**Title:** Cytosine deamination in SARS-CoV-2 leads to progressive CpG depletion.

**Authors**: Mukhtar Sadykov[1*], Tobias Mourier[1*], Qingtian Guan[1], Arnab Pain[1,2,3**]
**Affiliation**:

[1]King Abdullah University of Science and Technology (KAUST), Pathogen Genomics Laboratory, Biological and Environmental Science and Engineering (BESE), Thuwal-Jeddah, 23955-6900, Saudi Arabia;
[2]Research Center for Zoonosis Control, Global Institution for Collaborative Research and Education (GI-CoRE); Hokkaido University, N20 W10 Kita-ku, Sapporo, 001-0020 Japan;
[3]Nuffield Division of Clinical Laboratory Sciences (NDCLS), University of Oxford, Headington, Oxford, OX3 9DU, United Kingdom

*These authors contributed equally
****Correspondence:** Arnab Pain Email: arnab.pain@kaust.edu.sa. King Abdullah University of Science and Technology, Jeddah, Saudi Arabia. Phone: (+966) 54 470 0687

**Abbreviations**: C>U stands for cytosine to uracil substitution, the same applies to other nucleotide substitutions; APOBEC - Apolipoprotein B Editing Complex; ZAP - zinc antiviral protein.

## Abstract

RNA viruses use CpG reduction to evade the host cell defense, but the driving mechanism is still largely unknown. To address this, we used a rapidly growing genomic dataset of SARS-CoV-2 with relevant metadata information. SARS-CoV-2 genomes show a progressive increase of C-to-U substitutions resulting in CpG loss over just a few months. This is consistent with APOBEC-mediated RNA editing resulting in CpG reduction, thus allowing the virus to escape ZAP-mediated RNA degradation. Our results thus link the dynamics of target sequences in the viral genome for two known host molecular defense mechanisms, mediated by the APOBEC and ZAP proteins.

## Introduction

Viruses utilize numerous mechanisms to avoid the host cell defense. One such mechanism is the CpG dinucleotide reduction observed in many single-stranded RNA viruses[1]. Low CpG abundance is typically a consequence of two factors: constraints of amino-acid conservation and selection pressure from diverse host cell defense mechanisms targeting CpG[2].

Prior to the involvement of the host immune system, the infected cells have to rely on their own antiviral restriction factors. These factors are mostly proteins induced by interferon signaling cascade[3]. The antiviral proteins directly involved in the restriction of viral replication and transcription are zinc antiviral protein (ZAP) and Apolipoprotein B Editing Complex (APOBEC)[4]. The defensive role of ZAP on HIV-1 manifestation in the host cell has been extensively studied[5]. ZAP selectively binds to viral CpG regions and recruits RNA exosome complex resulting in viral RNA degradation[6,7]. Recently, ZAP was shown to restrict SARS-CoV-2 production upon the induction by interferons[8].

The APOBEC protein family is another example of host antiviral enzymes. Seven human APOBEC proteins are known for catalyzing cytosine to uracil deamination in foreign single-stranded DNA (ssDNA) and RNA (ssRNA)[7,9–11]. Enzymatic target motifs for most of the APOBEC enzymes have been experimentally identified, among which the most common were 5'-[T/U]C-3' and 5'-CC-3' for DNA/RNA substrates[9–13]. Some members of APOBEC proteins inhibit replication of human coronavirus, HCoV-NL63, but their role in hypermutation of the viral genome is not clear[14]. However, it was recently suggested that the SARS-CoV-2 undergoes genome editing by host-dependent RNA-editing proteins such as APOBEC (C>U) and ADAR (A>G)[15-16]. mRNA transcriptional response data from COVID-19 patients demonstrated that APOBEC3A (A3A) belongs to one of the most expressed genes in response to SARS-CoV-2, and the highest among APOBEC family[17].

SARS-CoV-2 is an enveloped positive-sense ssRNA virus, belonging to *Coronaviridae* family, which is characterized by large genomes (~30K nucleotide long), high recombination frequencies and genomic proofreading mechanism[18]. SARS-CoV-2 is believed to have a zoonotic origin and show high genetic similarity to bat-related coronaviruses[19]. The COVID-19 pandemic brought intensive scientific attention to the virus, and thousands of viral genomic sequences with metadata became publicly available[20,21]. Given a large amount of available data and the relatively low mutation rate of the virus[22], we aimed to monitor its genomic evolution on a very brief time scale during the COVID-19 pandemic. We show progressive C>U substitutions in SARS-CoV-2 genome within the timeframe of fewer than five months. We highlight the role of C>U substitutions in the reduction of CpG motifs and hypothesize that this progressive decrease in CpG motifs is driven by host APOBEC enzymes.

## Main

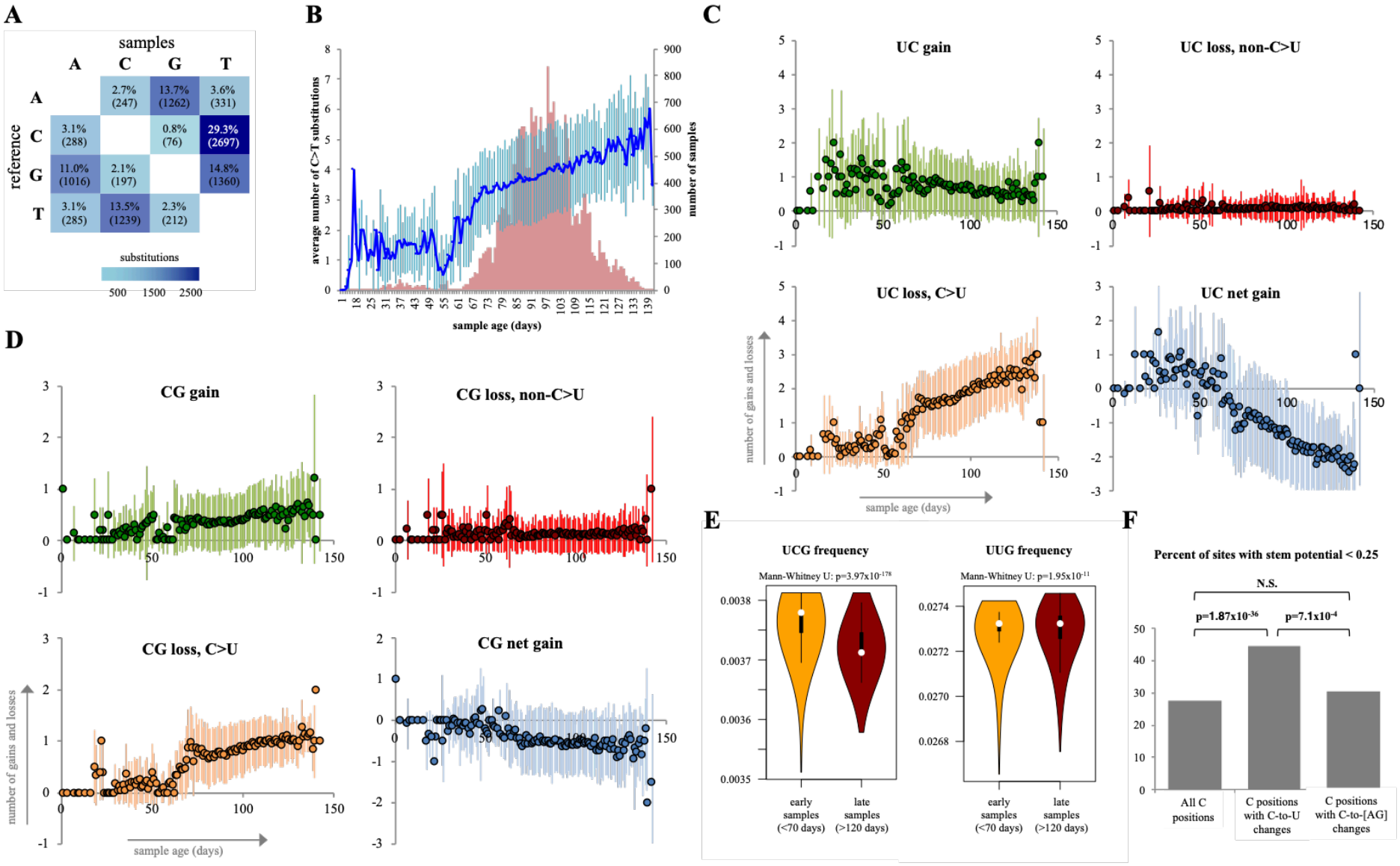

A
samples
A C G T
reference
A
C
G
T
2.7% (247)
13.7% (1262)
3.6% (331)
3.1% (288)
0.8% (76)
29.3% (2697)
11.0% (1016)
2.1% (197)
14.8% (1360)
3.1% (285)
13.5% (1239)
2.3% (212)
substitutions
500 1500 2500
B
average number of C>T substitutions
number of samples
sample age (days)
C
UC gain
UC loss, non-C>U
UC loss, C>U
UC net gain
number of gains and losses
sample age (days)
D
CG gain
CG loss, non-C>U
CG loss, C>U
CG net gain
number of gains and losses
sample age (days)
E
UCG frequency
Mann-Whitney U: p=3.97x10-178
UUG frequency
Mann-Whitney U: p=1.95x10-11
early samples (<70 days)
late samples (>120 days)
F
Percent of sites with stem potential < 0.25
N.S.
p=1.87x10-36
p=7.1x10-4
All C positions
C positions with C-to-U changes
C positions with C-to-[AG] changes

**Figure 1.** (A) SNV events observed between individual SARS-CoV-2 sample sequences (n=22,164) and the reference genome. (B) The number of C>U substitutions across sample dates. The average number of substitutions for each sampling day is plotted (blue line, left y-axis) with plus/minus one standard deviations as error bars. The number of samples for each day is shown as red bars (right y-axis). (C) The number of UpC gains and losses over time. The average number of gains and losses relative to the reference genome are shown as circles for each sampling day (x-axis). Error bars denote plus/minus one standard deviation. (D) Number of CpG gains and losses over time. Non-C>U corresponds to all the nucleotide substitution changes for both cytosine and guanine. (E) Comparison of early and late 5'-UCG-3' (left) and 5'-UUG-3' (right) motif frequencies. The median, interquartile ranges, and upper adjacent values is shown as the white circles, thick and thin black lines, respectively. (F) Folding potential of positions with C>U changes (see supplementary Methods). P-values from Fisher's exact test are shown above bars.

GISAID data of viral sequences including their date of collection allowed us to analyze viral nucleotide change dynamics over time. We have analyzed 22,164 SARS-CoV-2 genomes from GISAID deposited from December 24th, 2019 to May 23rd, 2020[20] (Figure S1 & S2; Table S1). By aligning these genomes to the SARS-CoV-2 reference (GenBank accession: NC_045512), we observed a total of 9,210 single nucleotide variations (SNVs) with C>U being the most abundant substitution and representing almost a third of all substitution events (Figure 1A). As almost half of the C positions in the reference are substituted to U in at least one sample, we expect a high degree of saturation. Downsizing the number of samples resulted in a consistent increase in C>U frequencies (Figure S3). The reported C>U frequency is therefore most likely an underestimate. Over this period of time, we find a steady increase in C>U substitutions (Figure 1B). One potential driver behind this increase could be the recently proposed APOBEC-mediated viral RNA editing[15-16].

As mentioned earlier, the majority of APOBEC enzymes preferentially acts on a 5'-UC-3' (5'-TC-3' for DNA substrate) motif[9–13]. We analyzed the context of genomic sites undergoing C>U changes and noticed a preference for 5'-UCG-3' motifs (Table S2). Next, we examined the 5'-UCG-3' motif by following the dynamics of UpC and CpG losses and gains over time (Figure 1C&D). We observe a progressive loss of both UpC and CpG dinucleotides due to C>U changes, but not other nucleotide changes (Figure 1C&D). The observed losses do not appear to be related to samples from specific continents (Figure S4). Plots for all dinucleotides are available in Figure S5. Further, the frequency of the full 5'-UCG-3' motif is found to be significantly lower in "early" (<70 days) compared to "late" (>120 days) samples (Figure 1E, left), while the frequency of 5'-UUG-3' motif is increased (Figure 1E, right).

Since APOBEC3 family members display a preference for RNA in open conformation as opposed to forming secondary structures[23], we calculated the folding potential of all genomic sites that include C>U substitutions (Figure 1F). Positions with C>U changes

are more often located in regions with low potential for forming secondary RNA structures. These observations are in agreement with the notion that members of the APOBEC family are the main drivers of cytosine deamination in SARS-CoV-2, thus contributing to the extensive viral genome editing[15-16].

Our CpG net gain analysis showed a progressive decline, indicating that the number of total CpG losses overrides their gain (Figure 1C). The observed CpG decline could indicate the ongoing adaptation of the virus to the new host or tissues where antiviral proteins targeting foreign CpG regions are differently expressed[24]. It was recently shown that synonymous substitutions disrupting CpG dinucleotides appear to contribute to the increased virulence of attenuated poliovirus[25], and we observe that little more than half of the C>U substitutions disrupting CpG dinucleotides were synonymous which is in agreement with Di Gioacchino and colleagues[26] (Figure S6). In the future, the SARS-CoV-2 virus may level up its CpG frequency with its host, achieving so-called host-cell mimicry[27], but the observed continuous CpG loss suggests that equilibrium between gains and losses have not yet been reached.

Previously, from a comparative analysis of different coronaviruses, Woo and colleagues concluded that CpG reduction and cytosine deamination are two independent selective forces that shape coronavirus codon bias[28]. Although selective forces may act differently upon motifs resulting from C>U changes, our results suggest a direct link between cytosine deamination and CpG reduction in SARS-CoV-2.

Previous studies reported that a reduced number of CpG motifs in HIV and other viruses played an important role in the viral replication inside the host cell, thus escaping ZAP protein activity[6]. Given the high expression levels of APOBEC and ZAP genes in COVID-19 patients[17], the involvement of ZAP-mediated restriction of SARS-CoV-2[8], and our observations, we hypothesize that as a consequence of APOBEC-mediated RNA editing, SARS-CoV-2 genome may escape host cell ZAP activity. Both APOBEC and ZAP are interferon-induced genes that act preferentially on ssRNA in open conformation[3,23,29]. Initially, APOBEC and ZAP enzymes may have overlapping preferred target motifs for their enzymatic functions (Figure 2). The catalytic activity of APOBEC on 5'-UC-3' leads to cytosine deamination which destroys ZAP's extremely specific acting site (CpG)[30]. The conversion of C>U allows viral RNA to escape from ZAP-mediated RNA destruction. Our analysis strongly supports the proposed model, but it merits future experimental validation. Since APOBEC enzymes can act on both RNA and DNA sequences, it is tempting to speculate that some other viruses may undergo CpG reduction similarly to SARS-CoV-2.

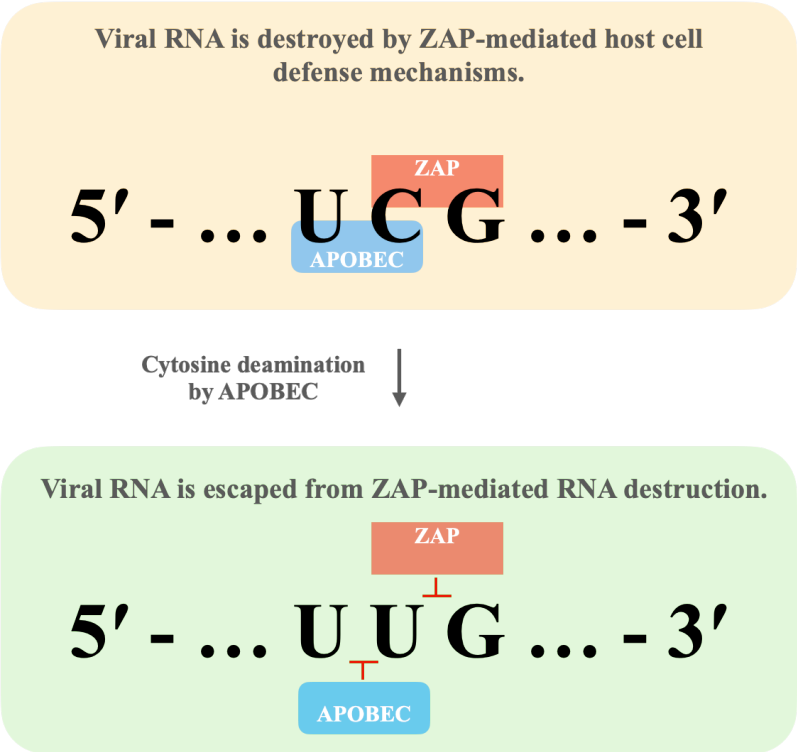



**Figure 2.** A model for the consequences of host-driven evolution by APOBEC enzymes on viral CpG dinucleotide composition.

In summary, the rapidly growing genomic data from SARS-CoV-2 (that importantly includes relevant metadata) and the moderate mutation rate of the virus has allowed us to observe a progressive increase of C-to-U substitutions resulting in CpG loss over a brief period of time. Our results thus link the dynamics of target sequences in the viral genome for two known host molecular defense mechanisms, the APOBEC and ZAP proteins.

## Materials and Methods

Samples and recording of substitutions

Samples were downloaded from GISAID[20] (https://www.gisaid.org/) on May 23rd, 2020. Only samples with available collection dates and with no more than 500 N's in the assembled sequence were used. December 24th, 2019 was set to day1. Assembled virus genomes were separately aligned to the reference genome (GenBank accession: NC_045512) using mafft[31], and mismatches were collected. Alignments with more than 500 gaps in the sample (due to incomplete assembly) and any alignments with gaps in the reference sequence were discarded. This resulted in 22,164 alignments. Only non-ambiguous bases and only mismatches without gaps in 3 positions up- and downstream were kept. To exclude alignment artifacts, mismatches in the first and last 100 positions in the reference genome were discarded.

The total number of substitutions is recorded from the combined set of pair-wise alignments, and a substitution at a given position in the reference is counted once regardless of the number of samples in which it is found. The timeline analysis of dinucleotide gains and losses were inferred from each sample's individual alignment to the reference.

Nucleotide motifs

For all di-, tri-, and tetra-nucleotide motifs containing C in the reference genome, the number of motifs with a C-to-U substitution, and the total number of the motifs were noted. The ratio between these two measures was compared to the expected ratio, defined as the number of C's with a C-to-U substitution divided by the total number of C's in the genome. The probability was calculated using a binomial distribution. All statistical tests were performed using Rstudio v. 1.1.414[32,33].

Folding potential

The reference sequence was divided into overlapping 500-nucleotides windows, each shifted by 10 nucleotides. These sequences were folded using RNAfold[34]. For each position in the reference genome, the number of windows in which the position was predicted to form a base-pair in a secondary structure was recorded. The folding potential of a position was defined as the fraction of windows with predicted base-pairing of that position. Hence, a position predicted to participate in a base-pair in all 500-nt windows has a folding potential of 1.

## Supplementary Material

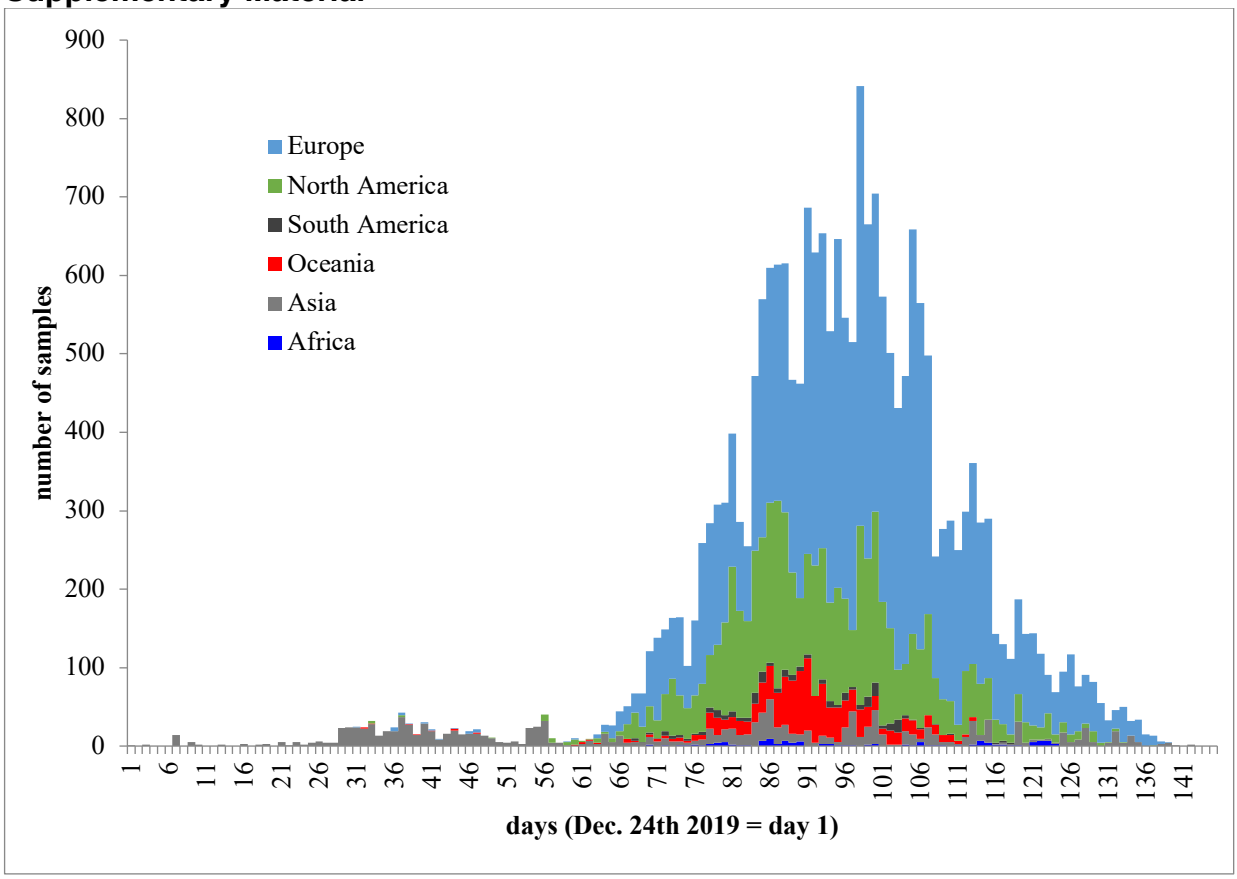



**Figure S1.** Number of samples from different time points. December 24th 2019 is set to day 1. Samples are colored according to their geographic region.

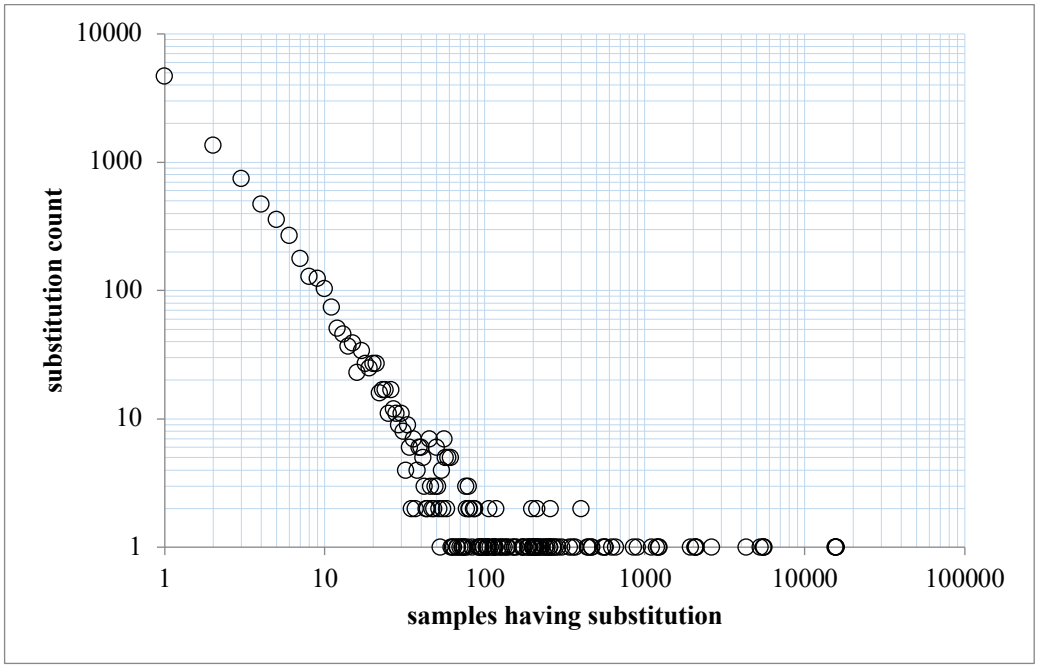



**Figure S2.** Number of samples in which the substitutions are observed.

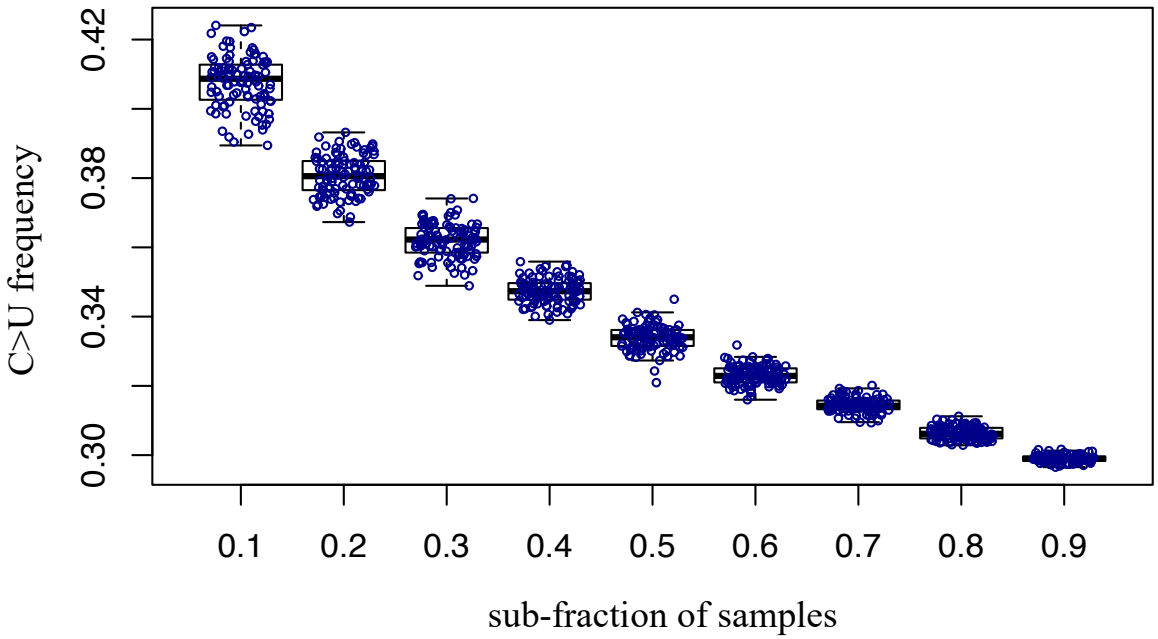
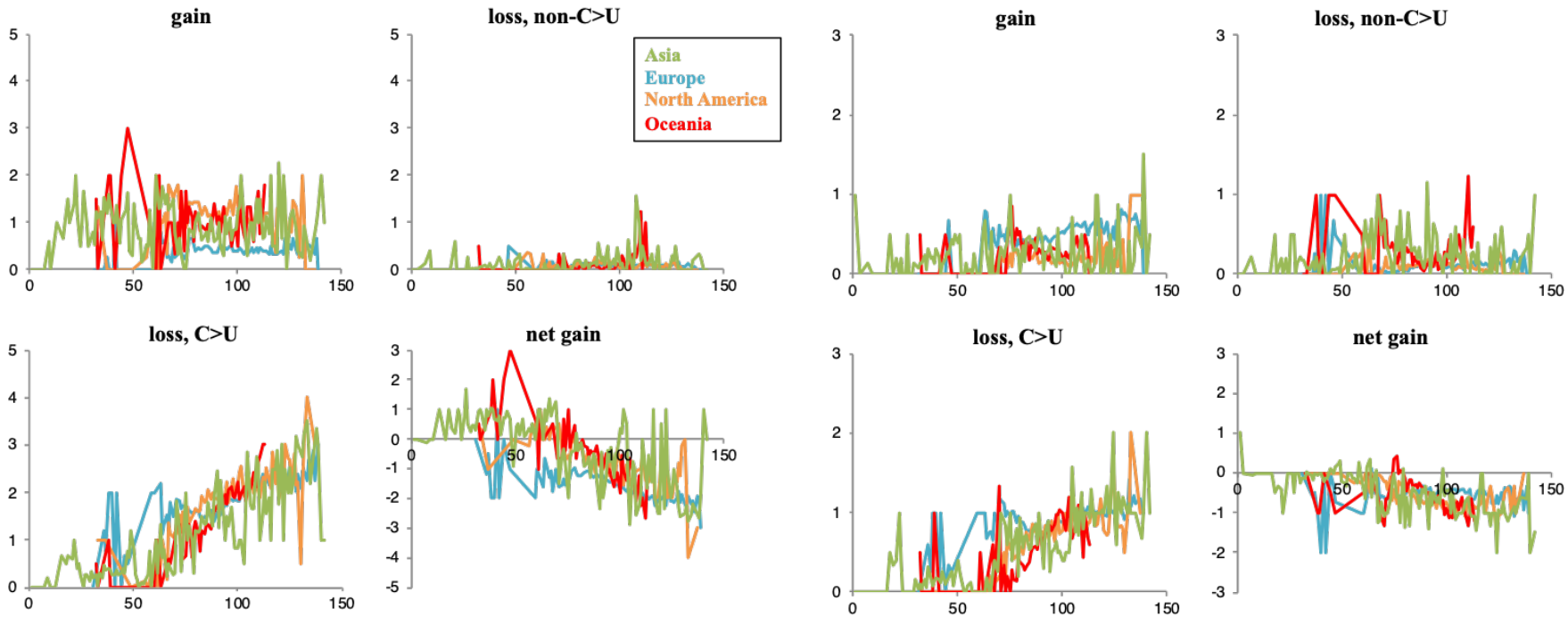



**Figure S3.** The inverse relationship of C>U frequency and the number of samples. Down-sizing the number of samples leads to higher C>U frequency (0.1 on the axis means 10% of the total samples). The blue dots are the individual simulated data sets, and the boxes denote their distribution.




220
221 **Figure S4.** UpC (left) and CpG (right) dynamics across continents.

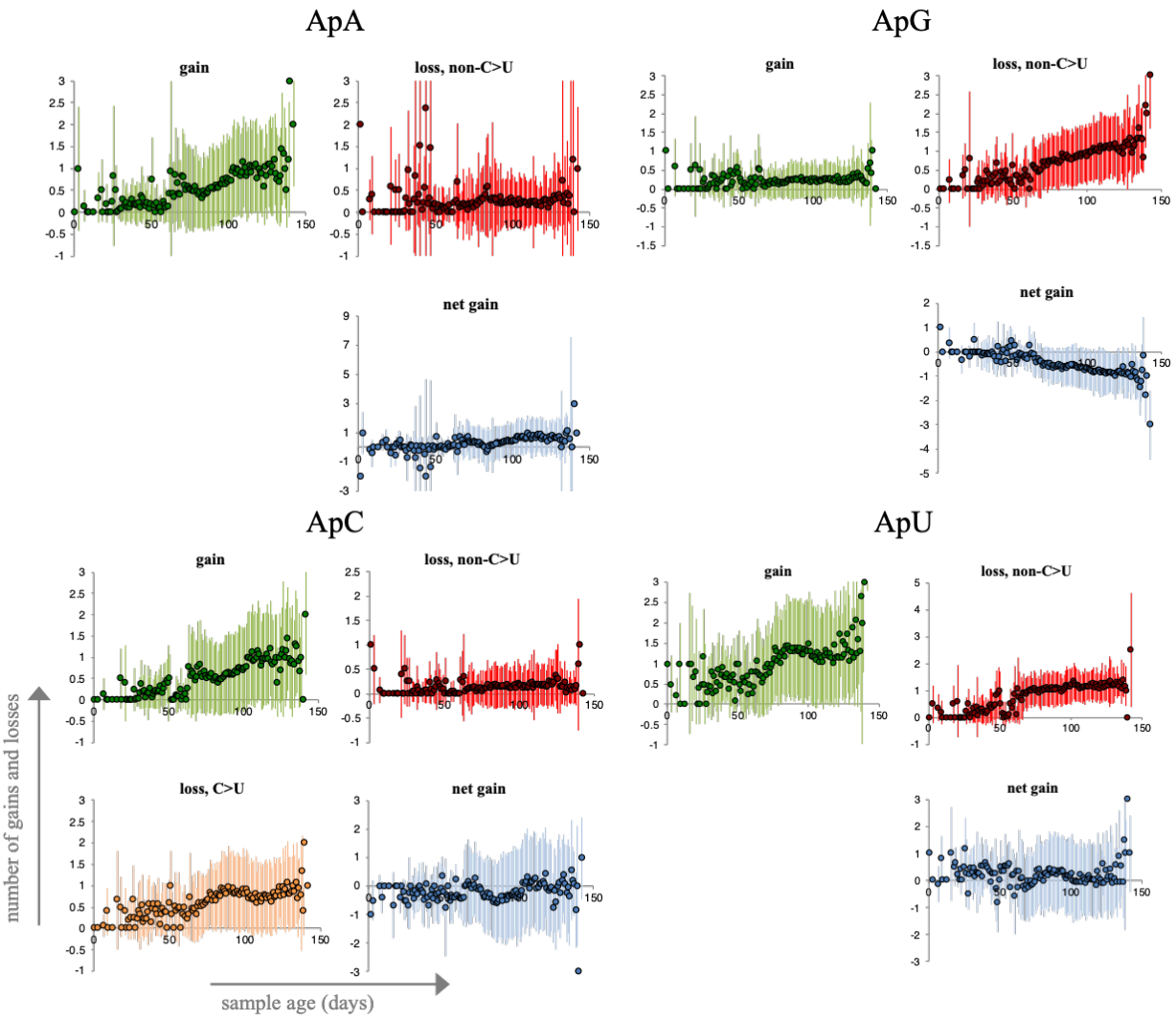



**Figure S5.** Number of dinucleotide gains and losses over time.

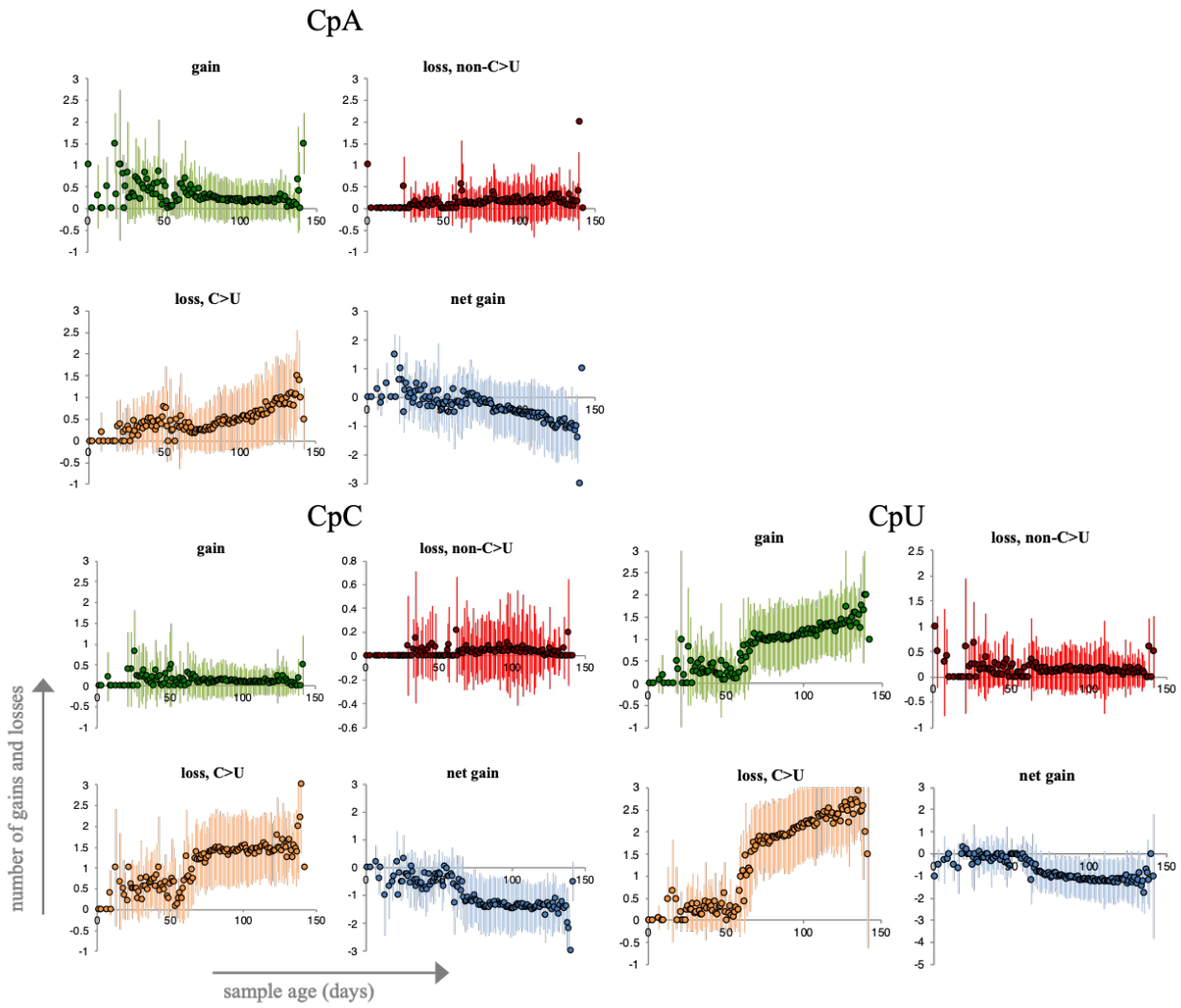



**Figure S5 (cont.).** Number of dinucleotide gains and losses over time.

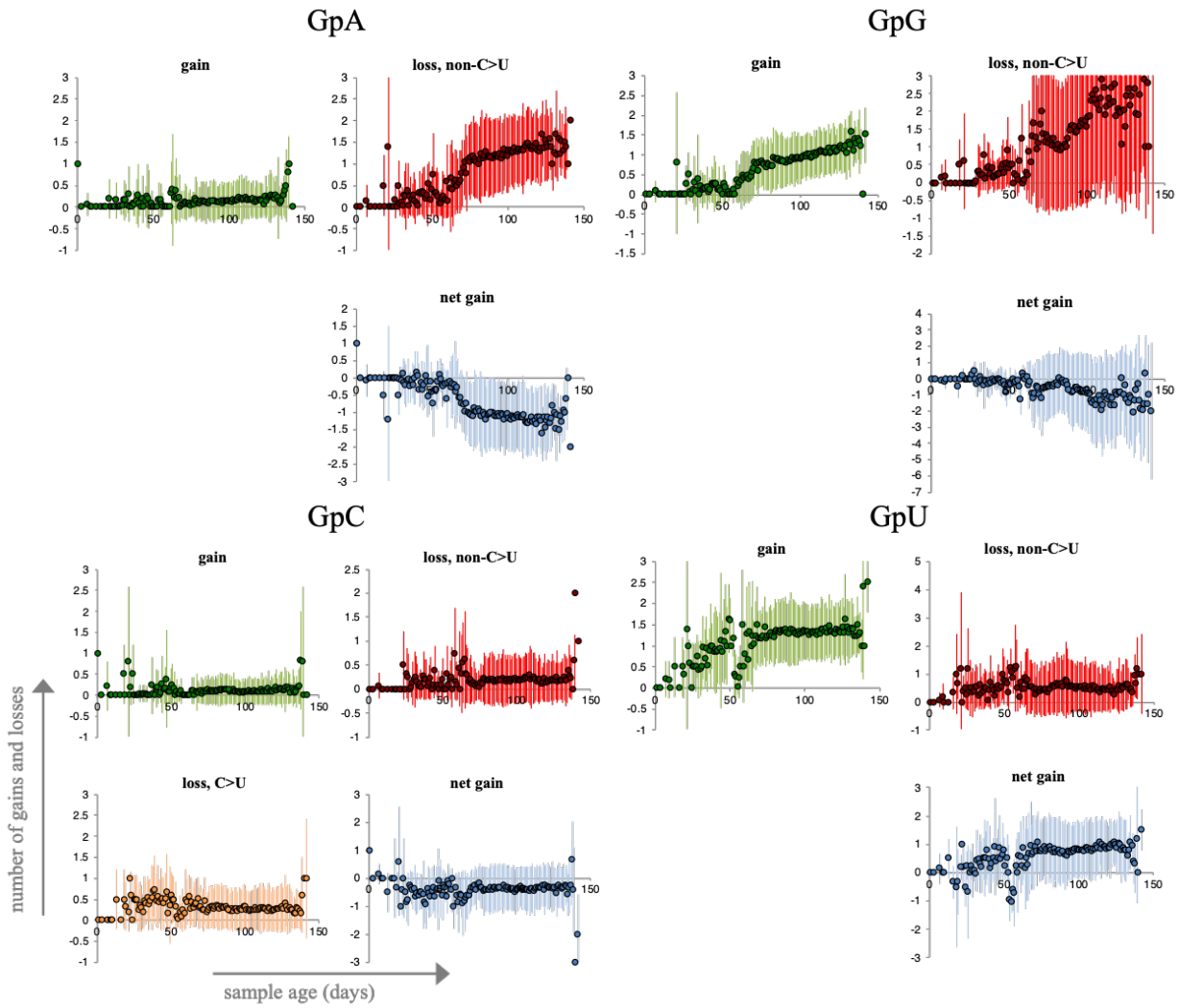



**Figure S5 (cont.).** Number of dinucleotide gains and losses over time.

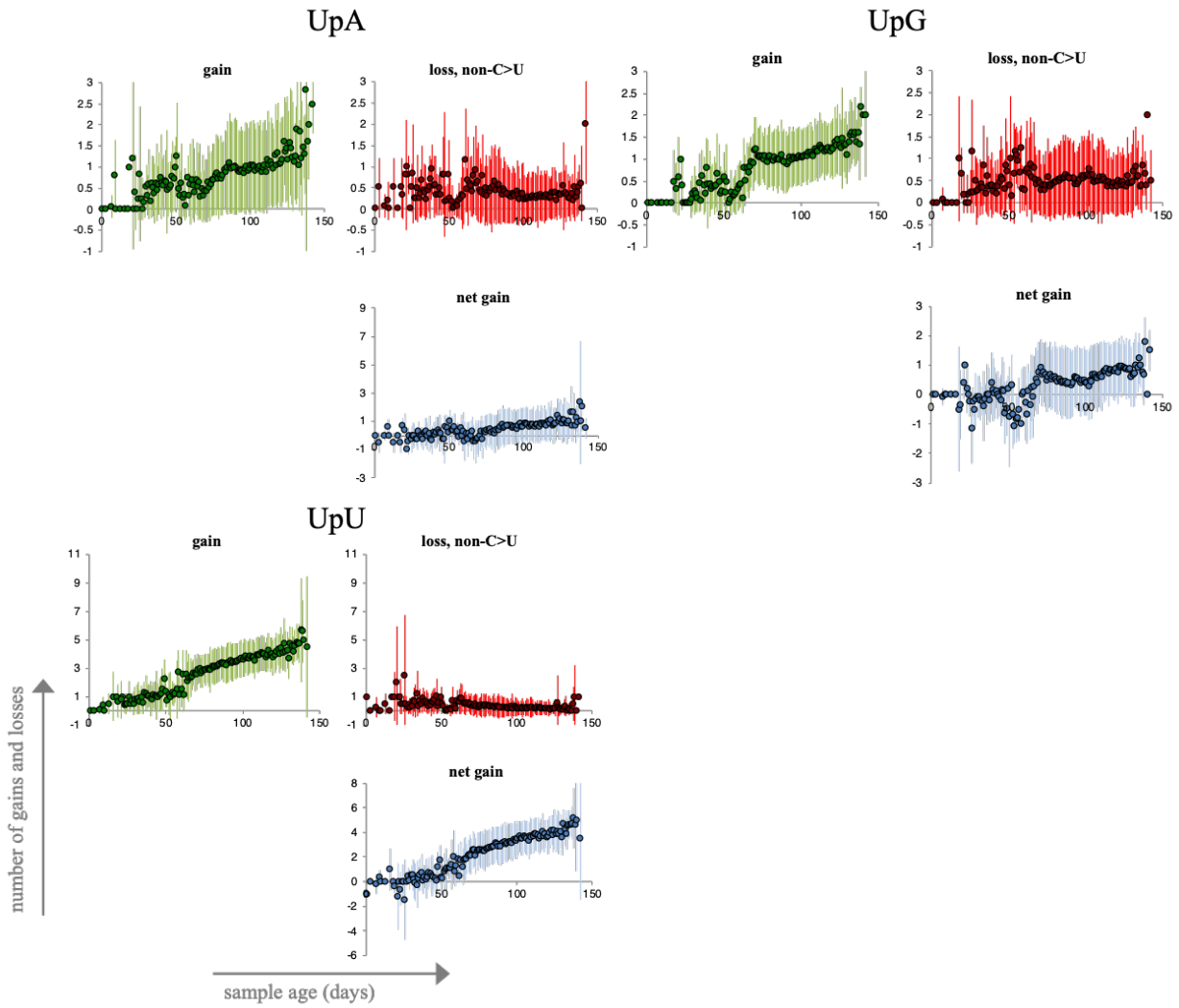



**Figure S5 (cont.).** Number of dinucleotide gains and losses over time.

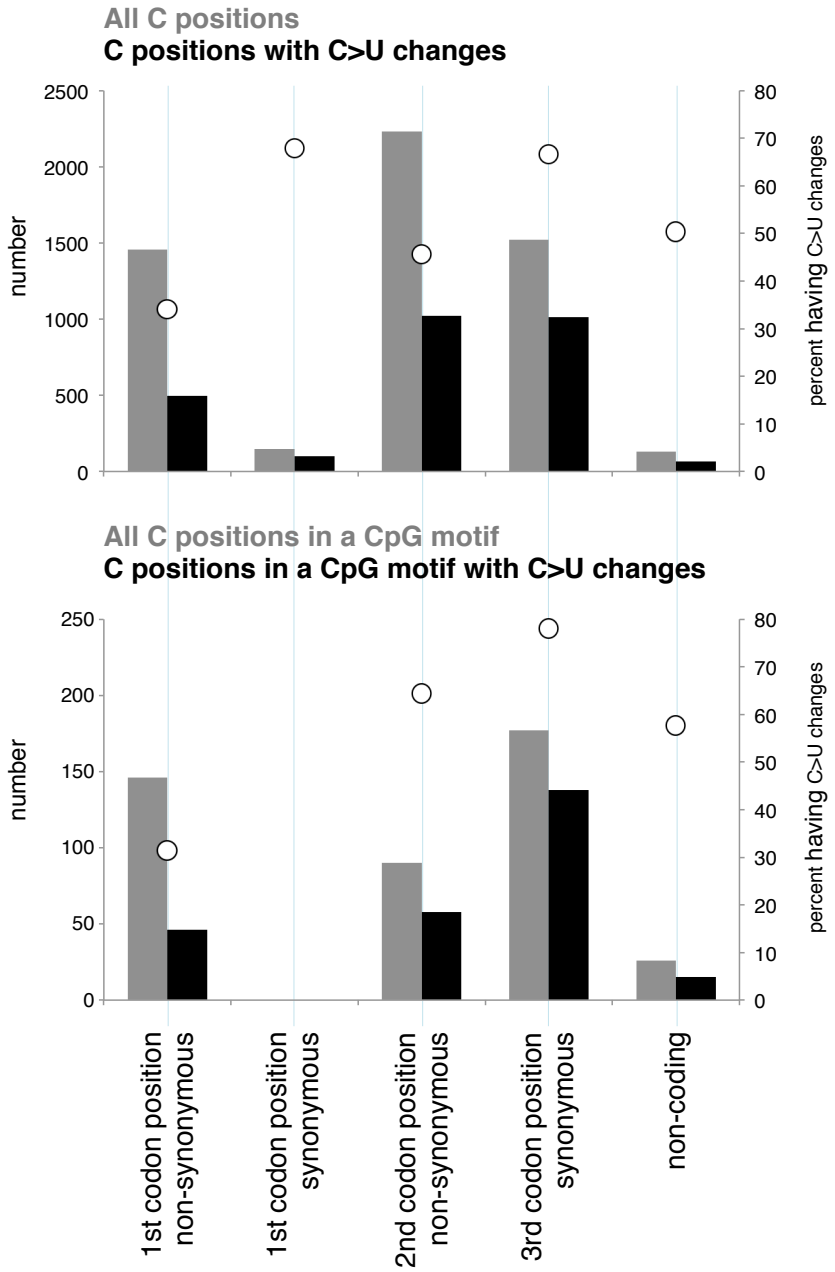



**Figure S6.** Plots showing the number C positions (top) and CpG positions (bottom) in the reference genome (black bars) and the number of these that have C>U changes (grey bars). Numbers are shown for different codon positions as indicated below the plots. CpG positions are defined as the position of the C in the motif, and positions with no changes are defined as either synonymous or non-synonymous based on the effect of a potential C>U change. The percentage of bases and motifs that have C>U changes are shown as open circles (right y-axis).

**Table S2.** The most enriched motifs undergoing C>U changes.

| C>U | | | |
|---|---|---|---|
| motif | motifs with C-to-U change/total number of motifs | odds-ratio | uncorrected p-value |
| UCG | 72/113 | 1.30 | 0.0012387 |
| UUCG | 27/39 | 1.41 | 0.0088244 |
| CG | 257/439 | 1.19 | 4.56E-05 |
| CGG | 55/76 | 1.47 | 3.19E-05 |
| CGGA | 13/16 | 1.65 | 0.0088056 |
| CGGU | 22/29 | 1.54 | 0.0030566 |
| GUGC | 75/113 | 1.35 | 0.0001577 |
| UGCU | 148/255 | 1.18 | 0.0026032 |
| CUA | 307/561 | 1.11 | 0.0044054 |
| UUAC | 137/241 | 1.16 | 0.0096264 |
| UGAC | 87/137 | 1.29 | 0.0004834 |
| CAAC | 112/194 | 1.18 | 0.0098096 |
| UAAC | 88/135 | 1.33 | 0.00012 |


## Acknowledgments

We thank all laboratories which have contributed sequences to the GISAID database and Zhadyra Yerkesh for giving her comments and helpful discussions.

This work was supported by funding from King Abdullah University of Science and Technology (KAUST), Office of Sponsored Research (OSR). Work in AP's laboratory is supported by the KAUST faculty baseline fund (BAS/1/1020-01-01).


## Author Contributions

A.P. supervised the project. M.S. and T.M. designed experiments. T.M. and QG performed bioinformatic analysis. M.S. wrote the draft of the manuscript. All authors discussed, edited, read, and agreed to the final version of the manuscript.

## References


1. Atkinson, N. J., Witteveldt, J., Evans, D. J. & Simmonds, P. The influence of CpG and UpA dinucleotide frequencies on RNA virus replication and characterization of the innate cellular pathways underlying virus attenuation and enhanced replication. *Nucleic Acids Res.* (2014) doi:10.1093/nar/gku075.
2. Gu, H., Fan, R. L. Y., Wang, D. & Poon, L. L. M. Dinucleotide evolutionary dynamics in influenza A virus. *Virus Evol.* (2019) doi:10.1093/ve/vez038.
3. Chemudupati, M. *et al.* From APOBEC to ZAP: Diverse mechanisms used by cellular restriction factors to inhibit virus infections. *Biochimica et Biophysica Acta - Molecular Cell Research* (2019) doi:10.1016/j.bbamcr.2018.09.012.
4. Kluge, S. F., Sauter, D. & Kirchhoff, F. SnapShot: Antiviral Restriction Factors. *Cell* (2015) doi:10.1016/j.cell.2015.10.019.
5. Gao, G., Guo, X. & Goff, S. P. Inhibition of retroviral RNA production by ZAP, a CCCH-type zinc finger protein. *Science (80-. ).* (2002) doi:10.1126/science.1074276.
6. Takata, M. A. *et al.* CG dinucleotide suppression enables antiviral defence targeting non-self RNA. *Nature* (2017) doi:10.1038/nature24039.
7. Harris, R. S. & Dudley, J. P. APOBECs and virus restriction. *Virology* (2015) doi:10.1016/j.virol.2015.03.012.
8. Nchioua, R. *et al.* The Zinc Finger Antiviral Protein restricts SARS-CoV-2. *bioRxiv* (2020) doi:10.1101/2020.06.04.134379.
9. Salter, J. D. & Smith, H. C. Modeling the Embrace of a Mutator: APOBEC Selection of Nucleic Acid Ligands. *Trends in Biochemical Sciences* (2018) doi:10.1016/j.tibs.2018.04.013.
10. Sharma, S. *et al.* APOBEC3A cytidine deaminase induces RNA editing in monocytes and macrophages. *Nat. Commun.* (2015) doi:10.1038/ncomms7881.
11. Sharma, S. *et al.* Mitochondrial hypoxic stress induces widespread RNA editing by APOBEC3G in natural killer cells. *Genome Biol.* (2019) doi:10.1186/s13059-019-1651-1.
12. Salter, J. D., Bennett, R. P. & Smith, H. C. The APOBEC Protein Family: United by Structure, Divergent in Function. *Trends in Biochemical Sciences* (2016) doi:10.1016/j.tibs.2016.05.001.
13. Thielen, B. K. *et al.* Innate immune signaling induces high levels of TC-specific deaminase activity in primary monocyte-derived cells through expression of APOBEC3A isoforms. *J. Biol. Chem.* (2010) doi:10.1074/jbc.M110.102822.
14. Milewska, A. *et al.* APOBEC3-mediated restriction of RNA virus replication. *Sci. Rep.* (2018) doi:10.1038/s41598-018-24448-2.
15. Giorgio, S. Di, 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) doi:10.1126/sciadv.abb5813.
16. Simmonds, P. Rampant C->U hypermutation in the genomes of SARS-CoV-2 and other coronaviruses - causes and consequences for their short and long evolutionary trajectories. *biorxiv* (2020) doi:10.1101/2020.05.01.072330.

17. Blanco-Melo, D. *et al.* Imbalanced Host Response to SARS-CoV-2 Drives Development of COVID-19. *Cell* (2020) doi:10.1016/j.cell.2020.04.026.
18. Fehr, A. R. & Perlman, S. Coronaviruses: An overview of their replication and pathogenesis. in *Coronaviruses: Methods and Protocols* (2015). doi:10.1007/978-1-4939-2438-7_1.
19. Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C. & Garry, R. F. The proximal origin of SARS-CoV-2. *Nature Medicine* (2020) doi:10.1038/s41591-020-0820-9.
20. Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob. Challenges* (2017) doi:10.1002/gch2.1018.
21. Hadfield, J. *et al.* NextStrain: Real-time tracking of pathogen evolution. *Bioinformatics* (2018) doi:10.1093/bioinformatics/bty407.
22. Salemi, M. *et al.* Severe Acute Respiratory Syndrome Coronavirus Sequence Characteristics and Evolutionary Rate Estimate from Maximum Likelihood Analysis. *J. Virol.* (2004) doi:10.1128/jvi.78.3.1602-1603.2004.
23. McDaniel, Y. Z. *et al.* Deamination hotspots among APOBEC3 family members are defined by both target site sequence context and ssDNA secondary structure. *Nucleic Acids Res.* (2020) doi:10.1093/nar/gkz1164.
24. Xia, X. Extreme genomic CpG deficiency in SARS-CoV-2 and evasion of host antiviral defense. *Mol. Biol. Evol.* (2020) doi:10.1093/molbev/msaa094.
25. Stern, A. *et al.* The Evolutionary Pathway to Virulence of an RNA Virus. *Cell* (2017) doi:10.1016/j.cell.2017.03.013.
26. Di Gioacchino, A. *et al.* The Heterogeneous Landscape and Early Evolution of Pathogen-Associated CpG Dinucleotides in SARS-CoV-2. *SSRN Electron. J.* (2020) doi:10.2139/ssrn.3611280.
27. Greenbaum, B. D., Levine, A. J., Bhanot, G. & Rabadan, R. Patterns of evolution and host gene mimicry in influenza and other RNA viruses. *PLoS Pathog.* (2008) doi:10.1371/journal.ppat.1000079.
28. Woo, P. C. Y., Wong, B. H. L., Huang, Y., Lau, S. K. P. & 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) doi:10.1016/j.virol.2007.08.010.
29. Luo, X. *et al.* Molecular Mechanism of RNA Recognition by Zinc-Finger Antiviral Protein. *Cell Rep.* (2020) doi:10.1016/j.celrep.2019.11.116.
30. Meagher, J. L. *et al.* Structure of the zinc-finger antiviral protein in complex with RNA reveals a mechanism for selective targeting of CG-rich viral sequences. *Proc. Natl. Acad. Sci. U. S. A.* (2019) doi:10.1073/pnas.1913232116.
31. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* (2013) doi:10.1093/molbev/mst010.
32. Booth, D. S. *et al.* RStudio: Integrated Development for R. *Nature* (2018) doi:10.1108/eb003648.
33. R Development Core Team. R: A language and environment for statistical

computing. *Vienna, Austria* (2017) doi:R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

34. Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms Mol. Biol.* (2011) doi:10.1186/1748-7188-6-26.