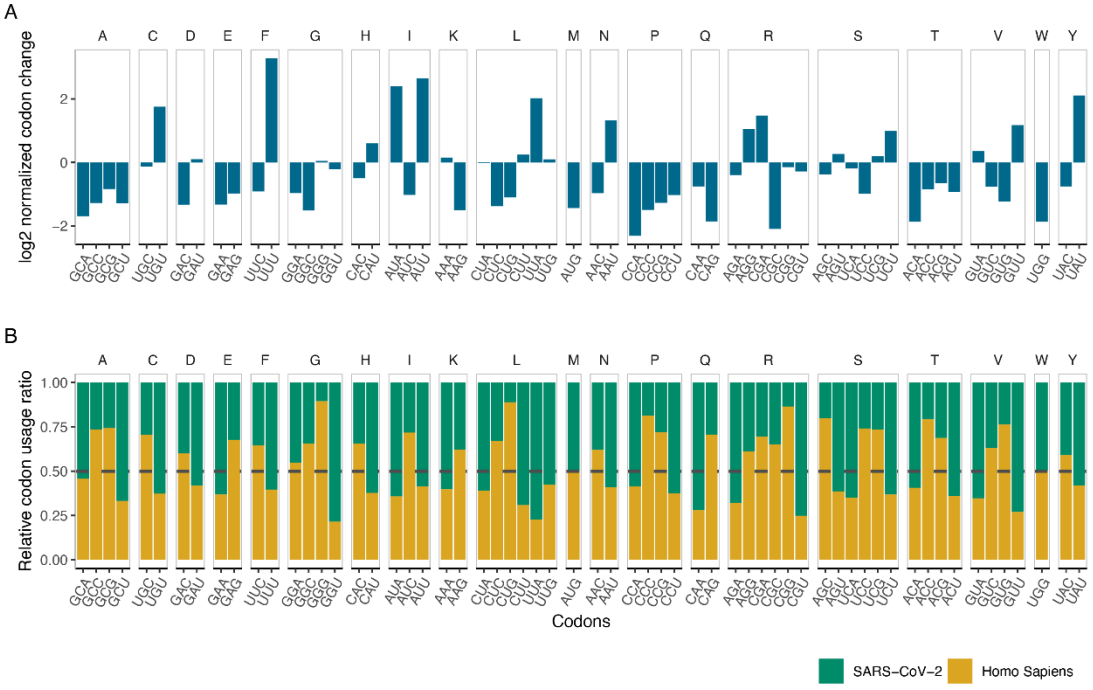

Figure 3. Comparison of codon variations in SARS-CoV-2 phylogenetic tree, human genome, and SARS-CoV-2reference genome. (**A**) Codon variations from mutations in the phylogenetic tree, represented by the ratio of formations over deformations per codon in log2 formation. (**B**) Relative codon usage percentage between Human and SARS-CoV-2 reference genome.

3.3. CG Nucleotide Deforms, While UU Nucleotide Forms

After observing excessive mutations in trinucleotides that contain CG at their second and third positions, and higher deformation in G and C containing codons, we examined the deformation (Figure 4A) and formation (Figure 4B) of dinucleotides. Because deformation of a dinucleotide is dependent on its occurrence in the genome, we normalized the deformed value of each dinucleotide with respect to its occurrence in the reference genome. As suggested by others [36,38], CG dinucleotide is the most deformed among all (Figure 4A). Xia et al. attributed the reduction in CG dinucleotide to a protein called zinc finger antiviral protein (ZAP), which binds and mediates the degradation of the viral genome [36]. This study indicates that SARS-CoV-2 is the most CG deficient betacoronavirus [36]. Thus, high CG deformation might be an adaptation of SARS-CoV-2 to escape ZAP under high purifying selection. In addition, UU dinucleotide is formed more than all other dinucleotides. In general, A and U containing dinucleotides are formed, meanwhile C and G containing dinucleotides are deformed.

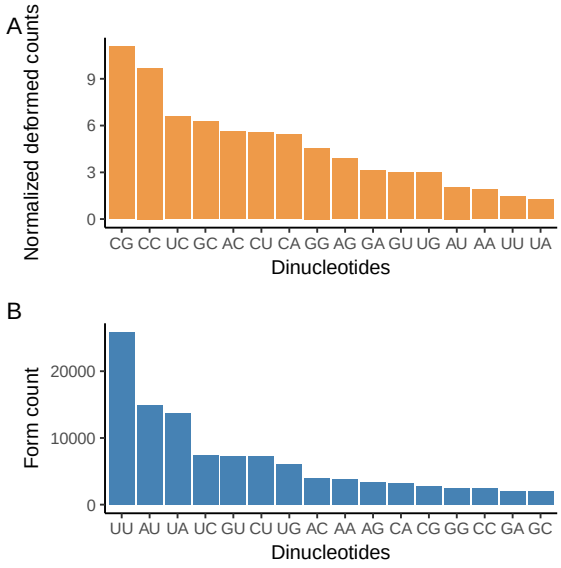

Figure 4. Comparison of dinucleotide formations and deformations retrieved from phylogenetic trees. Deformation ratio of dinucleotides is represented as the ratio of deformation count in the tree over dinucleotide's abundance in the reference genome (**A**). As a result of the mutations, the relative dinucleotides are formed (**B**).

4. Discussion

The COVID-19 pandemic has been spreading aggressively, killing thousands of people and affecting the daily lives of many more. Moreover, the evolutionary behavior of SARS-CoV-2 might potentially weaken the efficiency of the current treatments and vaccines. Here, we performed a phylogenetic tree-based mutational analysis to assess the contribution of mutagenesis and selection mechanism to SARS-CoV-2 mutation profiles.

The mutation profile of SARS-CoV-2 revealed that C > U, G > U, U > C, and A > G are the predominant substitutions. Based on these mutational patterns, we compiled some potential mechanisms that might be influencing the SARS-CoV-2 viral genome (Figure 5), which are namely APOBEC, ADAR, and ZAP. These mechanisms were linked to SARS-CoV-2 mutagenesis in previous studies as well [22,36,39,40]. In addition, we suspect that ROS might be a driver of G > U substitutions, however, more studies should be conducted to link ROS to SARS-CoV-2.

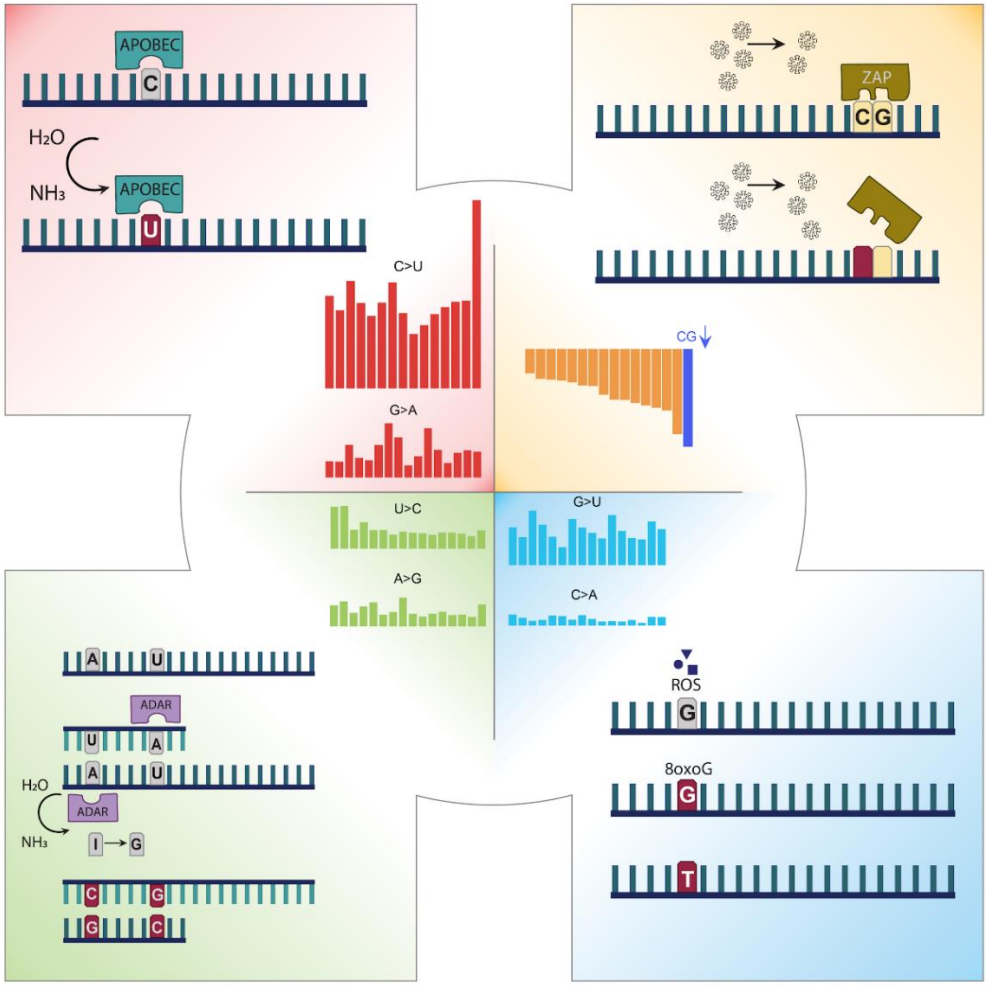

Figure 5. Mechanisms that can alter the sequence context of SARS-CoV-2. (i) APOBEC-caused mutations correlated with the enzyme signature dominantly on the plus RNA strand; (ii) ADAR-caused mutations equivalently affecting both RNA strands due to its mechanism of action; (iii) drop of CG dinucleotide targeted by ZAP through selection; (iv) ROS effect shown on the plus RNA strand.

Another aim of this study was to examine the main driver of the mutational patterns of SARS-CoV-2; whether the viral genome is inclined to converge into the host genome or the mechanisms we have discussed are the only contributors to the mutational patterns. Analyses on formed and deformed codons exhibit an increase of A and U and a decrease of G and C containing codons. Furthermore, the comparison between human and SARS-CoV-2 codon usage does not reveal a strong correlation between codon usage percentages and SARS-CoV-2 formation tendency. These results combined suggest that SARS-CoV-2 genome diverges through RNA editing mechanisms of the host, independently of any adaptive mechanism to increase its genomic similarity to the host genome, which was suggested in another study as well [37]. Then, we examined the formation tendency of dinucleotides. In general, we observed a decrease of G and C, and an increase of A and U containing dinucleotides. Strikingly, the deform rate of CG dinucleotides and formation of UU dinucleotides are extremely high. This phenomenon, which was observed in most human viruses [41], was previously associated with the reduction of the hydrogen bonds between strands to achieve more efficient gene expression [38].

In conclusion, the mutational profile we generated supported the potential biological mechanisms contributing to the genome diversity of SARS-CoV-2 genomes. Strand asymmetry of some mutation signatures suggested the mechanism acting on the plus RNA strand only. Strand-wise equivalent mutation signature attributed to ADAR is in

agreement with its mechanism of action where RNA is affected in the double-strand form. Antiviral responses and selection cannot be distinguished from each other. Host responses against the virus cause mutations in one hand, and the reduced targets in the virus genome make it less susceptible to the same antiviral attacks. Although we don't suggest a direct antiviral mechanism to reduce CG content, the reduced CG content can be explained by an adaptation to the host antiviral mechanism by ZAP. So far, the virus has been affected by the host antiviral mechanisms. Although there are several Spike protein amino acid substitutions that are likely to provide a selection advantage [8,42], selection hasn't been the major driving force of the genome-wide mutagenesis until the date of data collection. In the coming months, with a wide administration of the vaccines, it might be possible to see the effect of the vaccination and selection pressure by observing amino acid changes providing an advantage in escaping from immunized hosts.

Author Contributions: Conceptualization, O.A.; methodology, C.A., O.A.; software, C.A., Z.K., B.T.; formal analysis, investigation, C.A., Z.K., B.T., and D.C.; resources, O.A.; writing—original draft preparation, C.A.; writing—review and editing, visualization, C.A., Z.K., B.T., D.C., and O.A.; supervision, project administration, funding acquisition, O.A. All authors have read and agreed to the published version of the manuscript.

Funding: This study was partly supported by EMBO Installation Grant (grant 4163), which is funded by Scientific and Technological Research Council of Turkey (TÜBİTAK).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All the codes and the processed data are publicly available at https://github.com/CompGenomeLab/SARS-CoV-2_Mutational_Profile.

Acknowledgments: We thank Stuart James Lucas for his helpful feedback. Z.K., B.T., and D.C. were supported by TUBITAK STAR program. O.A. is supported by the Science Academy, Turkey (BAGEP, Young Scientist Award).

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Hu, B.; Guo, H.; Zhou, P.; Shi, Z.L. Characteristics of SARS-CoV-2 and COVID-19. *Nat. Rev. Microbiol.* **2020**, 1–14.
2. Shah, B.; Modi, P.; Sagar, S.R. In silico studies on therapeutic agents for COVID-19: Drug repurposing approach. *Life Sci.* **2020**, *252*, 117652. [CrossRef] [PubMed]
3. Li, C.; Yang, Y.; Ren, L. Genetic evolution analysis of 2019 novel coronavirus and coronavirus from other species. *Infect. Genet. Evol.* **2020**, *82*, 104285. [CrossRef] [PubMed]
4. Wu, F.; Zhao, S.; Yu, B.; Chen, Y.-M.; Wang, W.; Song, Z.-G.; Hu, Y.; Tao, Z.-W.; Tian, J.-H.; Pei, Y.-Y.; et al. Author Correction: A new coronavirus associated with human respiratory disease in China. *Nature* **2020**, *580*, E7. [CrossRef]
5. Zhou, P.; Yang, X.-L.; Wang, X.-G.; Hu, B.; Zhang, L.; Zhang, W.; Si, H.-R.; Zhu, Y.; Li, B.; Huang, C.-L.; et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **2020**, *579*, 270–273. [CrossRef]
6. Zhu, N.; Zhang, D.; Wang, W.; Li, X.; Yang, B.; Song, J.; Zhao, X.; Huang, B.; Shi, W.; Lu, R.; et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N. Engl. J. Med.* **2020**, *382*, 727–733. [CrossRef]
7. Hou, Y.J.; Chiba, S.; Halfmann, P.; Ehre, C.; Kuroda, M.; Dinnon, K.H.; Leist, S.R.; Schäfer, A.; Nakajima, N.; Takahashi, K.; et al. SARS-CoV-2 D614G variant exhibits efficient replication ex vivo and transmission in vivo. *Science* **2020**, *370*, 1464–1468. [CrossRef] [PubMed]
8. Volz, E.; Hill, V.; McCrone, J.T.; Price, A.; Jorgensen, D.; O'Toole, Á.; Southgate, J.; Johnson, R.; Jackson, B.; Nascimento, F.F.; et al. Evaluating the Effects of SARS-CoV-2 Spike Mutation D614G on Transmissibility and Pathogenicity. *Cell* **2020**, *184*, 64–75. [CrossRef] [PubMed]
9. Adebali, O.; Bircan, A.; Circi, D.; İşlek, B.; Kilinc, Z.; Selcuk, B.; Turhan, B. Phylogenetic analysis of SARS-CoV-2 genomes in Turkey. *Turk. J. Biol.* **2020**, *44*, 146–156. [CrossRef]
10. Popa, A.; Genger, J.W.; Nicholson, M.D.; Penz, T.; Schmid, D.; Aberle, S.W.; Agerer, B.; Lercher, A.; Endler, L.; Colaço, H.; et al. Genomic epidemiology of superspreading events in Austria reveals mutational dynamics and transmission properties of SARS-CoV-2. *Sci. Transl. Med.* **2020**, *12*, 573. [CrossRef]

11. Shu, Y.; McCauley, J. GISAID: Global initiative on sharing all influenza data-from vision to reality. *Euro. Surveill.* **2017**, *22*, 30494. [CrossRef] [PubMed]
12. Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* **2012**, *28*, 3150–3152. [CrossRef] [PubMed]
13. Hadfield, J.; Megill, C.; Bell, S.M.; Huddleston, J.; Potter, B.; Callender, C.; Sagulenko, P.; Bedford, T.; Neher, R.A. Nextstrain: Real-time tracking of pathogen evolution. *Bioinformatics* **2018**, *34*, 4121–4123. [CrossRef]
14. Katoh, K.; Standley, D.M. A simple method to control over-alignment in the MAFFT multiple sequence alignment program. *Bioinformatics* **2016**, *32*, 1933–1942. [CrossRef]
15. Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis*; Springer: New York City, NY, USA, 2016.
16. Nakamura, Y.; Gojobori, T.; Ikemura, T. Codon usage tabulated from international DNA sequence databases: Status for the year 2000. *Nucleic Acids Res.* **2000**, *28*, 292. [CrossRef] [PubMed]
17. Sagulenko, P.; Puller, V.; Neher, R.A. TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evol.* **2018**, *4*, vex042. [CrossRef] [PubMed]
18. Harris, R.S.; Dudley, J.P. APOBECs and virus restriction. *Virology* **2015**, *479–480*, 131–145. [CrossRef]
19. Sharma, S.; Patnaik, S.K.; Taggart, R.T.; Kannisto, E.D.; Enriquez, S.M.; Gollnick, P.; Baysal, B.E. APOBEC3A cytidine deaminase induces RNA editing in monocytes and macrophages. *Nat. Commun.* **2015**, *6*, 6881. [CrossRef] [PubMed]
20. Woo, P.C.; Wong, B.H.; Huang, Y.; Lau, S.K.; Yuen, K.-Y. Cytosine deamination and selection of CpG suppressed clones are the two major independent biological forces that shape codon usage bias in coronaviruses. *Virology* **2007**, *369*, 431–442. [CrossRef] [PubMed]
21. Di Giorgio, S.; Martignano, F.; Torcia, M.G.; Mattiuz, G.; Conticello, S.G. Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2. *Sci. Adv.* **2020**, *6*, eabb5813. [CrossRef] [PubMed]
22. Graudenzi, A.; Maspero, D.; Angaroni, F.; Piazza, R.; Ramazzotti, D. Mutational signatures and heterogeneous host response revealed via large-scale characterization of SARS-CoV-2 genomic diversity. *BioRxiv* **2020**, 102116.
23. Simmonds, P. Rampant C–>U Hypermutation in the Genomes of SARS-CoV-2 and Other Coronaviruses: Causes and Consequences for Their Short- and Long-Term Evolutionary Trajectories. *mSphere* **2020**, *5*, 3. [CrossRef]
24. McDaniel, Y.Z.; Wang, D.; Love, R.P.; Adolph, M.B.; Mohammadzadeh, N.; Chelico, L.; Mansky, L.M. Deamination hotspots among APOBEC3 family members are defined by both target site sequence context and ssDNA secondary structure. *Nucleic Acids Res.* **2020**, *48*, 1353–1371. [CrossRef]
25. Schmidt, N.; Lareau, C.A.; Keshishian, H.; Ganskih, S.; Schneider, C.; Hennig, T.; Melanson, R.; Werner, S.; Wei, Y.; Zimmer, M.; et al. The SARS-CoV-2 RNA-protein interactome in infected human cells. *Nat. Microbiol.* **2020**, *6*, 339–353. [CrossRef]
26. Niocel, M.; Appourchaux, R.; Nguyen, X.-N.; Delpeuch, M.; Cimarelli, A. The DNA damage induced by the Cytosine Deaminase APOBEC3A Leads to the production of ROS. *Sci. Rep.* **2019**, *9*, 4714. [CrossRef] [PubMed]
27. Molteni, C.G.; Principi, N.; Esposito, S. Reactive oxygen and nitrogen species during viral infections. *Free Radic. Res.* **2014**, *48*, 1163–1169. [CrossRef]
28. Waris, G.; Ahsan, H. Reactive oxygen species: Role in the development of cancer and various chronic conditions. *J. Carcinog.* **2006**, *5*, 14. [CrossRef] [PubMed]
29. Toyoshima, Y.; Nemoto, K.; Matsumoto, S.; Nakamura, Y.; Kiyotani, K. SARS-CoV-2 genomic variations associated with mortality rate of COVID-19. *J. Hum. Genet.* **2020**, *65*, 1075–1082. [CrossRef] [PubMed]
30. Berrio, A.; Gartner, V.; Wray, G.A. Positive selection within the genomes of SARS-CoV-2 and other Coronaviruses independent of impact on protein function. *PeerJ* **2020**, *8*, e10234. [CrossRef]
31. Van Dorp, L.; Acman, M.; Richard, D.; Shaw, L.P.; Ford, C.E.; Ormond, L.; Owen, C.J.; Pang, J.; Tan, C.C.S.; Boshier, F.A.T.; et al. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. *Infect. Genet. Evol.* **2020**, *83*, 104351. [CrossRef] [PubMed]
32. Korber, B.; Fischer, W.M.; Gnanakaran, S.; Yoon, H.; Theiler, J.; Abfalterer, W.; Hengartner, N.; Giorgi, E.E.; Bhattacharya, T.; Foley, B.; et al. Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell* **2020**, *182*, 812–827. [CrossRef]
33. Yin, C. Genotyping coronavirus SARS-CoV-2: Methods and implications. *Genomics* **2020**, *112*, 3588–3596. [CrossRef]
34. Plante, J.A.; Liu, Y.; Liu, J.; Xia, H.; Johnson, B.A.; Lokugamage, K.G.; Zhang, X.; Muruato, A.E.; Zou, J.; Fontes-Garfias, C.R.; et al. Spike mutation D614G alters SARS-CoV-2 fitness. *Nature* **2020**, 1–6. [CrossRef] [PubMed]
35. Berkhout, B.; van Hemert, F. On the biased nucleotide composition of the human coronavirus RNA genome. *Virus Res.* **2015**, *202*, 41–47. [CrossRef]
36. Xia, X. Extreme Genomic CpG Deficiency in SARS-CoV-2 and Evasion of Host Antiviral Defense. *Mol. Biol. Evol.* **2020**, *37*, 2699–2705. [CrossRef] [PubMed]
37. Li, Y.; Yang, X.; Wang, N.; Wang, H.; Yin, B.; Yang, X.; Jiang, W. GC usage of SARS-CoV-2 genes might adapt to the environment of human lung expressed genes. *Mol. Genet. Genom.* **2020**, *295*, 1537–1546. [CrossRef] [PubMed]
38. Wang, Y.; Mao, J.M.; Wang, G.D.; Luo, Z.P.; Yang, L.; Yao, Q.; Chen, K.P. Human SARS-CoV-2 has evolved to reduce CG dinucleotide in its open reading frames. *Sci. Rep.* **2020**, *10*, 1–10. [CrossRef] [PubMed]
39. Kosuge, M.; Furusawa-Nishii, E.; Ito, K.; Saito, Y.; Ogasawara, K. Point mutation bias in SARS-CoV-2 variants results in increased ability to stimulate inflammatory responses. *Sci. Rep.* **2020**, *10*, 17766. [CrossRef]

40. Klimczak, L.J.; Randall, T.A.; Saini, N.; Li, J.L.; Gordenin, D.A. Similarity between mutation spectra in hypermutated genomes of rubella virus and in SARS-CoV-2 genomes accumulated during the COVID-19 pandemic. *PLoS ONE* **2020**, *15*, e0237689. [CrossRef]
41. Caudill, V.R.; Qin, S.; Winstead, R.; Kaur, J.; Tisthammer, K.; Pineda, E.G.; Solis, C.; Cobey, S.; Bedford, T.; Carja, O.; et al. CpG-creating mutations are costly in many human viruses. *Evol. Ecol.* **2020**, *34*, 339–359. [CrossRef]
42. Chand, M. *Investigation of Novel SARS-COV-2 Variant: Variant of Concern 202012/01*; Public Health England: London, UK, 2020.