APOBEC-mediated Editing of SARS-CoV-2 Genomic RNA Impacts Viral Replication and Fitness

One-sentence summary: Editing of SARS-CoV-2 genomic RNA by Host APOBEC Enzymes and Its Impacts on the Viral Replication and Emergence of New Strains in COVID-19 Pandemic

Kyumin Kim[1], Peter Calabrese[2], Shanshan Wang[1], Chao Qin[3], Youliang Rao[3], Pinghui Feng[3], Xiaojiang S. Chen[1,4,5,6 *]

[1]Molecular and Computational Biology Section, University of Southern California, Los Angeles, CA 90089, USA;

[2]Quantitative and Computational Biology Section, University of Southern California, Los Angeles, CA 90089, USA;

[3]Section of Infection and Immunity, Herman Ostrow School of Dentistry, Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, CA 90089, USA;

[4]Genetic, Molecular and Cellular Biology Program, Keck School of Medicine, University of Southern California, Los Angeles, CA 90089, USA;

[5]Center of Excellence in NanoBiophysics, University of Southern California, Los Angeles, CA 90089, USA;

[6]Norris Comprehensive Cancer Center; University of Southern California, Los Angeles, CA 90089, USA.

*To whom correspondence should be addressed.
Tel: +1 213 740-5487;   FAX: +1 213 740 4340;
Email: xiaojiac@usc.edu

ABSTRACT

During COVID-19 pandemic, mutations of SARS-CoV-2 produce new strains that can be more virulent and evade vaccines. Viral RNA mutations can arise from misincorporation by RNA-polymerases and modification by host factors. Recent SARS-CoV-2 sequence analyses showed a strong bias toward C-to-U mutation, suggesting that host APOBEC cytosine deaminases with immune functions may cause the mutation. We report the experimental evidence demonstrating that APOBEC3A and APOBEC1 can efficiently edit SARS-CoV-2 RNA to produce C-to-U mutation at specific sites. However, APOBEC-editing does not inhibit the viral RNA accumulation in cells. Instead, APOBEC3A-editing of SARS-CoV-2 promotes viral replication/propagation, suggesting that SARS-CoV-2 utilizes the APOBEC-mediated mutations for fitness and evolution. Unlike the unpredictability of random mutations, this study has significant implications in predicting the potential mutations based on the UC/AC motifs and surrounding RNA structures, thus offering a basis for guiding future antiviral therapies and vaccines against the escape mutants.

INTRODUCTION

The causative agent of the COVID-19 pandemic, the severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2), is a member of the enveloped Coronaviridae family that has a single-stranded positive-sense RNA genome (*1, 2*). Unlike most RNA viruses that exhibit high mutation rates ((*3*) and references therein), SARS-CoV-2 and other coronaviruses have moderate genetic variability because they have proofreading machinery to correct the errors caused by their RNA-dependent RNA-polymerase (RdRP) (*4*).

However, enormous sequencing data of SARS-CoV-2 revealed its persistent accumulation of new mutations, leading to the emergence of new virus strains that can be more transmissible or virulent (*5, 6*) or evade vaccines (*7, 8*), highlighting the importance of understanding the evolution of SARS-CoV-2 genome. There are two main sources for SARS-CoV-2 viral mutations: Spontaneous random errors that are not corrected by the build-in proofreading mechanism and the Host-driven viral genome mutations ((*3*) and references therein). Recent analysis of SARS-CoV-2 and rubella vaccine virus genome data show predominant mutational patterns with specific signatures rather than random genetic variations, suggesting that the host-factor induced mutations play an important role in shaping the viral genomic RNA mutational outcome and evolution (*9-12*).

The host responses that can cause mutations on SARS-CoV-2 include reactive oxygen species (ROS) ((*13*) and references therein) and two families of human RNA deaminases: ADARs (the adenosine deaminases acting on RNA) and APOBECs (the apolipoprotein-B (*ApoB*) mRNA editing enzyme, catalytic polypeptide-like proteins) (*9, 10, 12*). The ROS could oxidize nucleic acids to cause viral mutations, which is proposed to be related to the G-to-U and C-to-A mutations (*3, 14*). The ADAR enzymes modify adenosine to inosine to cause A-to-G mutations in double-stranded RNA (dsRNA), which play important roles in immune regulation ((*15, 16*) and references therein).

The APOBEC proteins are a family of cytosine deaminases that can deaminate cytosine to uracil (C-to-U) in single-stranded nucleic acids and function in a variety of biological processes, including innate and adaptive immune responses to viral pathogens ((*17-20*) and references therein). The seven APOBEC3 subfamilies (including A3A to A3H) are reported to have restriction activities to DNA and RNA viruses (reviewed in (*19, 20*)). While most APOBECs use single-stranded DNA (ssDNA) as substrate for C-deamination (reviewed in (*17*)), three APOBECs, APOBEC1 (A1) (*21, 22*), APOBEC3A (A3A) (*23*), and APOBEC3G (A3G) (*24*), are also shown to deaminate certain cellular single-stranded RNA (ssRNA) targets to cause C-to-U editing. Interestingly, the database analysis of SARS-CoV-2 genomic variations showed an overwhelmingly high C-to-U mutation rate, account for about 40% of all single nucleotide variations, which was interpreted as a result of RNA editing by APOBECs rather than random mutations (*9, 10, 12, 25*). However, there is no report on the direct experimental evidence to demonstrate if APOBECs can edit the SARS-CoV-2 genome and, if so, what is the effect of the APOBEC-editing on the virus. In this study, we investigated whether APOBEC proteins can directly edit the RNA sequence of SARS-CoV-2 to generate C-to-U mutations and how such mutations may impact viral replication and viral progeny production in an experimental system.

RESULTS

Experimental design for testing APOBEC-mediated editing of SARS-CoV-2 RNA

Our goal is to design an experimental system to address whether a particular APOBEC enzyme can edit SARS-CoV-2 RNA and the potential effect of APOBEC-mediated RNA editing on the virus. We adapted a cell-based RNA editing system as described in (*26*) to examine the ability of APOBEC proteins to edit SARS-CoV-2 genomic RNA to cause C-to-U mutations. Because A1+A1CF, A3A, and A3G are the three APOBEC proteins shown to possess RNA editing activities (*22-24*), we tested each of these three APOBEC proteins for their ability to edit SARS-CoV-2 RNA. Due to technical and budget limitations for the next generation deep-sequencing (NGS), we selected seven 200 nt-long RNA segments across the SARS-

CoV-2 genome for the APOBEC-editing study **(Fig. 1A)**. These seven segments are distributed from the 5' to 3' end of the genome to include various viral genes, including the 5'-untranslated region (5'UTR)**.** The selected 200 nt viral RNA segments were constructed into DNA reporter vector fused to the C-terminus of eGFP coding sequence that can be transcribed into RNA under a constitutive promoter (**Fig. 1B**). An AAV intron is inserted in the middle of eGFP that can be useful to differentiate the mature mRNA transcript (with the intron spliced out) from the coding DNA containing the intron. A primer annealing to the exon-exon junction on the mature mRNA **(JUNC, Fig. 1B)** can specifically amplify the RNA, but not the coding DNA, by PCR, making it possible to rule out the C-to-U deamination on DNA from the direct C-to-U RNA editing by APOBECs. The reporter vector was co-transfected with the APOBEC editor vector to express the selected APOBEC protein in HEK293T cells **(Fig. 1C)**. Total RNAs were extracted from the cells for cDNA preparation for sequencing using the Safe-Sequencing-System (SSS) as described below.

To minimize the sequencing errors when evaluating the C-to-U RNA editing, we employed the SSS system, a targeted next generation deep-sequencing system, with slightly adapted protocols **(Fig. 1D**, see Methods) (*27, 28*). This SSS method involves the following four critical steps. First, the AccuScript high-fidelity reverse transcriptase (known to have ~10^{-4} - 10^{-5} error rates) was used for the initial reverse transcription of the target SARS-CoV-2 RNA transcripts from the cells to single-stranded cDNA. The JUNC forward primer is used to ensure only mature spliced mRNA segments of SARS-CoV-2 were amplified **(Fig. 1B)**. Second, 2 cycles of initial PCR amplification of the cDNA were performed with a both forward and reverse primers containing a Unique IDentifier (UID), a string of 15 nt randomized sequences, to attach the large family of different UID barcodes (~4^{30}), discerning each original target molecule. Third, the initial 2 cycle-PCR products were purified and then amplified with Illumina adaptors (PCR error rate is ~10^{-7}). Finally, the errors from paired-end Illumina sequencing (PE150, ~10^{-2} -10^{-3} error rates) are minimized by eliminating random mutations in the same UID family (see Methods).

Sequence motifs near the APOBEC-edited C on SARS-CoV-2 RNA

In our SSS system to identify the APOBEC-edited C-to-U mutations, the average number of UID families for each of the 28 experiments (seven segments each with three different APOBEC enzymes and one control) has an average of about 130,000 (minimum 85,000, maximum 187,000) from a total of ~ 484 million (paired) reads **(Table S1)**. The C-to-U editing by all three APOBECs is detected, with A1+A1CF and A3A showing much higher editing than A3G **(Fig S1)**. The C-to-U editing levels by each APOBEC are normalized by the control group **(Table S2)**. Here, we define the significant target C site where the C-to-U editing efficiency is at least 3 times higher than control. Out of 307 total C in the selected viral RNA segments, the number of significant target sites with A1+A1CF is 135, A3A is 67, and A3G is 11 **(Fig. 2A and Table S2)**. Analysis of the sequence contexts around the significant target C sites (with ± 5 nucleotides from target C) showed that A3A prefers for an UCa/u trinucleotide motif and A1+A1CF prefers ACu/a motif **(Fig. 2A)**, which is consistent with the reported motif preference for RNA editing by A3A and A1 **(*23, 29*)**. However, A3G did not show a clear motif preference here, possibly due to generally inefficient editing by A3G on a small number of edited sites (n=11) in this study **(Fig. 2A)**.

We also performed the same analysis of the sequence contexts with the top 30% editing efficiency by APOBECs (or hotspot editing sites), which is translated to 38 times higher than control for A1+A1CF, or 15 times higher for A3A, or 6 times higher for A3G) (**Fig. 2B**). It distinctly shows that A3A strongly prefers the UC motif, whereas A1+A1CF has a strong bias toward the AC motif, and A3G prefers CC motifs on the viral RNA. These results suggest that the observed AC-to-AU mutations are most likely generated by A1 plus cofactor A1CF (or other A1 cofactors, such as RBM47 (***30***)); the UC-to-UU mutations by A3A, and CC-to-CU mutation by A3G in the sequence variation detected on the SARS-CoV-2 RNA segments (**Fig. 2C, D)**. Interestingly, among all the C-to-U variations of the SARS-CoV-2 sequences in our analysis, AC-to-AU (preferred by A1) and UC-to-UU (preferred by A3A) account for 38.23% and 31.83%, respectively, significantly higher than GC-to-GU and CC-to-CU that account for 15.44% and 14.50%, respectively **(Fig S2)**, indicating the significance of APOBEC editing on the viral mutation.

Features of the efficiently APOBEC-edited RNA sites on SARS-CoV-2

Although each of the three APOBEC proteins showed a strong preference for specific dinucleotide sequence motifs (such as AC, UC, or CC sequence motifs) for editing on the viral RNA, the relative editing efficiency of these motif sites vary greatly, such as between 0.0041% to 22.15% for A1+A1CF, and between 0.0040% to 4.46% for A3A **(Table S2)**. Furthermore, many of these AC, UC, and CC motif sites on the viral RNA have no detectable editing (AC: 20.0%, UC: 41.4%, CC: 90.9%) by A1+A1CF, A3A, and A3G, respectively. These results suggest that other RNA features around the preferred sequence motifs, such as the secondary and tertiary structures, must play a role in the editing efficiency of a particular motif site by the respective APOBECs.

The RNA editing by A1+A1CF was previously reported to require a so-called mooring sequence that has a general stem-loop structure around the target C and contains relatively high U/G/A content downstream of the target C (*31, 32*). However, this requirement for the mooring sequence and stem-loop structures are shown to be quite relaxed and still needs further characterization (*26, 33*). We analyzed the RNA features around the top 3 AC sites with the highest editing efficiency by A1+A1CF (**Fig. 3A**). The result showed that they could form a relatively stable stem-loop structure, with relatively high U/G/A contents downstream of the target C **(Fig. 3A)**. Among these top 3 editing sites, editing at C16054 is significantly higher than at C23170 (**Fig. 3A**), suggesting that the editing efficiency also depends on the detailed stem-loop structure and the position of the target C.

Sharma and colleagues recently discovered RNA editing activities by A3A and A3G in human transcripts through RNAseq using NGS technology (*23, 24, 34*). More than half of target cellular RNA substrates have a stem-loop secondary structure, and the target C locates in the loop region. Interestingly, our top 3 highest A3A-mediated editing sites on SARS-CoV-2 RNA reported here all have the U$\underline{\text{C}}$ motifs in the loop of a predicted stem-loop secondary structure **(Fig. 3B)**. Again, the editing efficiency at C16063 (4.5%) is about 3-fold higher than the third efficiently edited site at C23453 (1.6%) **(Fig. 3B)**. Analysis of the top 3 sites edited by A3G also showed the CC (or UC) motifs in the single-stranded loop region of a predicted stem-loop secondary structure (**Fig. 3C**).

To rule out the possibility that the RNA C-to-U mutations result from DNA C-to-U deamination instead of the direct RNA editing by APOBECs, we sequenced the DNA on the reporter vector and its corresponding RNA transcript corresponding to an Orf1b region of SARS-CoV-2 (15,968-16,167 nt). The reporter DNA and the RNA transcript extracted from the cells expressing APOBECs were PCR amplified using the forward prime annealing to the AAV-intron specific for the DNA or to the JUNC specific for the spliced RNA only. The PCR products from the DNA and mRNA were subjected to Sanger sequencing, and the C-to-U changes were analyzed. No DNA C-to-U mutation was detected, but specific RNA C-to-U changes on the mRNA transcript were present in a frequency consistent with our SSS results obtained in the presence of A1+A1CF (e.g., C16049, C16054, and C16092, etc.) or A3A (C16063) **(Fig S3)**. These data indicate that the C-to-U changes on the RNA are not caused by DNA mutation, but are the result of direct RNA editing mediated by A1+A1CF and A3A in our cell-based assay system.

The potential effect of APOBEC-mediated RNA editing on SARS-CoV-2 variants

With the direct experimental evidence that APOBECs can target specific sites on SARS-CoV-2 for editing, we next examined whether APOBEC-mediated RNA editing of the SARS-CoV-2 genome may affect the viral strain variants and fitness. We analyzed the publicly available SARS-CoV-2 genome sequence data (the Nextstrain datasets from Dec. 2019 to Apr. 30th, 2021 downloaded from the UCSC Genome Browser, https://nextstrain.org/ncov/global)(*35*). The analysis revealed that the C-to-U is the predominant mutation for the entire genome, accounting for ~53.8% among all single nucleotide variants (SNVs) within the SARS-CoV-2 5'UTR-Orf1a region (142-341nt, the segment tested in our reporter 1 vector). Of particular interest is the prominent mutation occurring at C203, C222, and C241 in these virus variants **(Fig. 4A and Fig S4A)**, as these 3 sites all feature UC motifs and showed significant C-to-U editing by A3A in our assay **(Fig. 4B, Fig S4B, and Table S2)**. These results suggest that A3A generated these mutations on the viral RNA genome, and the mutations can be maintained, possibly because these mutations generated by A3A editing are not determinantal to the virus. The two C-to-U variants at U**C**203 and U**C**222 were detected in some of the SARS-CoV-2 sequences in late 2020 but are not persistently present in the main circulating strains, suggesting these two mutations may be neutral for the viral fitness. Surprisingly, the C-to-U variant at U**C**241 occurred in early January 2020 and has rapidly become a signature of the dominant strain that spread worldwide **(Fig. 4C and Fig S4C)**, strongly suggesting that this C-to-U mutation at U**C**241 generated by A3A editing may contribute to the better fitness for SARS-CoV-2. Although C241 to U mutation is within 5'UTR, the correlation of this mutation with the dominant new strains is reminiscent of that of the D614G mutation of the spike protein-coding region **(*5, 36*)**. However, because 5'UTR has an important regulatory function for the replication of SARS-CoV-2 RNAs and for the expression of viral proteins **(*37, 38*)**, the U**C**241 mutation may affect one or several aspects of these important functions of the 5'UTR in the viral infection steps.

SARS-CoV-2 replication and progeny yield in cells overexpressing APOBECs

To examine if the three APOBEC proteins can affect SARS-CoV-2 replication and progeny yield in a well-controlled experimental setting, we used the human colon epithelial cell line Caco-2 that expresses ACE2 receptor and thus is suitable for studying SARS-CoV-2 replication **(*39*)**. Because Caco-2 cell lines have no detectible endogenous expression of A1, A3A, and A3G, we first constructed Caco-2 stable cell lines expressing one of the three APOBEC genes. The externally inserted APOBECs are under tetracycline-controlled promoter so that their expression can be induced by doxycycline **(Fig. 5A, Fig S5)**. The Caco-2-APOBEC stable cell lines were then infected by SARS-CoV-2, and the viral RNA replication and progeny yield were measured and compared with the control cell line without APOBEC expression. The viral RNA abundance as an indicator for RNA replication was measured using real-time quantitative PCR to detect the RNA levels using primers specific for amplifying three viral regions: the Nsp12 region, the S region, and the N region. The viral progeny yield was assayed through plaque assay using the virions produced from the Caco-2-APOBEC stable cell lines at different time points post-infection **(Fig. 5A)**.

The replication assay results showed that, compared with the control Caco-2 cell line, the abundance of viral RNAs has increased significantly with A3A overexpression at 72 hours and 96 hours post SARS-CoV-2 infection. In contrast, no significant change of viral RNA level was detected with the overexpression of A1+A1CF or A3G even up to 96 hours post-infection **(Fig. 5B)**. These results suggest that, despite the general viral restriction function of APOBECs, the presence of A1+A1CF and A3G and their RNA editing activity do not restrict viral replication. Furthermore, A3A-mediated editing even appears to endow an advantage for the viral RNA replication, which is consistent with the observation that the UC241 mutation caused by A3A at the 5'UTR region in the widely circulating SARS-CoV-2 strains after January 2020 (**Fig. 4A-C)**.

Interestingly, the increased viral RNA replication with A3A overexpression also correlates with higher viral progeny yield. While no difference in viral titer was observed at 48 hours from all cell lines with or without APOBEC expression, the virus titer from the cell line expressing A3A showed an approximately 100-fold higher than the control and A1+A1CF and A3G expressing cell lines at 72 hours **(Fig. 5C)**. This result suggests that A3A expressing promotes viral replication and progeny production. These results collectively provided direct experimental evidence that SARS-CoV-2 may take advantage of A3A-mediated mutational forces for their fitness and propagation.

DISCUSSION

In this ongoing unprecedented COVID-19 pandemic, several new viral strains have emerged due to viral mutation. More viral strains are expected to evolve in the future due to continuous virus mutations and the added selection pressure from the clinical usage of vaccines and antiviral drugs. Prior bioinformatic analysis of the SARS-CoV-2 sequences suggested that some of the C-to-U mutations may be caused by APOBEC proteins instead of the random mutations by viral RNA polymerase during replication or by ROS (*9, 10, 12, 25, 40, 41*). Here we described the first experimental evidence demonstrating APOBEC enzymes, A1+A1CF, A3A, and A3G, can target specific SARS-CoV-2 viral sequences for RNA editing, and the resulting mutations likely contribute to the viral replication and fitness.

Using an APOBEC-RNA editing assay in a cell-based system (*26*), we examined the C-to-U editing on selected SARS-CoV-2 RNA segments in HEK293T cells by A1+A1CF, A3A, and A3G. The results demonstrated that the editing efficiency at specific viral RNA sites could reach as high as 22.2% for A1+A1CF, 4.5% for A3A, and 0.18% for A3G, respectively, on the transcribed reporter viral RNA populations (Fig. 3A-C). This editing efficiency at a specific site is very high compared with the estimated ~10^{-6} – 10^{-7} random incorporation error rate for Coronaviruses replication (*42*). A3A showed a preferred dinucleotide U$\underline{C}$ motif, consistent with previous reports about the preferred motif for A3A editing (*23, 43*).

Furthermore, sequence analysis of viral strains also revealed a high mutation rate of C in AC motif in SARS-CoV-2 and Rubella Virus, but with no mutational driver assigned for such mutations (*3, 9, 25*). The data here demonstrate that A1, together with its cofactor A1CF, can edit specific AC motif sties of the viral RNA with high efficiency **(Fig. 2-3)**.

However, the UC and AC motifs alone are not sufficient for efficient editing by A3A or A1+A1CF. Many of the UC and AC motifs in the SARS-CoV-2 RNA showed only the background level of C-to-U mutation. Additional RNA structural features around the UC/AC motifs should play an important role in dictating whether a UC or AC site can be efficiently edited **(Fig. 3)**. Prior reports show that some cellular RNA targets with stem-loop structures are favored by A3A and A1 editing (*23, 26, 32, 43*). However, the exact RNA secondary/tertiary structural features that can dictate the editing efficiency of a UC/AC site in the SARS-CoV-2 genomic RNA by A3A/A1 remains to be addressed.

While the editing of SARS-CoV-2 RNA by A3A, A1+A1CF, and A3G has been demonstrated in our cell culture system, the analysis of the expression profiles reveals that A3A and A1+A1CF, but not A3G, are expressed in the human organs and cell types infected by SARS-CoV-2 **(Fig S6A, B)**. Such expression profiles make it possible for A3A and A1+A1CF to edit the viral RNA genome in the real world. Many human cell types expressing ACE2 in multiple organs can be infected by SARS-CoV-2, including (but not limited to) the lungs, heart, small intestine, and liver (*44, 45*). A3A expresses in lung epithelial cells, and, importantly, the A3A expression level is significantly stimulated by SARS-CoV-2 infection in patients **(*46-48*) (Fig S6A, B)**. A1 and its two known cofactors, A1CF and RBM47, are not expressed in the lungs but are expressed in the small intestine or liver (*49*) that can also be infected by SARS-CoV-2 **(Fig S6B)**.

A1 was not considered previously as a candidate which could edit the SARS-CoV-2 genome because of its target specificity and lack of expression in lungs, the primary target tissues of SARS-CoV-2 infection (*50*). In our experimental system, the editing efficiency of A1+A1CF on the AC motif is much higher than A3A on the UC motif **(Fig. 2C, D)**. Interestingly, analysis of SARS-CoV-2 variants from the database also showed higher AC motif mutations (38.3%) than UC motif mutations (31.2%) **(Fig S2)**. These results indicate that many of the AC-to-AU mutations in the SARS-CoV-2 genome from patients can be driven by A1+cofactor-mediated RNA editing in the small intestine and liver when SARS-CoV-2 infects them. Interestingly, A1CF and another A1-cofactor, RBM47, were shown to physically interact with SARS-CoV-2 RNA in an interactome study (*51*), suggesting that these RNA-binding cofactors may recruit A1 to edit SARS-CoV-2 RNA in the infected tissue.

Since APOBEC proteins play an important role in immune responses against DNA and RNA viral pathogens APOBECs ((*17-20*) and references therein), we also investigated the possible effects of the three

APOBECs on SARS-CoV-2 replication and progeny yield in our experimental system **(Fig. 5A)**. Even though no apparent differences in SARS-CoV-2 replication and progeny production were observed in the control and the A1+A1CF or A3G-expressing stable cell lines, significant increases of the viral RNA replication and the viral yield were observed in the A3A-expressing cell line **(Fig. 5B and 5C)**. These results contrast to the known antiviral effect of APOBECs (*17-20*). Currently, it is unclear exactly how the A3A-overexpression can promote the replication and progeny production of SARS-CoV-2. One plausible explanation is that, while most APOBEC-mediated C-to-U mutations may be detrimental to the virus and are lost in the viral infection cycle, some mutations beneficial to the virus's fitness will be selected for in the new viral strains. If so, SARS-CoV-2 can then turn the tables on the APOBEC mutational defense system for its evolution, including (but not limited to) the improvement of viral RNA replication, protein expression, evasion of host immune responses, and receptor binding and cell entry.

Even though the mechanisms by which A3A-editing promotes SARS-CoV-2 replication and progeny production may be complex and require future investigation, we have detected an example that could offer insights into how SARS-CoV-2 takes advantage of A3A-mediated mutations. Among many of the detected A3A edited sites on SARS-CoV-2 RNA in our study are three C-to-U mutations located in the 5'UTR region, UC203, UC222, UC241 (**Fig. 4A, B**). The SARS-CoV-2 from patients detected all of these three mutations at different stages since early 2020, but only major circulating viral strains since early 2020 acquired the mutation at UC241 (**Fig. 4C, Fig S4C)**, suggesting that C-to-U mutation at U**C**241 is selected for better viral fitness. This observation is surprising considering UC241 is in the non-coding 5'UTR region and, thus, not affecting the ORF coding of a protein, as in the case of the previously reported D614G mutation of the spike protein for viral fitness (*5, 36*). Therefore, this C241 to U mutation should not be related to a change of a viral protein function, such as cell surface receptor binding or polyprotein processing. Mutations outside protein-coding ORFs of SARS-CoV-2 were previously designated as non-functional changes (*3*). However, the 5'UTR region of SARS-CoV-2 has important function in regulating protein expression and viral RNA replication and virion packaging (*37, 38, 52*). Thus, this UC241 mutation in the 5'UTR region may impact the viral RNA stability, replication, packaging, or the translation efficiency of the downstream proteins to endow the virus with better fitness.

In summary, we report here the experimental evidence demonstrating that SARS-CoV-2 genomic RNA can be directly edited with high efficiency by A3A and A1+A1CF to cause C-to-U mutation. Critical factors dictating the RNA-editing sites include an UC for A3A or AC motif for A1 as well as certain RNA structural features around the target C. Even though some APOBECs, including A3A, are regarded as host antiviral factors, we show here that RNA editing of SARS-CoV-2 by A3A somehow can promote viral replication/propagation. These results suggest that SARS-CoV-2 can take advantage of APOBEC-mediated mutation for their fitness and evolution. Unlike the unpredictability of the random mutations caused by RNA replication or ROS, this study has significant implications in predicting potential locations of mutations in the SARS-CoV-2 genome based on the UC/AC motifs and the surrounding RNA structures. Such prediction of possible mutations in the coding and non-coding regions is especially meaningful with the new selection pressure from vaccines and antiviral drugs. The risk of novel immune escape or drug-resistant SARS-CoV-2 strains increases with the continued large-scale circulation of the delta and other variants among unvaccinated and vaccinated people. Thus, we envision that this study will provide a basis for future prediction of the potential mutations on the virus by APOBECs and other host factors for guiding future vaccine design and antiviral therapies against the immune escape and drug-resistant mutants.

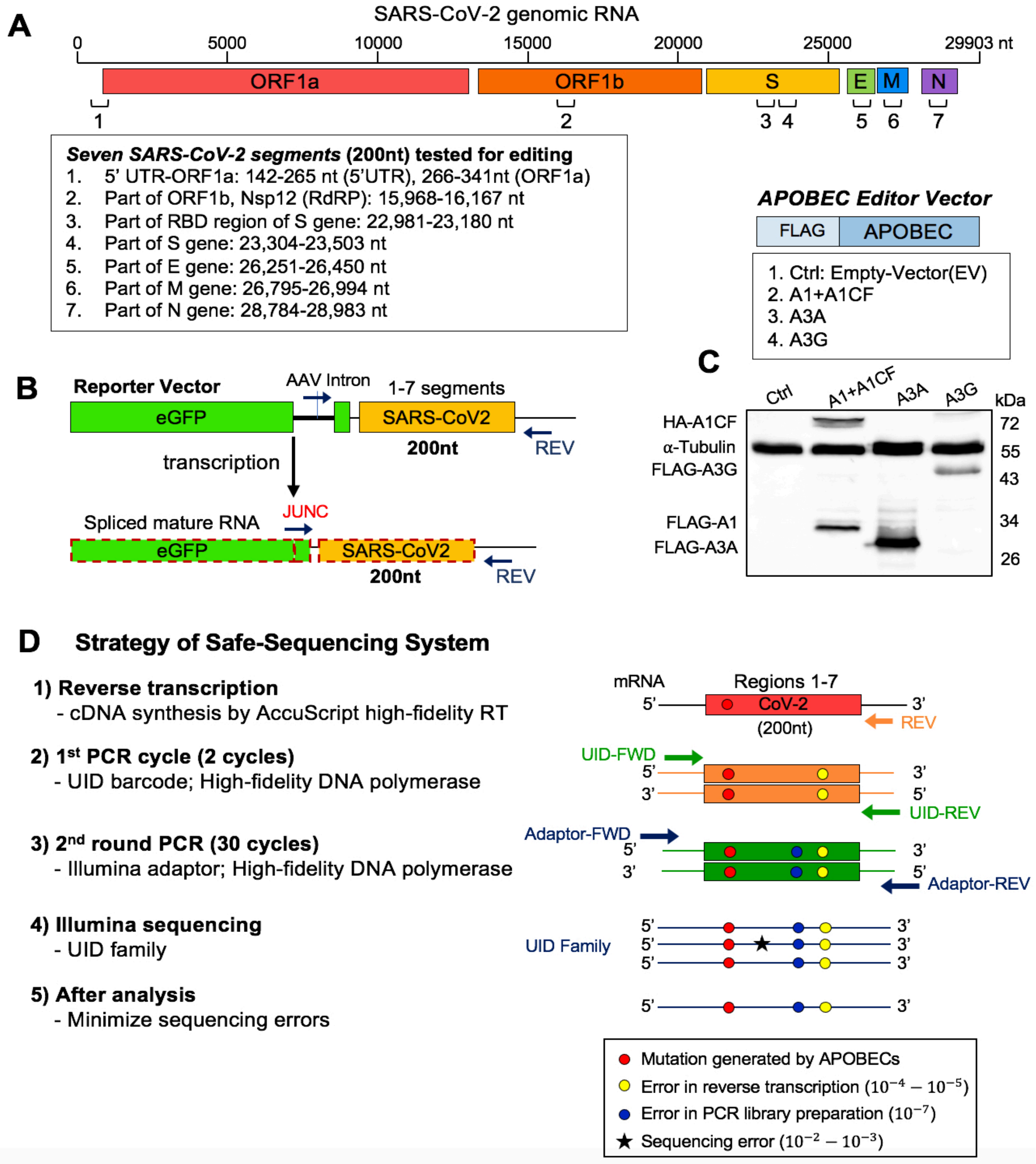

A
SARS-CoV-2 genomic RNA
0
5000
10000
15000
20000
25000
29903 nt
ORF1a
ORF1b
S
E
M
N
1
2
3 4
5 6
7
Seven SARS-CoV-2 segments (200nt) tested for editing
1. 5' UTR-ORF1a: 142-265 nt (5'UTR), 266-341nt (ORF1a)
2. Part of ORF1b, Nsp12 (RdRP): 15,968-16,167 nt
3. Part of RBD region of S gene: 22,981-23,180 nt
4. Part of S gene: 23,304-23,503 nt
5. Part of E gene: 26,251-26,450 nt
6. Part of M gene: 26,795-26,994 nt
7. Part of N gene: 28,784-28,983 nt
APOBEC Editor Vector
FLAG
APOBEC
1. Ctrl: Empty-Vector(EV)
2. A1+A1CF
3. A3A
4. A3G
B
Reporter Vector
AAV Intron
1-7 segments
eGFP
SARS-CoV2
200nt
REV
transcription
JUNC
Spliced mature RNA
eGFP
SARS-CoV2
200nt
REV
C
Ctrl
A1+A1CF
A3A
A3G
kDa
HA-A1CF
α-Tubulin
FLAG-A3G
FLAG-A1
FLAG-A3A
72
55
43
34
26
D
Strategy of Safe-Sequencing System
1) Reverse transcription
- cDNA synthesis by AccuScript high-fidelity RT
2) 1st PCR cycle (2 cycles)
- UID barcode; High-fidelity DNA polymerase
3) 2nd round PCR (30 cycles)
- Illumina adaptor; High-fidelity DNA polymerase
4) Illumina sequencing
- UID family
5) After analysis
- Minimize sequencing errors
mRNA
Regions 1-7
5'
CoV-2
3'
(200nt)
REV
UID-FWD
UID-REV
Adaptor-FWD
Adaptor-REV
UID Family
5'
3'
● Mutation generated by APOBECs
● Error in reverse transcription ($10^{-4} - 10^{-5}$)
● Error in PCR library preparation (10^{-7})
★ Sequencing error ($10^{-2} - 10^{-3}$)

Fig. 1. Experimental design of APOBEC-mediated editing of SARS-CoV-2 RNA. (A) Diagram of the entire SARS-CoV-2 genomic RNA, showing the nucleotide (nt) positions (box) of the seven RNA segments (1-7) selected for studying the RNA editing by APOBECs. **(B)** Reporter vector that contain each of the seven selected viral RNA segments that can be transcribed into an mRNA containing an AAV intron between the eGFP and the viral RNA segment. The AAV intron will be spliced out in a mature mRNA containing a junction sequence (JUNC) that differs from its coding DNA, which can be used to selectively amplify either the mature mRNA or the coding DNA by using a primer anneal to the JUNC or to the AAV intron. **(C)** Three APOBEC editor vectors (top, A1+A1CF, A3A, and A3G) and the Western blot showing their expression in 293T cells (bottom). **(D)** Strategy of the Safe-Sequencing-System (SSS) used in this study to minimize random errors from PCR amplification and sequencing. After the SARS-CoV-2 RNAs from cell extracts are reverse transcribed, the cDNAs are sequentially amplified by the UID barcode (2 cycles) and the Illumina adapter (30 cycles). Sequencing analysis will identify the APOBEC-mediated C-to-U editing within the same UID family (See Methods).

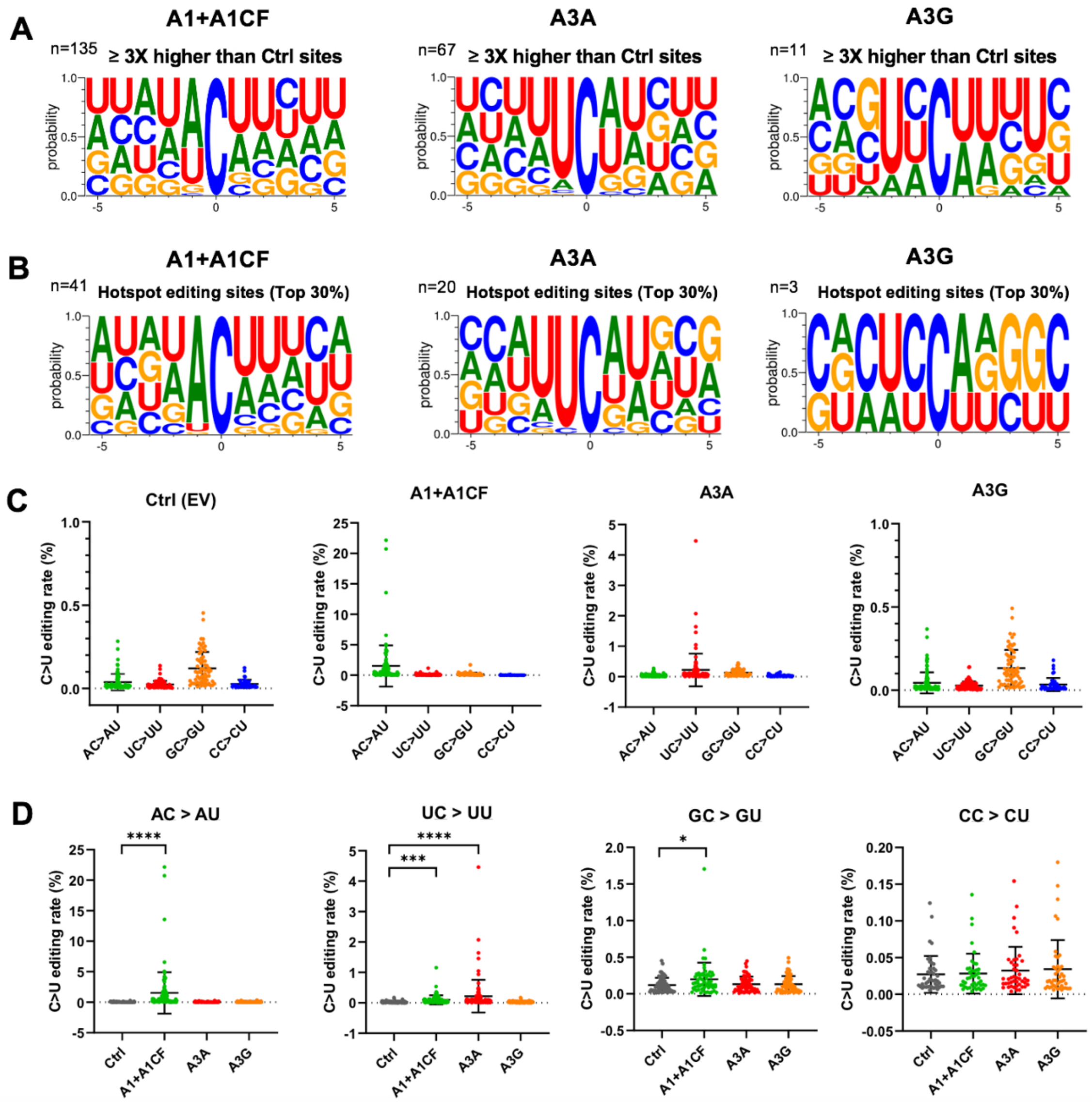
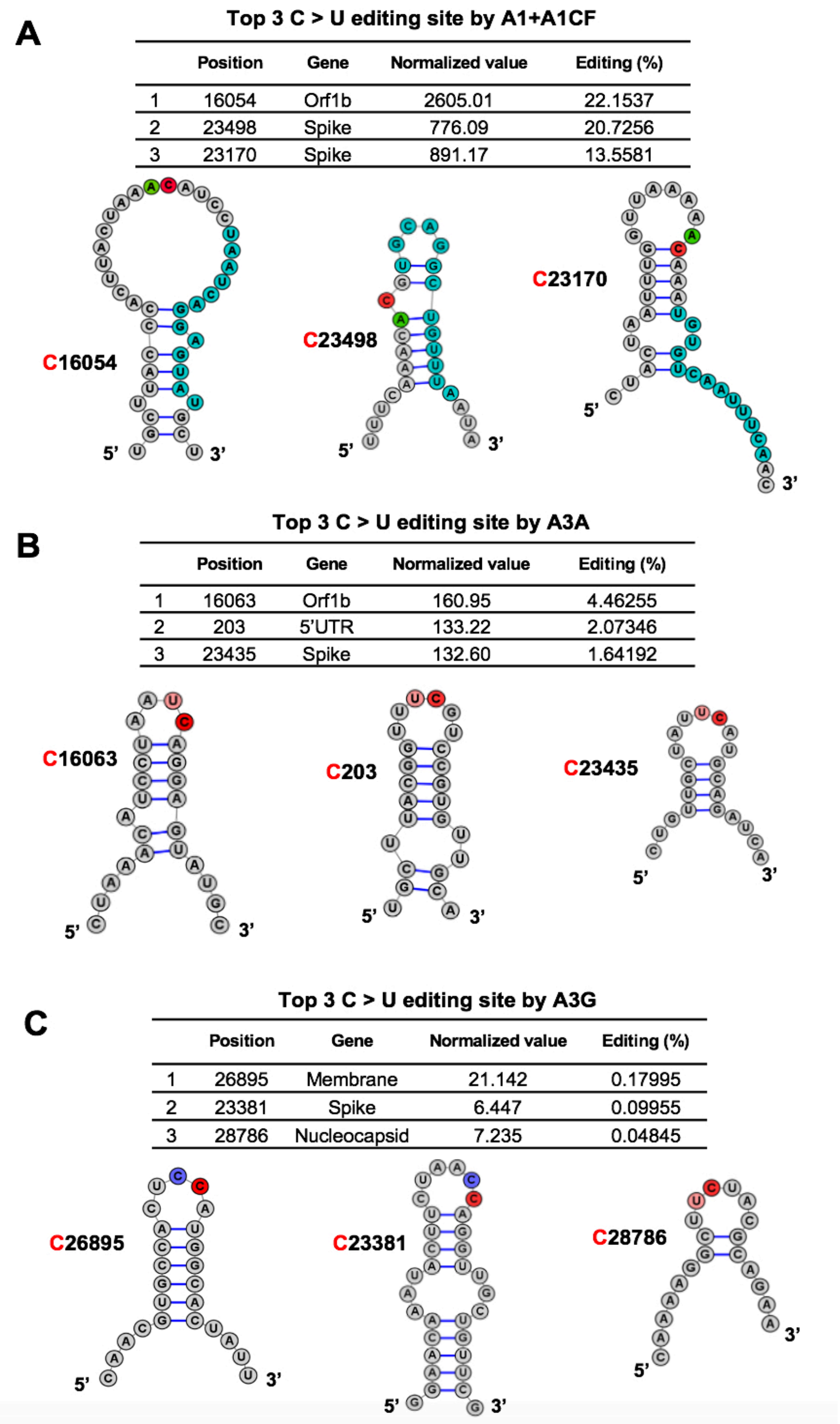


Fig. 2. Local sequence context at the APOBEC-edited C sites on SARS-CoV-2 RNA. (A) Local sequences around the significantly edited target C sites (± 5 nucleotides from target C at position 0) by A1+A1CF, A3A, or A3G. The editing level of each C site was normalized to the Ctrl, and only sites with 3x or higher editing levels than the normalized value were defined as significant editing sites. **(B)** Analysis of local sequences around the top 30% edited C sites (or hotspot editing sites) among the significantly edited sites, showing predominantly AC motif for A1+A1CF, UC for A3A, and CC for A3G. **(C-D)** Comparison of the C-to-U editing rates (%) of different dinucleotide motifs by a particular APOBEC RNA editor (panel-C) and C-to-U editing rates (%) of a particular dinucleotide motif by the three APOBEC RNA editors (panel-D). Each dot represents the C-to-U editing level obtained from the SSS results. In panel-D, statistical significance was calculated by unpaired two-tailed student's t-test with *P*-values represented as: P > 0.05 = not significant; not indicated, * = P < 0.05, *** = P < 0.001, **** = P < 0.0001.



Top 3 C > U editing site by A1+A1CF

	Position	Gene	Normalized value	Editing (%)
1	16054	Orf1b	2605.01	22.1537
2	23498	Spike	776.09	20.7256
3	23170	Spike	891.17	13.5581

Top 3 C > U editing site by A3A

	Position	Gene	Normalized value	Editing (%)
1	16063	Orf1b	160.95	4.46255
2	203	5'UTR	133.22	2.07346
3	23435	Spike	132.60	1.64192

Top 3 C > U editing site by A3G

	Position	Gene	Normalized value	Editing (%)
1	26895	Membrane	21.142	0.17995
2	23381	Spike	6.447	0.09955
3	28786	Nucleocapsid	7.235	0.04845

Fig. 3. Overall features of the RNA around the most preferred APOBEC-edited sites on SARS-CoV-2. The predicted RNA secondary structures (*53*) of the sequences near the target C for the top 3 highest editing sites by A1+A1CF **(A)**, the top 3 highest editing sites by A3A **(B)**, and the top 3 highest editing sites by A3G **(C)**. (See related Table S2). The editing efficiency of each site is listed at the top of each panel. In the secondary structure, the target C sites are highlighted in red, and -1 positions of the target C sites are highlighted in green for A, pink for U, and blue for C, respectively. In panel-A, the proposed canonical mooring sequences (highlighted in sky blue) contain relatively high U/A/G contents downstream of the target C.

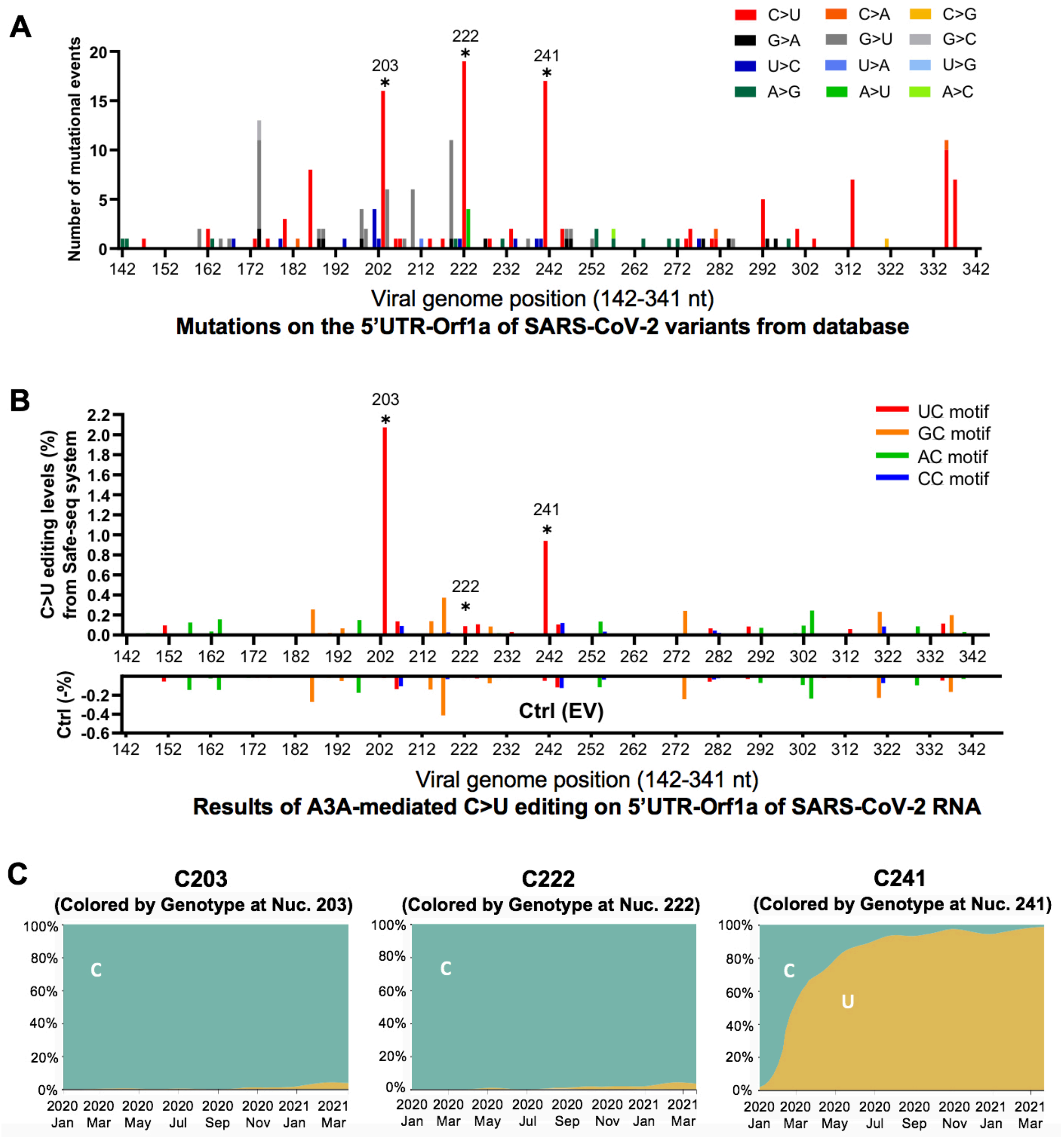


Fig. 4. The potential effect of APOBEC-mediated editing on SARS-CoV-2 mutations and fitness. (A) The number of mutational events (all single nucleotide variants) on SARS-CoV-2 RNA segment 5'UTR-Orf1a (142-341nt), see related Fig. 1A) from the publicly available SARS-CoV-2 genome sequence data (the Nextstrain datasets from Dec. 2019 to Apr. 30th, 2021 downloaded from the UCSC Genome Browser, https://nextstrain.org/ncov/global) and https://nextstrain.org/ncov/global). The C203, C222, and C241 represent many of the C-to-U mutational events (asterisks) with the U<u>C</u> motif in the SARS-CoV-2 variants. **(B)** A3A-mediated C-to-U editing results on the 5'UTR-Orf1a (142-341nt) of SARS-CoV-2 obtained from our cell-based editing system and the SSS analysis. The Ctrl (EV) editing levels (or background error rates) of the corresponding region are presented as negative values (%). Four different dinucleotide motifs are indicated, with U<u>C</u> in red, G<u>C</u> in orange, A<u>C</u> in green, and C<u>C</u> in blue. The C203, C222, and C241 (asterisks) all showed significant editing by A3A. **(C)** The C-to-U mutation prevalence over time at C203, C222, and C241. The sequencing frequency is represented by C in blue and U in yellow (referred to the Nextstrain datasets: https://nextstrain.org/ncov/global). This analysis showed that SARS-CoV-2 started to acquire the C-to-U mutation at C241 in January 2020. By July 2020, 90% of the circulating viral variants carry this mutation at C241. By March 2021, almost all circulating viral variants have this mutation, suggesting the C241 to U mutation in the 5'UTR is beneficial to the viral fitness.

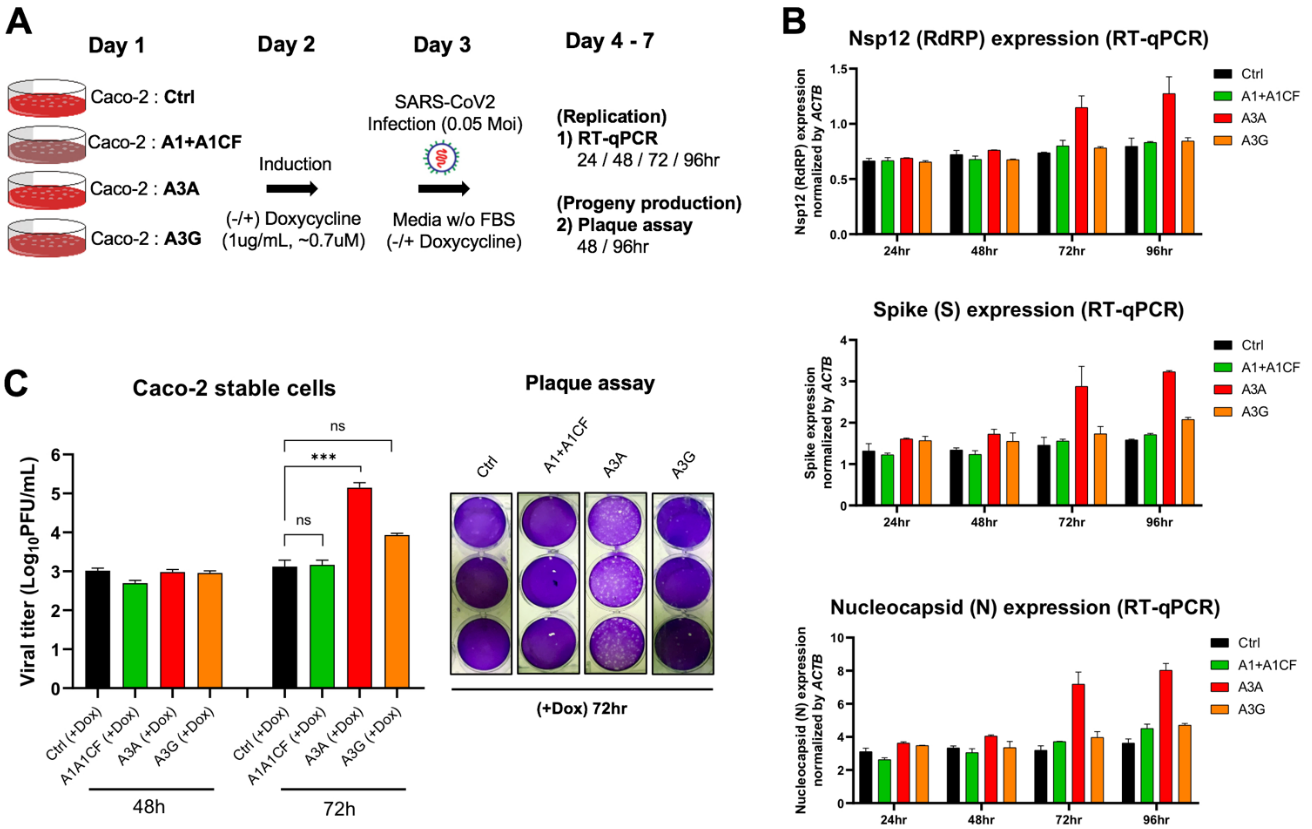


Fig. 5. SARS-CoV-2 replication in cells overexpressing APOBECs. (A) Overview of experiments for SARS-CoV-2 replication in the presence of APOBECs. The Caco-2 stable cell lines were constructed to express A1+A1CF, A3A, or A3G under a tetracycline-controlled promoter through doxycycline induction. The Caco-2-APOBEC stable cell lines were then infected with SARS-CoV-2 (MOI = 0.05), and the viral RNA replication and progeny production were measured at different time points post-infection. **(B)** Effect of each APOBEC overexpression on SARS-CoV-2 viral RNA replication. Measurement of relative viral RNA abundance at different time points after viral infection of Caco-2-APOBEC stable cell lines with the expression of A1+A1CF, A3A, or A3G. The viral RNA abundance was measured using real-time quantitative PCR (qPCR) to detect RNA levels by using specific primers to amplify three separate viral regions, the *Nsp12*, *S,* or *N* coding regions (see Methods). **(C)** Effect of each APOBEC overexpression on SARS-CoV-2 progeny production. Infectious viral progeny yield harvested in the medium at 48 hrs and 72 hrs post-infection was determined by plaque assay (see Methods). Statistical significance was calculated by unpaired two-tailed student's t-test with *P*-values represented as: P > 0.05 = not significant, *** = P < 0.001.

REFERENCES

1. B. Hu, H. Guo, P. Zhou, Z. L. Shi, Characteristics of SARS-CoV-2 and COVID-19. *Nat Rev Microbiol*, (2020).
2. N. Zhu *et al.*, A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N Engl J Med* **382**, 727-733 (2020).
3. Z. W. Kockler, D. A. Gordenin, From RNA World to SARS-CoV-2: The Edited Story of RNA Viral Evolution. *Cells* **10**, (2021).
4. F. Robson *et al.*, Coronavirus RNA Proofreading: Molecular Basis and Therapeutic Targeting. *Mol Cell* **80**, 1136-1138 (2020).
5. B. Korber *et al.*, Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell* **182**, 812-827 e819 (2020).
6. R. N. Tasakis *et al.*, SARS-CoV-2 variant evolution in the United States: High accumulation of viral mutations over time likely through serial Founder Events and mutational bursts. *PLoS One* **16**, e0255169 (2021).
7. W. F. Garcia-Beltran *et al.*, Multiple SARS-CoV-2 variants escape neutralization by vaccine-induced humoral immunity. *Cell* **184**, 2523 (2021).
8. W. T. Harvey *et al.*, SARS-CoV-2 variants, spike mutations and immune escape. *Nat Rev Microbiol* **19**, 409-424 (2021).
9. S. Di Giorgio, F. Martignano, M. G. Torcia, G. Mattiuz, S. G. Conticello, Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2. *Sci Adv* **6**, eabb5813 (2020).
10. P. Simmonds, 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* **5**, (2020).
11. L. Perelygina *et al.*, Infectious vaccine-derived rubella viruses emerge, persist, and evolve in cutaneous granulomas of children with primary immunodeficiencies. *PLoS Pathog* **15**, e1008080 (2019).
12. R. Matyasek, A. Kovarik, Mutation Patterns of Human SARS-CoV-2 and Bat RaTG13 Coronavirus Genomes Are Strongly Biased Towards C>U Transitions, Indicating Rapid Evolution in Their Hosts. *Genes (Basel)* **11**, (2020).
13. T. Mourier *et al.*, Host-directed editing of the SARS-CoV-2 genome. *Biochem Biophys Res Commun* **538**, 35-39 (2021).
14. A. Graudenzi, D. Maspero, F. Angaroni, R. Piazza, D. Ramazzotti, Mutational signatures and heterogeneous host response revealed via large-scale characterization of SARS-CoV-2 genomic diversity. *iScience* **24**, 102116 (2021).
15. S. R. Gonzales-van Horn, P. Sarnow, Making the Mark: The Role of Adenosine Modifications in the Life Cycle of RNA Viruses. *Cell Host Microbe* **21**, 661-669 (2017).
16. C. X. George, Z. Gan, Y. Liu, C. E. Samuel, Adenosine deaminases acting on RNA, RNA editing, and interferon action. *J Interferon Cytokine Res* **31**, 99-117 (2011).
17. M. E. Olson, R. S. Harris, D. A. Harki, APOBEC Enzymes as Targets for Virus and Cancer Therapy. *Cell Chem Biol* **25**, 36-49 (2018).
18. J. D. Salter, R. P. Bennett, H. C. Smith, The APOBEC Protein Family: United by Structure, Divergent in Function. *Trends Biochem Sci* **41**, 578-594 (2016).
19. R. S. Harris, J. P. Dudley, APOBECs and virus restriction. *Virology* **479-480**, 131-145 (2015).
20. X. S. Chen, Insights into the Structures and Multimeric Status of APOBEC Proteins Involved in Viral Restriction and Other Cellular Functions. *Viruses* **13**, (2021).
21. D. M. Driscoll, J. K. Wynne, S. C. Wallis, J. Scott, An in vitro system for the editing of apolipoprotein B mRNA. *Cell* **58**, 519-525 (1989).
22. B. Teng, C. F. Burant, N. O. Davidson, Molecular cloning of an apolipoprotein B messenger RNA editing protein. *Science* **260**, 1816-1819 (1993).
23. S. Sharma *et al.*, APOBEC3A cytidine deaminase induces RNA editing in monocytes and macrophages. *Nat Commun* **6**, 6881 (2015).
24. S. Sharma, S. K. Patnaik, R. T. Taggart, B. E. Baysal, The double-domain cytidine deaminase APOBEC3G is a cellular site-specific RNA editing enzyme. *Sci Rep* **6**, 39100 (2016).

25. L. J. Klimczak, T. A. Randall, N. Saini, J. L. Li, D. A. Gordenin, Similarity between mutation spectra in hypermutated genomes of rubella virus and in SARS-CoV-2 genomes accumulated during the COVID-19 pandemic. *PLoS One* **15**, e0237689 (2020).
26. A. D. Wolfe, D. B. Arnold, X. S. Chen, Comparison of RNA Editing Activity of APOBEC1-A1CF and APOBEC1-RBM47 Complexes Reconstituted in HEK293T Cells. *J Mol Biol* **431**, 1506-1517 (2019).
27. I. Kinde, J. Wu, N. Papadopoulos, K. W. Kinzler, B. Vogelstein, Detection and quantification of rare mutations with massively parallel sequencing. *Proc Natl Acad Sci U S A* **108**, 9530-9535 (2011).
28. J. Eboreime, S. K. Choi, S. R. Yoon, N. Arnheim, P. Calabrese, Estimating Exceptionally Rare Germline and Somatic Mutation Frequencies via Next Generation Sequencing. *PLoS One* **11**, e0158340 (2016).
29. B. R. Rosenberg, C. E. Hamilton, M. M. Mwangi, S. Dewell, F. N. Papavasiliou, Transcriptome-wide sequencing reveals numerous APOBEC1 mRNA-editing targets in transcript 3' UTRs. *Nat Struct Mol Biol* **18**, 230-236 (2011).
30. N. Fossat *et al.*, C to U RNA editing mediated by APOBEC1 requires RNA-binding protein RBM47. *EMBO Rep* **15**, 903-910 (2014).
31. M. Sowden, J. K. Hamm, H. C. Smith, Overexpression of APOBEC-1 results in mooring sequence-dependent promiscuous RNA editing. *J Biol Chem* **271**, 3011-3017 (1996).
32. C. Maris, J. Masse, A. Chester, N. Navaratnam, F. H. Allain, NMR structure of the apoB mRNA stem-loop and its interaction with the C to U editing APOBEC1 complementary factor. *RNA* **11**, 173-186 (2005).
33. A. D. Wolfe, S. Li, C. Goedderz, X. S. Chen, The structure of APOBEC1 and insights into its RNA and DNA substrate selectivity. *NAR Cancer* **2**, zcaa027 (2020).
34. S. Sharma, B. E. Baysal, Stem-loop structure preference for site-specific RNA editing by APOBEC3A and APOBEC3G. *PeerJ* **5**, e4136 (2017).
35. J. Hadfield *et al.*, Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121-4123 (2018).
36. J. A. Plante *et al.*, Spike mutation D614G alters SARS-CoV-2 fitness. *Nature* **592**, 116-121 (2021).
37. Z. Miao, A. Tidu, G. Eriani, F. Martin, Secondary structure of the SARS-CoV-2 5'-UTR. *RNA Biol* **18**, 447-456 (2021).
38. J. Nomburg, M. Meyerson, J. A. DeCaprio, Pervasive generation of non-canonical subgenomic RNAs by SARS-CoV-2. *Genome Med* **12**, 108 (2020).
39. H. Chu *et al.*, Comparative tropism, replication kinetics, and cell damage profiling of SARS-CoV-2 and SARS-CoV with implications for clinical manifestations, transmissibility, and laboratory studies of COVID-19: an observational study. *Lancet Microbe* **1**, e14-e23 (2020).
40. M. Sadykov, T. Mourier, Q. Guan, A. Pain, Short sequence motif dynamics in the SARS-CoV-2 genome suggest a role for cytosine deamination in CpG reduction. *J Mol Cell Biol*, (2021).
41. R. Wang, Y. Hozumi, Y. H. Zheng, C. Yin, G. W. Wei, Host Immune Response Driving SARS-CoV-2 Evolution. *Viruses* **12**, (2020).
42. E. C. Smith, N. R. Sexton, M. R. Denison, Thinking Outside the Triangle: Replication Fidelity of the Largest RNA Viruses. *Annu Rev Virol* **1**, 111-132 (2014).
43. R. Buisson *et al.*, Passenger hotspot mutations in cancer driven by APOBEC3A and mesoscale genomic features. *Science* **364**, (2019).
44. Y. Y. Zheng, Y. T. Ma, J. Y. Zhang, X. Xie, COVID-19 and the cardiovascular system. *Nat Rev Cardiol* **17**, 259-260 (2020).
45. W. Trypsteen, J. Van Cleemput, W. V. Snippenberg, S. Gerlo, L. Vandekerckhove, On the whereabouts of SARS-CoV-2 in the human body: A systematic review. *PLoS Pathog* **16**, e1009037 (2020).
46. T. S. Heng, M. W. Painter, C. Immunological Genome Project, The Immunological Genome Project: networks of gene expression in immune cells. *Nat Immunol* **9**, 1091-1094 (2008).
47. D. Blanco-Melo *et al.*, Imbalanced Host Response to SARS-CoV-2 Drives Development of COVID-19. *Cell* **181**, 1036-1045 e1039 (2020).
48. Y. Xiong *et al.*, Transcriptomic characteristics of bronchoalveolar lavage fluid and peripheral blood mononuclear cells in COVID-19 patients. *Emerg Microbes Infect* **9**, 761-770 (2020).
49. V. Blanc *et al.*, Apobec1 complementation factor (A1CF) and RBM47 interact in tissue-specific regulation of C to U RNA editing in mouse intestine and liver. *RNA* **25**, 70-81 (2019).

50. J. Ratcliff, P. Simmonds, Potential APOBEC-mediated RNA editing of the genomes of SARS-CoV-2 and other coronaviruses and its impact on their longer term evolution. *Virology* **556**, 62-72 (2021).
51. N. Schmidt *et al.*, The SARS-CoV-2 RNA-protein interactome in infected human cells. *Nat Microbiol* **6**, 339-353 (2021).
52. B. J. Guan, Y. P. Su, H. Y. Wu, D. A. Brian, Genetic evidence of a long-range RNA-RNA interaction between the genomic 5' untranslated region and the nonstructural protein 1 coding region in murine and bovine coronaviruses. *J Virol* **86**, 4631-4643 (2012).
53. J. S. Reuter, D. H. Mathews, RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics* **11**, 129 (2010).
54. Y. Rao *et al.*, Targeting CTP Synthetase 1 to Restore Interferon Induction and Impede Nucleotide Synthesis in SARS-CoV-2 Infection. *bioRxiv*, (2021).
55. Y. Wei, J. R. Silke, P. Aris, X. Xia, Coronavirus genomes carry the signatures of their habitats. *PLoS One* **15**, e0244025 (2020).

Acknowledgments: This work is supported in part by NIH AI150524 to X.S.C, and the startup fund to P.F. from Herman Ostrow School of Dentistry at the University of Southern California.

Authors contributions: K. Kim, P. Calabrese and X.S. Chen designed the experiments; K. Kim, S. Wang, P. Calabrese, C. Qin, Y. Rao. performed the experiments and data analysis; K. Kim wrote the manuscript draft; P. Calabrese, S. Wang, P. Feng, commented the manuscript; and X.S. Chen revised the manuscript.

Competing interests: The authors have no competing interests.

Data and materials availability: All data is available in the manuscript or the supplementary materials. All materials used in the study are available to any researchers upon request.

Supplementary Materials
Materials and Methods
Fig S1-S6
References (26-28, 35, 39, 48, 54-55)

Table S1
Table S2
Table S3

Supplementary Materials

APOBEC-mediated Editing of SARS-CoV-2 Genomic RNA Impacts Viral Replication and Fitness

MATERIALS AND METHODS

The cell-based RNA editing system

The Cell-based RNA editing system is adapted from previously reported in reference (*26*). Briefly, reporter vectors containing DNA corresponding to the different RNA segments of SARS-CoV-2 (NC_045512.2) (see Fig. 1A, 1B) and the APOBEC (A1+A1CF, A3A, and A3G) editor vectors (see Fig. 1C) were constructed. A1+A1CF is constructed as one open reading frame (ORF) with a translational terminator sequence T2A inserted between A1 and A1CF (A1-T2A-A1CF), which will produce individual A1 and A1CF proteins in a 1:1 ratio. HEK293T cells were cultured in DMEM medium supplemented with 10% FBS, streptomycin (100 µg/mL), and penicillin (100U/mL) and maintained at 37 °C, 5% CO_2. One day before transfection, the cells (250 µL) were seeded at an approximate concentration of 250,000 cells/mL on an 8-well glass chamber (CellVis). The cells were then transfected with a mixture (25 µL) of an APOBEC editor vector (500 ng) and a SARS-CoV-2 reporter vector (50 ng) and 1.5 µL of X-tremeGENE 9 transfection reagent (Sigma) and incubated for 48 hrs. After harvesting the cells, RNA extraction with Trizol (Thermo Fisher) and DNA extraction with QuickExtract (EpiCentre) was performed, respectively, according to the manufacturer's recommended instructions. The editor and reporter vectors used in this study were listed in Table S3.

Sequencing library preparation

The extracted RNA was reverse transcribed with Accuscript High-Fidelity Reverse Transcriptase (Agilent) to produce the single-stranded cDNA using a specific primer annealing to the downstream sequence of SARS-CoV-2 reporter segments. The reaction was performed in a volume of 20µl containing 1µg of total RNA, 100 µM of reverse primer, 1X Accuscript buffer, 10 mM dNTP, 0.1M DTT, 8U RNase Inhibitor, and 1µl of Accuscript High-Fidelity Reverse Transcriptase (Agilent) for 1 hr at 42 °C. The cDNA was then amplified for 2 cycles by adding a forward primer annealing to the junction region (JUNC, Fig. 1B), where the AAV intron is spliced out. In this first 2-cycle PCR amplification, the forward and reverse primers were attached to barcodes consists of 15 randomized nucleotides as the Unique Identifier (UID), plus four tri-nucleotides designating four different experimental conditions: TGA for A1+A1CF; CAT for A3A; GTC for A3G; and ACG for Ctrl. Phusion® High-Fidelity DNA Polymerase (NEB) was used for this PCR reaction: 98 °C 5 min - (98 °C 30 sec, 71.4 °C 30 sec, 72 °C 1 min) x2 – 72 °C 5 min. This PCR product (330 bp) was then cleaned up using a spin column PCR cleanup kit (Thermo) to remove the free first-round barcode primers. The second-round PCR was performed for 30 cycles with Illumina flowcell adaptor primers using Phusion® High-Fidelity DNA Polymerase (NEB): 98 °C 5 min - (98 °C 30 sec, 72 °C 1 min) x30 – 72 °C 5 min. All 28 (4 editors x 7 different SARS-CoV-2 substrates) of the different pooled PCR products (399 bp) were combined in equal amounts for the final libraries. The final libraries were subjected to a full HiSeq Lane (PE150, 370M paired reads, Novogene). The primers for the sequencing library preparation were listed in Table S3.

Analysis of Safe-Sequencing-System

To distinguish a true mutation from random mutation during PCR and sequencing errors, we followed the approach as reported in (*27*). The details of our implementation of the method was described in

(*28*). We wrote Python scripts to analyze the sequencing data. We only considered those sequencing reads such that (1) at least 85% of the bases matched the reference sequence, and (2) the quality scores for all the UID bases were 30 or greater (probability of a sequencing error < 0.001). We clustered reads with the same UID and barcode into UID families. We only considered those families with at least three reads with the same UID and barcode. At each nucleotide site, the mutation frequency is calculated by dividing a numerator by a denominator. The denominator is the number of UID families that, at this particular nucleotide site, have at least three reads with quality scores of at least 20 (probability of a sequencing error < 0.01; because of this quality restriction, the denominator may be different at different sites). The numerator is the number of UID families that, at this particular site, (1) have at least three reads with quality scores of at least 20, and (2) 95% of these reads have the same base, which is different than the reference. The probability that three out of three reads will all have the same sequencing error at a site is then 10^{-7} (=$(0.01^3)/(3^2)$).

Caco-2 Stable cell line expressing APOBEC proteins

We used lentiviral transfection to construct stable Caco-2 cell lines expressing A3A, A3G, and A1+A1CF to study the effect of APOBEC on SARS-CoV-2 replication because Caco-2 expresses the virus receptor ACE2 (***39***). Lentivirus was produced by lentiviral vector system pLVX-TetOne-Puro (Clon-tech) in HEK293T cells. The cells (about 2 x 10^6 cells) were seeded in a 100 mm plate one day before transfection. The cells were then co-transfected with lentiviral packaging vectors, 1.0 µg of pdR8.91 (Gag-Pol-Tat- Rev, Addgene), 0.5 µg of pMD2.G (VSV-G, Addgene), and 1.7 µg of the pLVX-TetOne-Puro vector encoding the APOBEC proteins, using 20 µL of X-tremeGENE 9 transfection reagent (Sigma). Lentivirus-containing supernatant from infected HEK293T cells was collected after 70 hrs and filtered through a 0.45 µm PVDF filter (Millipore). Virions were precipitated with NaCl (0.3 M final) and PEG-6000 (8.5% final) at 4°C for 6 hrs and centrifugated at 4000 rpm at 4°C for 30 min. The pelleted virions were resuspended in 100 µL of MEM medium. Caco-2 cells (human colon epithelial cell line, ATCC) were cultured in MEM medium supplemented with 10% FBS, streptomycin (100µg/mL), and penicillin (100U/mL), and maintained at 37°C, 5% CO_2. The Caco-2 stable cell lines were generated by transducing with the lentivirus for 24 hrs and selected with 5 µg/ml of puromycin. The expression of A1+A1CF, A3A, or A3G was induced by adding 1µg/mL doxycycline for 24 - 96 hrs. Expression of these APOBEC proteins was verified by Western blot.

SARS-CoV-2 virus replication and progeny production

SARS-CoV-2 propagation, infection, and viral titration were performed as previously described (*54*). All SARS-CoV-2 related experiments were performed in the biosafety level 3 (BSL-3) facility (USC). For SARS-CoV-2 propagation, Vero E6-hACE2 cells were used. The cells were plated at 1.5 x 10^6 cells in a T25 flask for 12 hr and infected with SARS-CoV-2 (isolate USA-WA1/2020) at MOI 0.005 in an FBS-free DMEM medium. Virus-containing supernatant was collected when virus-induced cytopathic effect (CPE) reached approximately 80%.

To assess the effect of APOBEC (A1+A1CF, A3A, and A3G) on SARS-CoV-2 RNA replication, the Caco-2-APOBEC stable cells (about 2 x 10^5 cells) were plated in 12-well plates. After 15 hours, cells were treated or untreated with Doxycycline for 24 hours before infection. Before viral infection, the cells were washed with an FBS-free medium once. Viral infection was incubated on a rocker for 45 min at 37 °C. The cells were washed and incubated in a medium containing 10% FBS with or without Doxycycline. Total cellular RNA was extracted from the infected cells at 24, 48, 72, 96 hrs. Real-time quantitative PCR (qPCR) was used to quantify the viral RNA abundance level at the four different time points using viral RNA-specific primers to detect the Nsp12, S, and N regions. The qPCR of the internal actin RNA abundance level is used as a control by using actin-specific primers.

To assess the effect of APOBEC (A1+A1CF, A3A, and A3G) on SARS-CoV-2 viral progeny production, plaque assay on Vero E6-hACD2 cells was used. Vero E6-hACE2 cells were seeded in 12-well plates. Once cell reached confluence, cells were infected with serially diluted SARS-CoV-2 virions collected from the infected Caco-2-APOBEC stable cells that express A1+A1CF, A3A, or A3G at 48 hrs and 72 hrs after viral infection. The medium was removed after infection, and overlay medium containing FBS-free 1 x DMEM and 1% low-melting-point agarose was added. At 48 and 72 h post-infection, cells were fixed with 4% paraformaldehyde (PFA) overnight and stained with 0.2% crystal violet. Plaques were counted on a lightbox.

Quantitative real-time PCR

Total RNA was extracted from the SARS-CoV-2 infected Caco-2 cells using Trizol (Thermo Fisher). The extracted RNA was then reverse transcribed with the reverse primers specific to Nsp12, S, and N coding regions of SARS-CoV-2, and β-Actin as an internal control, respectively, using the high-fidelity reverse transcriptase Protoscript II (NEB). The reaction was performed in a volume of 20 µl containing 1µg of total RNA, 100 µM reverse primer, 1X Protoscript II buffer, 10 mM dNTP, 0.1M DTT, 8U RNase Inhibitor (40U/µl), and 200U ProtoScript RT for 1 hr at 42 ℃. Quantitative real-time PCR was then performed with SYBR Green (PowerUp™ SYBR™ Green Master Mix, Thermo Fisher Scientific) in a volume of 10 µl/well containing 1µl of reverse transcribed cDNA product from above, 0.25 µl of forward and reverse primers (10 µM), and 5 µl of PowerUp™ SYBR™ Green Master Mix (2X) using a CFX Connected Real-Time PCR machine (Bio-Rad). Primers used in this study were listed in Table S3. The indicated gene (*Nsp12*, *S*, *N*) expression levels were calculated by the 2-ΔΔCt method and normalized by β-Actin expression level.

Western blot and antibodies

For Western blot analysis, cells were lysed in 1x RIPA buffer (Sigma). Western blot analysis were performed from three independent transfections using FLAG-tagged APOBECs and HA-tagged A1CF. α-Tubulin: internal loading control. The lysates were then subjected to Western blot with anti-FLAG M2 mAb (F3165, Sigma, 1:3,000), anti-HA mAb (HA.C5, Abcam, 1:3,000), and anti- α-tubulin mAb from mouse (GT114, GeneTex, 1:5,000) as primary antibodies. Cy3-labelled goat-anti-mouse mAb (PA43009, GE Healthcare, 1:3,000) was subsequently used as a secondary antibody. Cy3 signals were detected and visualized using Typhoon RGB Biomolecular Imager (GE Healthcare).

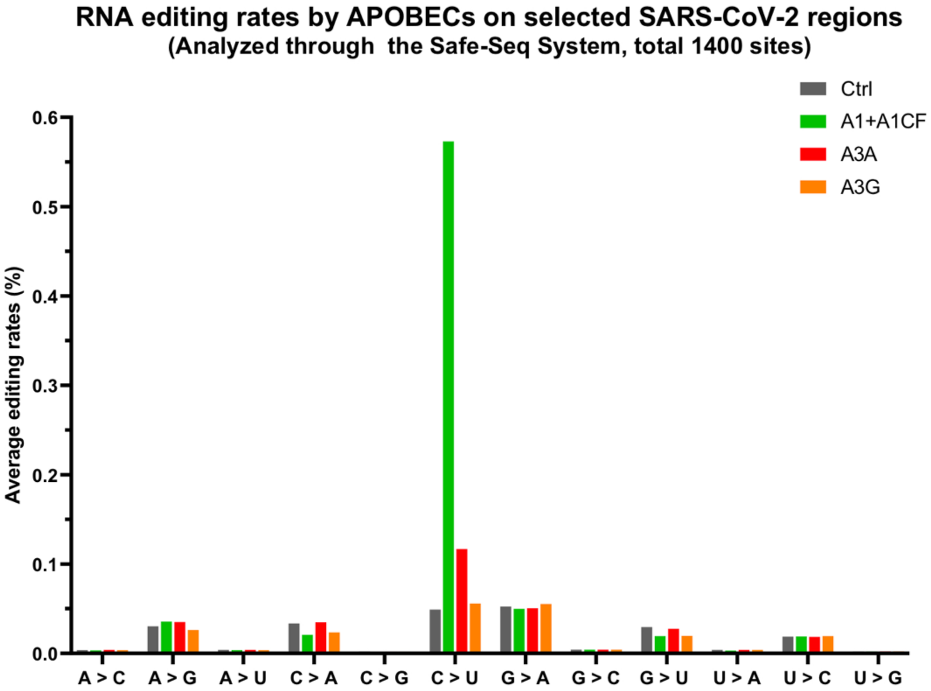
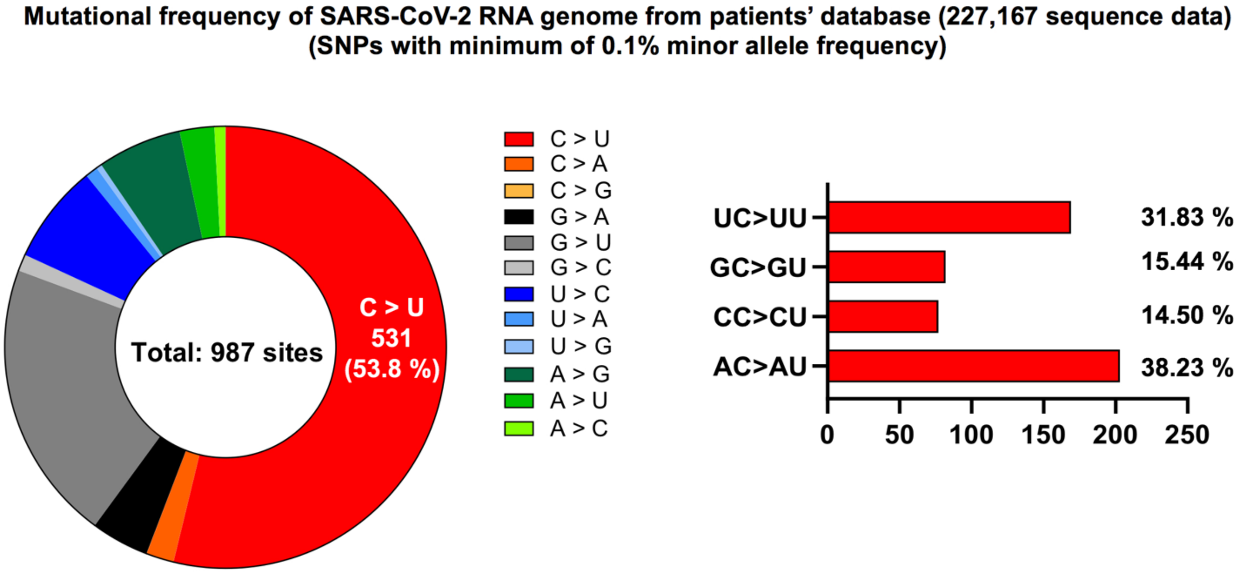


Fig. S1. RNA editing rates by APOBECs on the selected SARS-CoV-2 segments. Average rates (%) of all single nucleotide variations were analyzed through the Safe-Sequencing-System (SSS). See related Supplementary Table 1.



Mutational frequency of SARS-CoV-2 RNA genome from patients' database (227,167 sequence data)
(SNPs with minimum of 0.1% minor allele frequency)

Fig. S2. Single nucleotide variations (SNPs) of SARS-CoV-2 genome from patients' database. A total of 987 SNPs with minor allele frequencies > 0.1 % were counted from a total of 227,167 SARS-CoV-2 sequences on the UCSC genome browser (https://genome.ucsc.edu/covid19.html). The C-to-U mutation is the most common type with 53.8 % (top), of which the ratio according to the dinucleotide motifs including -1 position upstream of the mutated C are shown in the bottom chart.

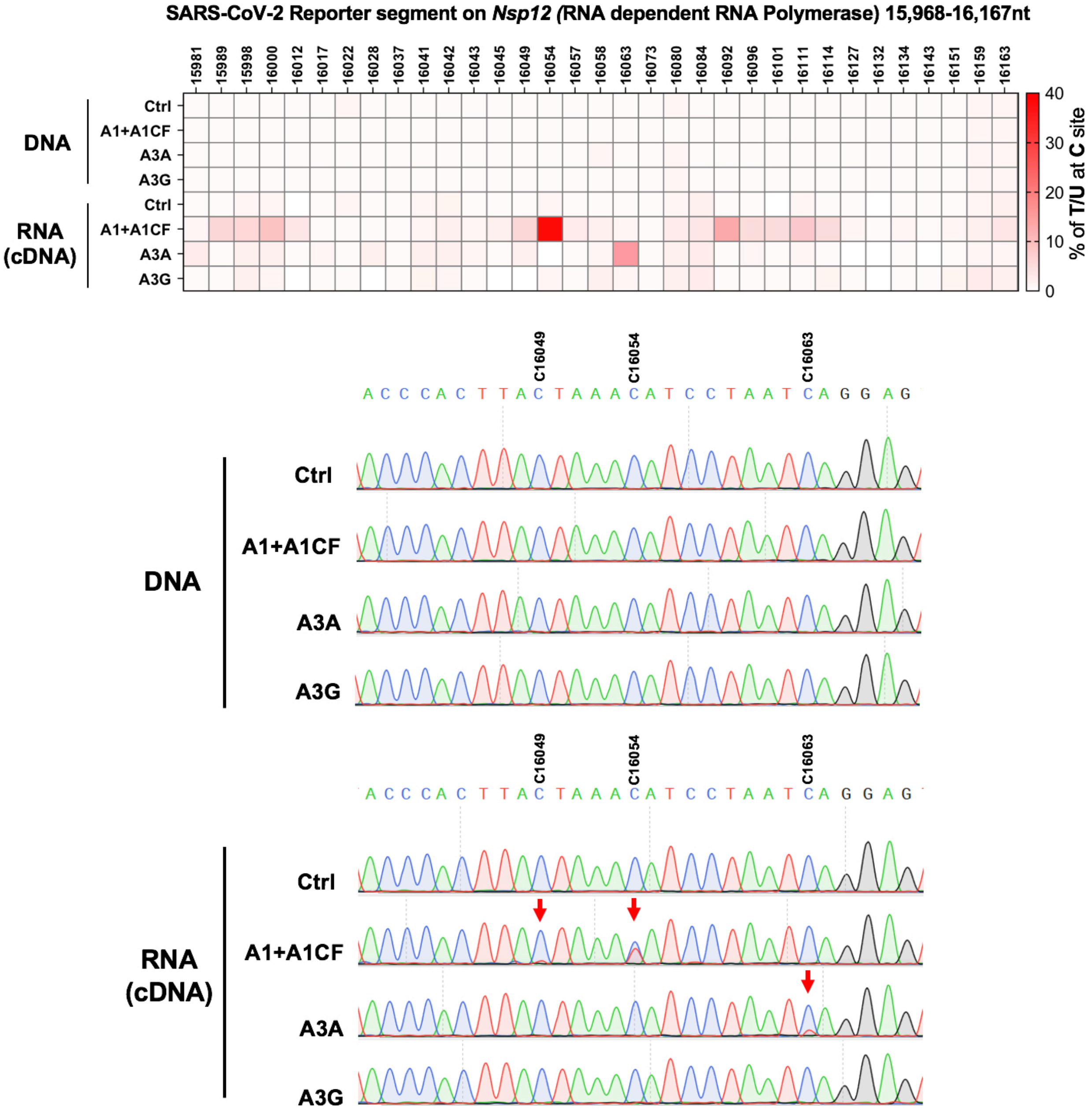


Fig. S3. Verification of C-to-U mutation caused by the APOBEC-mediated DNA deamination or direct RNA editing on a SARS-CoV-2 reporter segment on *Nsp12* (15968-16167 nt). The temperature-bar chart (top panel) shows the DNA and RNA C-to-T/U editing levels (%), which are based on the Sanger sequencing results of the DNA (middle panel) and the cDNA (RNA) (bottom panel). All C sites in this SARS-CoV-2 segment are marked with the virus nt sequence numbers on the top bar chart. Some of the RNA editing sites are indicated by red arrows (C16049, C16054, and C16063).

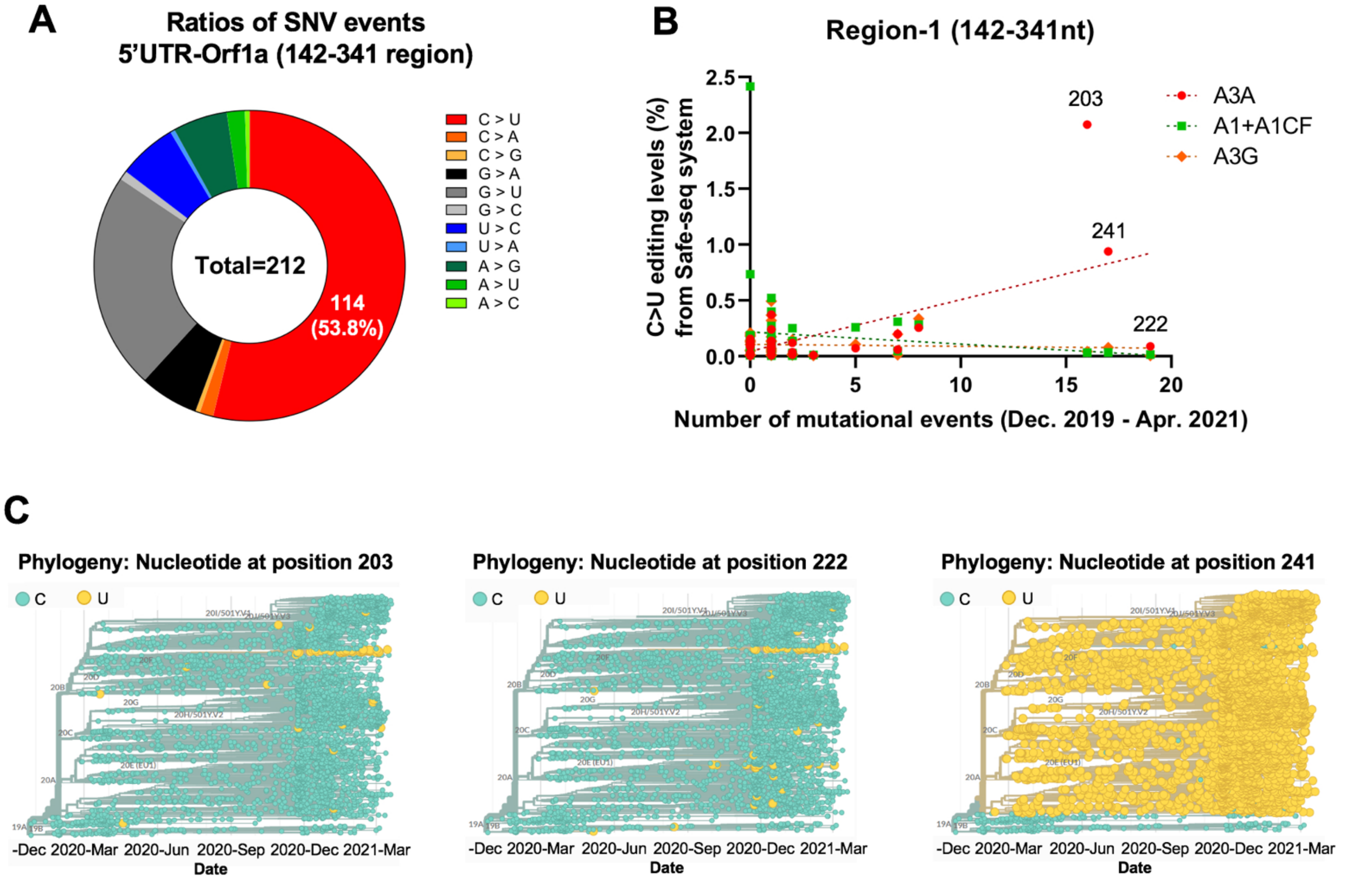


Fig S4. Comparison of SARS-CoV-2 variants and the APOBEC-mediated RNA editing sites on the viral 5'UTR-Orf1a segment (142-341). (A) Ratios of all SNVs events on the 5'UTR-Orf1a segment from the sequence database (referred to the Nextstrain datasets (*35*)). : https://nextstrain.org/ncov/global) (B) Correlation of C-to-U RNA editing levels by the three APOBECs identified by our Safe-sequencing system (Y-axis) and the mutational events of SARS-CoV-2 from the sequence database between Dec. 2019 to April. 2021 (X-axis). Dotted lines indicate linear regressions with 95% confidence, and the case of A3A shows a positive correlation, and A1+A1CF shows a negative correlation. (C) Phylogenetic trees for C-to-U variant at C203, C222, and C241 (referred to the Nextstrain datasets (*35*)). : https://nextstrain.org/ncov/global). These phylogenetic trees correlate well with the C-to-U mutation prevalence over time at C203, C222, and C241, as shown in Fig. 4C.

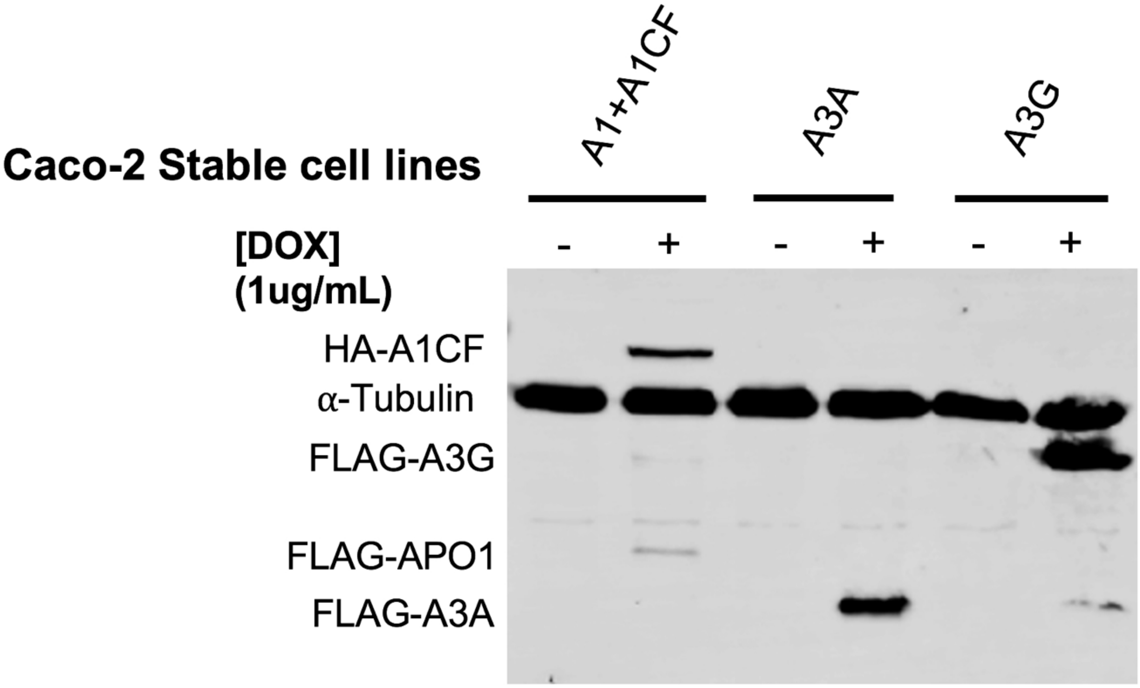


Fig. S5. Western blot analysis of APOBEC protein expression from Caco-2-APOBEC stable cell lines. N-terminal FLAG-tagged A1, A3A, and A3G, and HA-tagged A1CF were detected under doxycycline treatment (1μg/mL). α-Tubulin is the internal loading control.

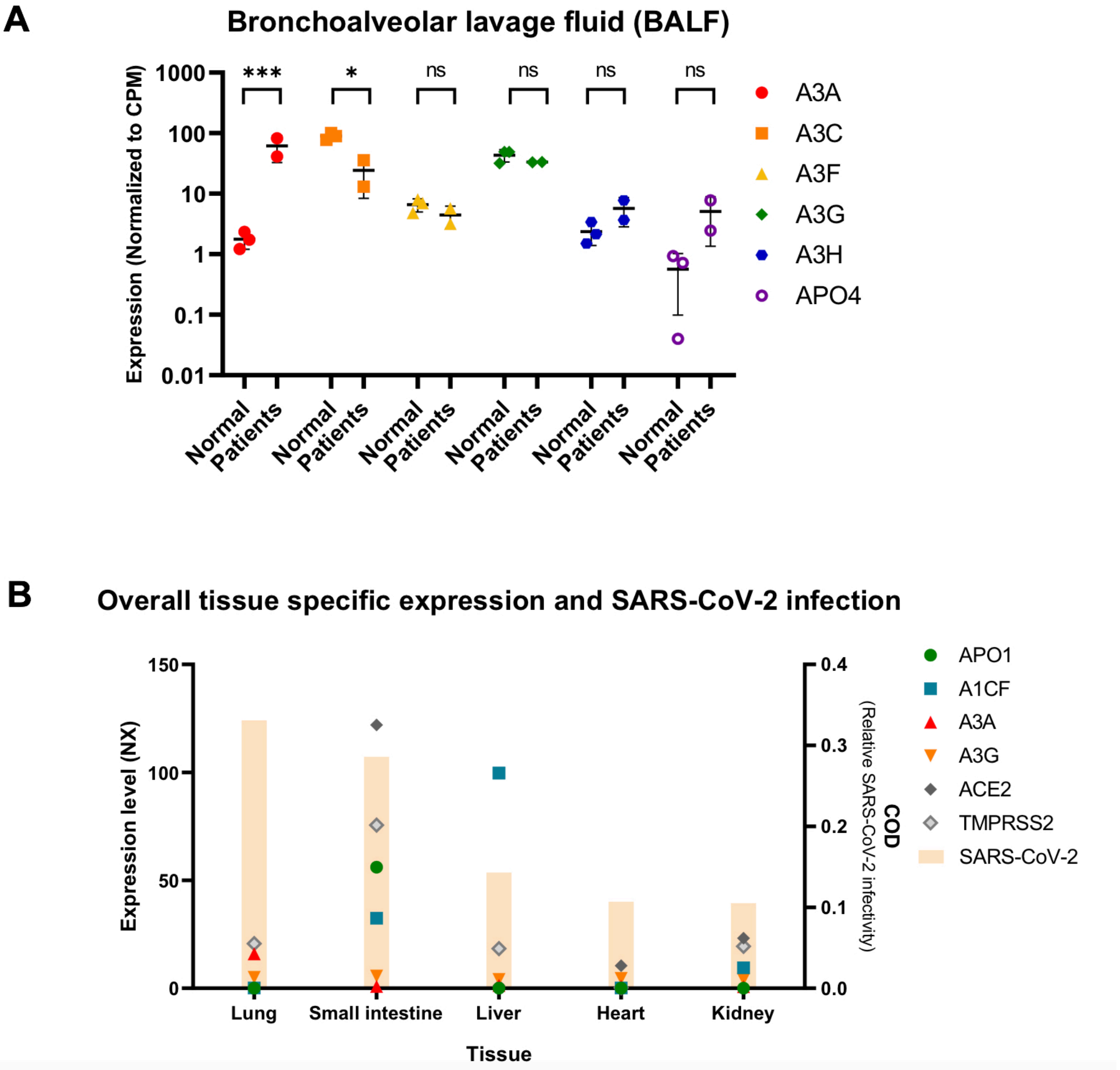


Fig S6. Relations between SARS-CoV-2 infection and APOBEC expression. (A) Data analysis of the expression level of six APOBECs in healthy people and COVID-19 infected patients in Bronchoalveolar lavage fluid (BALF) samples (referred to the RNAseq data from reference (*48*)). **(B)** Overall gene expressions of the three APOBECs (A1, A3A, A3G) and A1CF in the tissues that can be infected by SARS-CoV-2. The commonness of viral detection (COD, relative SARS-CoV-2 infectivity) score for each tissue is indicated by yellow shaded boxes (referred to the COD score based on reference (*55*)).). Each of gene expression values (NX) was retrieved from the human protein atlas (http://www.proteinatlas.org)