Title: Short sequence motif dynamics in the SARS-CoV-2 genome suggest a role for cytosine deamination in CpG reduction.

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-finger antiviral protein.

Keywords: virus evolution, genome evolution, genome biology, virus-host interaction.

ABSTRACT

RNA viruses use CpG reduction to evade the host cell defense, but the driving mechanisms are still largely unknown. In an attempt to address this we used a rapidly growing genomic dataset of SARS-CoV-2 with relevant metadata information. Remarkably, by simply ordering SARS-CoV-2 genomes by their date of collection, we find a progressive increase of C-to-U substitutions resulting in 5'-UCG-3' motif reduction that in turn have reduced the CpG frequency over just a few months of observation. 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 (Atkinson et al. 2014). 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 (Gu et al. 2019).

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 (Chemudupati et al. 2019). The antiviral proteins directly

43 involved in the restriction of viral replication and transcription are zinc-finger antiviral
44 protein (ZAP) and Apolipoprotein B Editing Complex (APOBEC) (Kluge et al. 2015). The
45 defensive role of ZAP on HIV-1 manifestation in the host cell has been extensively
46 studied (Ficarelli et al. 2019). ZAP selectively binds to viral CpG regions and recruits
47 RNA exosome complex resulting in viral RNA degradation (Harris and Dudley 2015;
48 Takata et al. 2017). Recently, ZAP was shown to restrict SARS-CoV-2 production upon
49 the induction by interferons (Nchioua et al. 2020).
50
51 The APOBEC protein family is another example of host antiviral enzymes. Seven
52 human APOBEC proteins are known for catalyzing cytosine to uracil deamination in
53 foreign single-stranded DNA (ssDNA) and RNA (ssRNA) (Sharma et al. 2015; Salter
54 and Smith 2018; Sharma et al. 2019; Giorgio et al. 2020). APOBEC proteins induce
55 extensive nucleotide changes, which is restrictive to viral replication, but they also
56 contribute to the genetic diversity and adaptation of a virus (Kim et al. 2014; Harris and
57 Dudley 2015). Enzymatic target motifs for most of the APOBEC enzymes have been
58 experimentally identified, among which the most common were 5'-[T/U]C-3' and 5'-CC-
59 3' for DNA/RNA substrates (Thielen et al. 2010; Sharma et al. 2015; Salter and Smith
60 2018; Sharma et al. 2019). Some members of APOBEC proteins inhibit replication of
61 human coronavirus, HCoV-NL63, but their role in hypermutation of the viral genome is
62 not clear (Milewska et al. 2018). However, it was recently suggested that the SARS-
63 CoV-2 undergoes genome editing by host-dependent RNA-editing proteins such as
64 APOBEC (C>U) and ADAR (A>G) (Giorgio et al. 2020; Simmonds 2020). mRNA

65 transcriptional response data from COVID-19 patients demonstrated that APOBEC3A
66 (A3A) belongs to one of the most expressed genes in response to SARS-CoV-2, and
67 the highest among APOBEC family (Blanco-Melo et al. 2020).
68
69 SARS-CoV-2 is an enveloped positive-sense ssRNA virus, belonging
70 to *Coronaviridae* family, which is characterized by large genomes (~30K nucleotide
71 long), high recombination frequencies and genomic proofreading mechanism (Fehr and
72 Perlman 2015). The COVID-19 pandemic is bringing intensive scientific attention to the
73 virus, and thousands of viral genomic sequences with metadata are publicly available
74 (Elbe and Buckland-Merrett 2017; Hadfield et al. 2018). Given the large amount of
75 available data and the relatively low mutation rate of the virus (Salemi et al. 2004), we
76 aimed to monitor its genomic evolution on a very brief time scale during the COVID-19
77 pandemic. We show progressive C>U substitutions in SARS-CoV-2 genome within the
78 timeframe of five months. We highlight the role of C>U substitutions in the reduction of
79 5'-UCG-3' motifs and hypothesize that this progressive decrease is driven by host
80 APOBEC and ZAP enzymes.

Results and Discussion

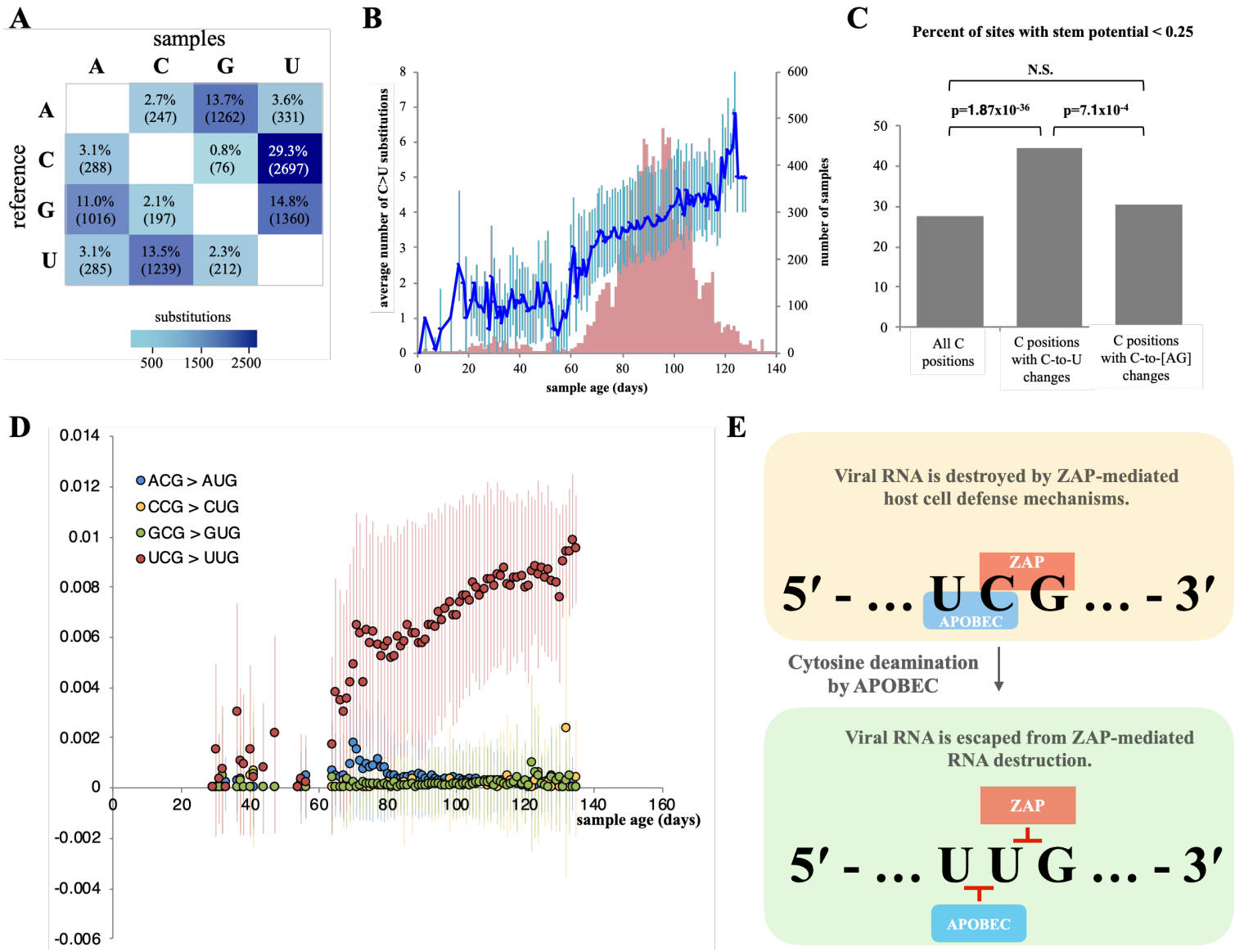


Figure 1. (A) SNV events observed between individual SARS-CoV-2 sample sequences (n=22,164) and the reference genome. (B) The number of C>U substitutions across sample dates. The average number of substitutions for each sampling day is plotted (blue line, left y-axis) with plus/minus one standard deviations as error bars. The number of samples for each day is shown as red bars (right y-axis). (C) Folding potential of positions with C>U changes (see Materials and Methods). P-values from Fisher's exact test are shown above bars. (D) The fraction of [A/C/G/U]CG triplets that are changed to [A/C/G/U]UG over time. The average fractions, relative to the reference genome, are shown as circles for each sampling day (x-axis). Error bars denote plus/minus one standard deviation. (E) A model for the consequences of host-driven evolution by APOBEC enzymes on viral CpG dinucleotide composition.

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 (Elbe and Buckland-Merrett 2017) (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 one third of all substitution events (Figure 1A). As nearly 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, suggesting that a considerable

105 fraction of the observed C>U changes represent multiple, independent events (Figure
106 S3). The reported C>U frequency is therefore most likely an underestimate. Over this
107 period of five months, we find a steady increase in C>U substitutions (Figure 1B). One
108 potential driver behind this increase could be the recently proposed APOBEC-mediated
109 viral RNA editing (Giorgio et al. 2020; Simmonds 2020). Since APOBEC3 family
110 members display a preference for RNA in open conformation as opposed to forming
111 secondary structures (McDaniel et al. 2020), we calculated the folding potential of all
112 genomic sites that include C>U substitutions (Figure 1C). Positions with C>U changes
113 are more often located in regions with low potential for forming secondary RNA
114 structures. These observations are in agreement with the notion that members of the
115 APOBEC family are the main drivers of cytosine deamination in SARS-CoV-2, thus
116 contributing to the extensive viral genome editing (Giorgio et al. 2020; Simmonds 2020).
117
118 The majority of APOBEC enzymes preferentially acts on a 5'-UC-3' (5'-TC-3' for DNA
119 substrate) motif (Thielen et al. 2010; Sharma et al. 2015; Salter and Smith 2018;
120 Sharma et al. 2019). We searched for possible APOBEC genetic footprints (5'-UC-3' >
121 5'-UU-3') in viral dinucleotide frequencies for the 5-month period (Figure S4). Among all
122 dinucleotides, UpC showed the highest degree of decrease, while UpU exerted the
123 highest rates of increase. CpG, CpA, CpC, CpU and GpA also showed a negative net
124 gain but not as prominent as UpC, CpA or ApG. We find that the majority of dinucleotide
125 losses were due to C>U changes, in agreement with a recent study by Rice and
126 colleagues (Rice et al. 2020). The remaining dinucleotides showed relative stability
127 across the time period. Interestingly, recent reports showed the reduction of CpG

128 dinucleotides in 5'-UTR and ORF1ab regions (Wang et al. 2020; Wei et al. 2020; Zhang
129 et al. 2020). When analyzing the context of genomic sites undergoing C>U changes we
130 noticed an enrichment for 5'-UCG-3' motifs (Table S2). To assess the contribution of
131 C>U changes in CpG loss, we examined the dynamics of [A/C/G/U]CG trinucleotides
132 over time (Figure 1D). The progressive change (~1% over a 5-month period) of 5'-UCG-
133 3' to 5'-UUG-3' is most striking when supported by a larger number of genomes (days
134 70 to 115), whereas no such pattern is observed for the other trinucleotides (Figure 1D).
135 The association between cytosine deamination and CpG loss is further underlined by
136 the rapid, progressive increase in 5'-UCG-3' > 5'-UUG-3' changes compared to other 5'-
137 UC[A/C/U]-3' motifs (Figure S5).
138
139 Our CpG net gain analysis showed a decline, indicating that the number of total CpG
140 losses overrides their gain (Figure S4), in agreement with recently published
141 observations (Wang et al. 2020; Wei et al. 2020; Zhang et al. 2020). Similar to other
142 coronavirus genomes, SARS-CoV-2 has a low CpG frequency around 1.5%, while the
143 expected frequency is 3.6% (Wang et al. 2020). The observed CpG decline could
144 indicate the ongoing adaptation of the virus to the new host or tissues where antiviral
145 proteins targeting foreign CpG regions are differently expressed (Xia 2020). It was
146 recently shown that synonymous substitutions disrupting CpG dinucleotides appear to
147 contribute to the increased virulence of attenuated poliovirus (Stern et al. 2017). We
148 observe that more than half of the C>U substitutions disrupting CpG dinucleotides were
149 synonymous, consistent with recent studies (Di Gioacchino et al. 2020; Rice et al.
150 2020) (Figure S6). In the future, the SARS-CoV-2 virus may level up its CpG frequency

151 with its host, achieving so-called host-cell mimicry (Greenbaum et al. 2008), but the
152 observed CpG loss suggests that equilibrium between gains and losses have not yet
153 been reached.
154
155 Previously, from a comparative analysis of different coronaviruses, Woo and colleagues
156 concluded that CpG reduction and cytosine deamination are two independent selective
157 forces that shape coronavirus codon bias (Woo et al. 2007). Although selective forces
158 may act differently upon motifs resulting from C>U changes, our results suggest a direct
159 link between cytosine deamination and CpG reduction in SARS-CoV-2.
160
161 Previous studies reported that a reduced number of CpG motifs in HIV and other
162 viruses played an important role in the viral replication inside the host cell, thus
163 escaping ZAP protein activity (Takata et al. 2017). Given the high expression levels of
164 APOBEC and ZAP genes in COVID-19 patients (Blanco-Melo et al. 2020), the
165 involvement of ZAP-mediated restriction of SARS-CoV-2 (Nchioua et al. 2020), and our
166 observations, we hypothesize that as a consequence of APOBEC-mediated RNA
167 editing, SARS-CoV-2 genome may escape host cell ZAP activity. Both APOBEC and
168 ZAP are interferon-induced genes that act preferentially on ssRNA in open conformation
169 (Chemudupati et al. 2019; Luo et al. 2020; McDaniel et al. 2020). Initially, APOBEC and
170 ZAP enzymes may have overlapping preferred target motifs for their enzymatic
171 functions (Figure 1E). The catalytic activity of APOBEC on 5'-UC-3' leads to cytosine
172 deamination, which destroys ZAP's extremely specific acting site (CpG) (Meagher et al.
173 2019). The conversion of C>U allows viral RNA to escape from ZAP-mediated RNA

174 destruction. A recent study hypothesized that both ZAP and APOBEC provide selective
175 pressure that drives the adaptation of SARS-CoV-2 to its host (Wei et al. 2020). Here
176 we provided one of the potential mechanisms that contribute to CpG reduction in SARS-
177 CoV-2. Our analysis together with other recent studies strongly supports the proposed
178 model, but it merits future experimental validation. Since APOBEC enzymes can act on
179 both RNA and DNA sequences, it is tempting to speculate that some other viruses may
180 undergo CpG reduction similarly to SARS-CoV-2.
181
182 In summary, the rapidly growing genomic data from SARS-CoV-2 (that importantly
183 includes relevant metadata) and the moderate mutation rate of the virus has allowed us
184 to observe a progressive increase of C-to-U substitutions resulting in CpG loss over a
185 brief period of time. Notably, this was achieved using a phylogeny-free approach,
186 exclusively considering the temporal aspect of the samples. Given that APOBEC
187 proteins induce extensive mutations in a viral genome which are mostly detrimental,
188 here we provide an example of a rather unexpected role of these proteins in facilitating
189 the adaptation of a virus to the host. To our knowledge, this is the first study linking the
190 dynamics of viral genome mutation to two known host molecular defense mechanisms,
191 involving the APOBEC and ZAP proteins.

192 **Materials and Methods**

193 Samples and recording of substitutions

194 Samples were downloaded from GISAID (https://www.gisaid.org/) on May 23rd, 2020
195 (Elbe and Buckland-Merrett 2017). Only samples with available collection dates and
196 with no more than 500 N's in the assembled sequence were used. December 24th,
197 2019 was set to day1. Assembled virus genomes were separately aligned to the
198 reference genome (GenBank accession: NC_045512) using MAFFT (Katoh and
199 Standley 2013), and mismatches were collected. Alignments with more than 500 gaps
200 in the sample (due to incomplete assembly) and any alignments with gaps in the
201 reference sequence were discarded. This resulted in 22,164 alignments. Only non-
202 ambiguous bases and only mismatches without gaps in 3 positions up- and downstream
203 were kept. To exclude alignment artifacts, mismatches in the first and last 100 positions
204 in the reference genome were discarded.

205

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

211

212 Nucleotide motifs

213 For all di-, tri-, and tetra-nucleotide motifs containing C in the reference genome, the
214 number of motifs with a C-to-U substitution, and the total number of the motifs were

215 noted. The ratio between these two measures was compared to the expected ratio,
216 defined as the number of C's with a C-to-U substitution divided by the total number of
217 C's in the genome. The probability was calculated using a binomial distribution. All
218 statistical tests were performed using Rstudio v. 1.1.414 (Booth et al. 2018).
219
220 Folding potential
221 The reference sequence was divided into overlapping 500-nucleotides windows, each
222 shifted by 10 nucleotides. These sequences were folded using RNAfold (Lorenz et al.
223 2011). For each position in the reference genome, the number of windows in which the
224 position was predicted to form a base-pair in a secondary structure was recorded. The
225 folding potential of a position was defined as the fraction of windows with predicted
226 base-pairing of that position. Hence, a position predicted to participate in a base-pair in
227 all 500-nt windows has a folding potential of 1.

Supplementary Material

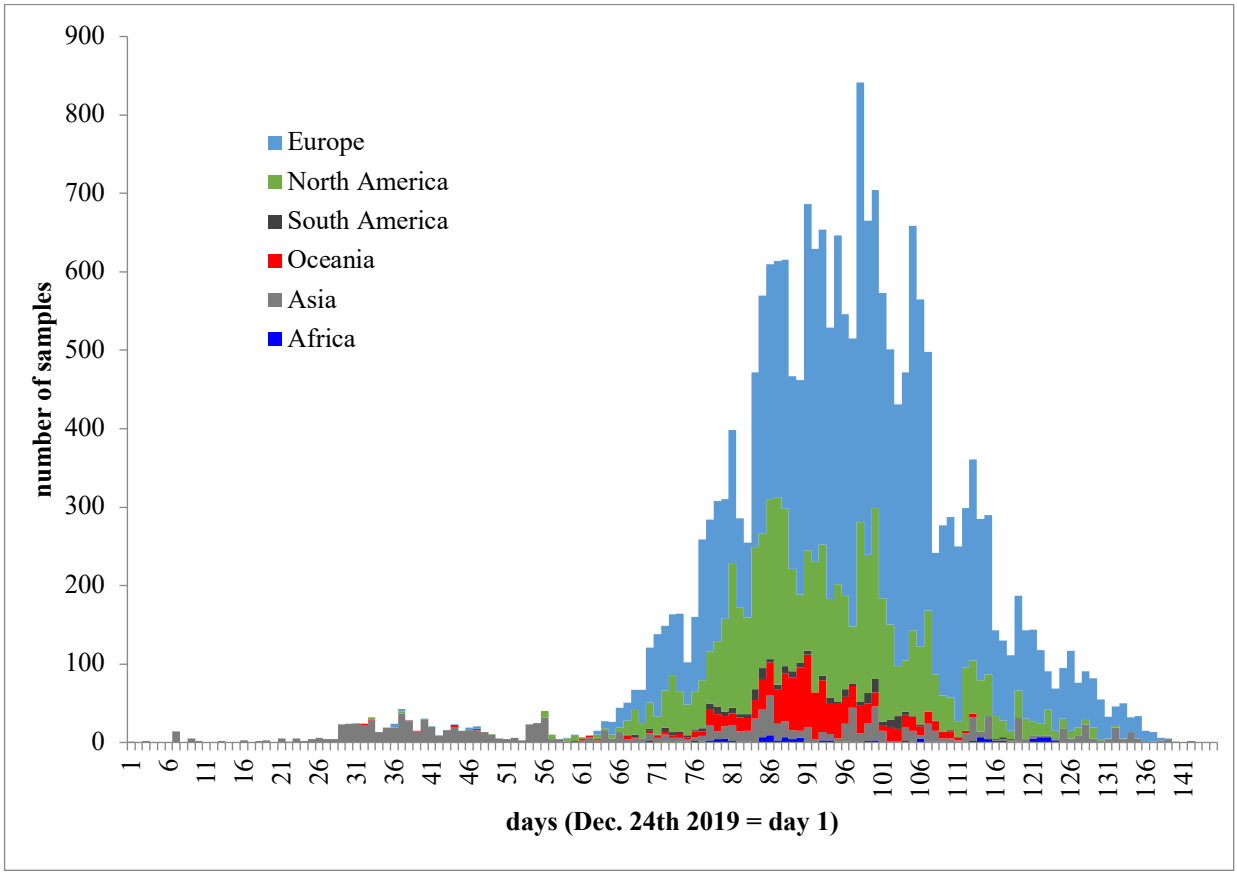


Figure S1. Number of samples from different time points. December 24th 2019 is set to day 1. Samples are colored according to their geographic region.

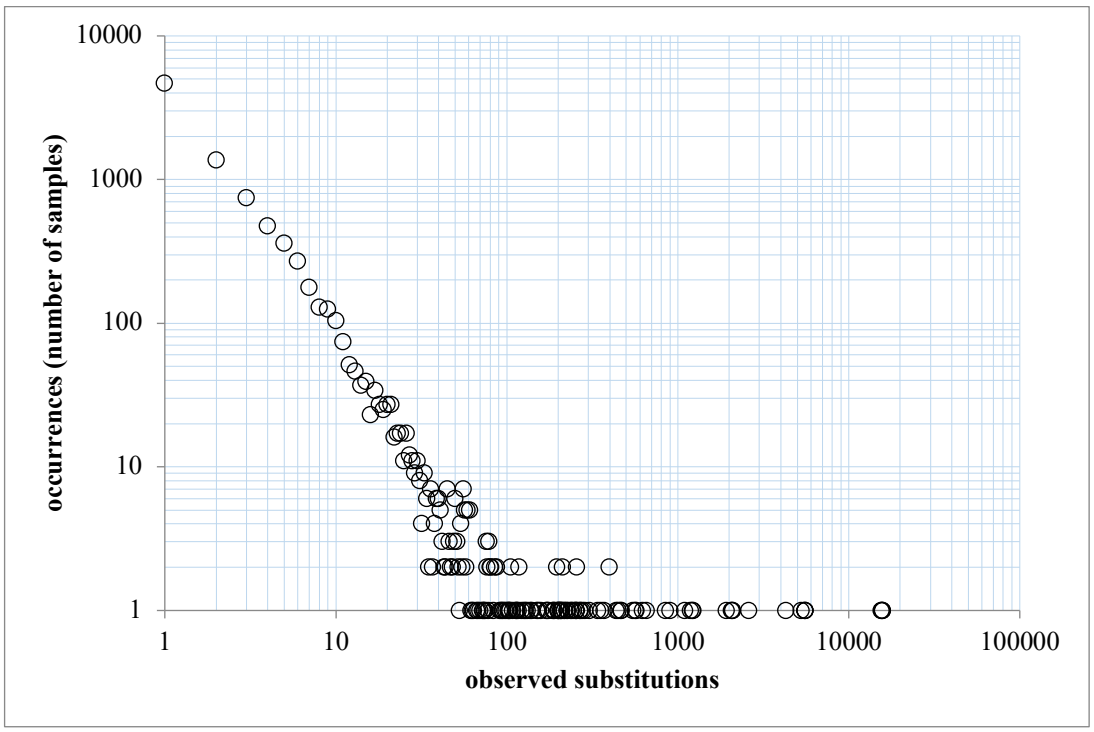


Figure S2. Substitution occurrences in samples. Most substitutions (15,750) were observed in just a single sample, whereas a single substitution was found in 4,683 samples.

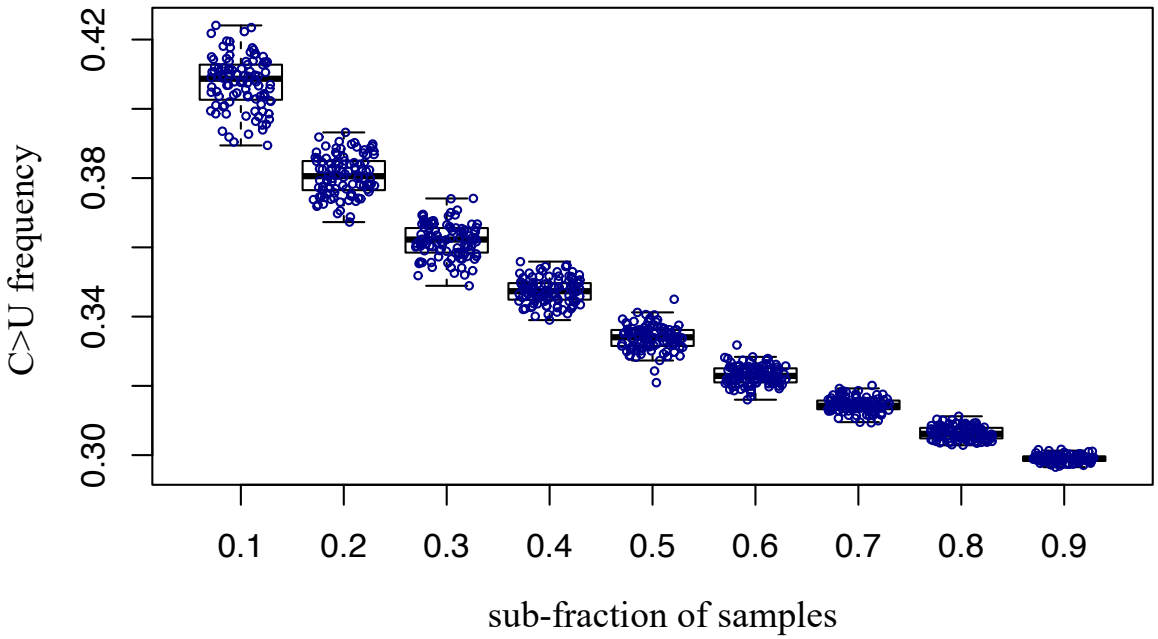


Figure S3. The inverse relationship of C>U frequency and the number of samples included. Down-sizing the number of samples leads to higher C>U frequency (0.1 on the x-axis means 10% of the total samples), suggesting that as more and more samples are included the prominent C>U changes happen as independent events at identical positions (hence only counted once in this analysis). As this saturation is expected to be lower for other, less abundant changes this will result in an apparent drop in the fraction of changes constituted by C>U. The blue dots are the individual simulated data sets, and the boxes denote their distribution.

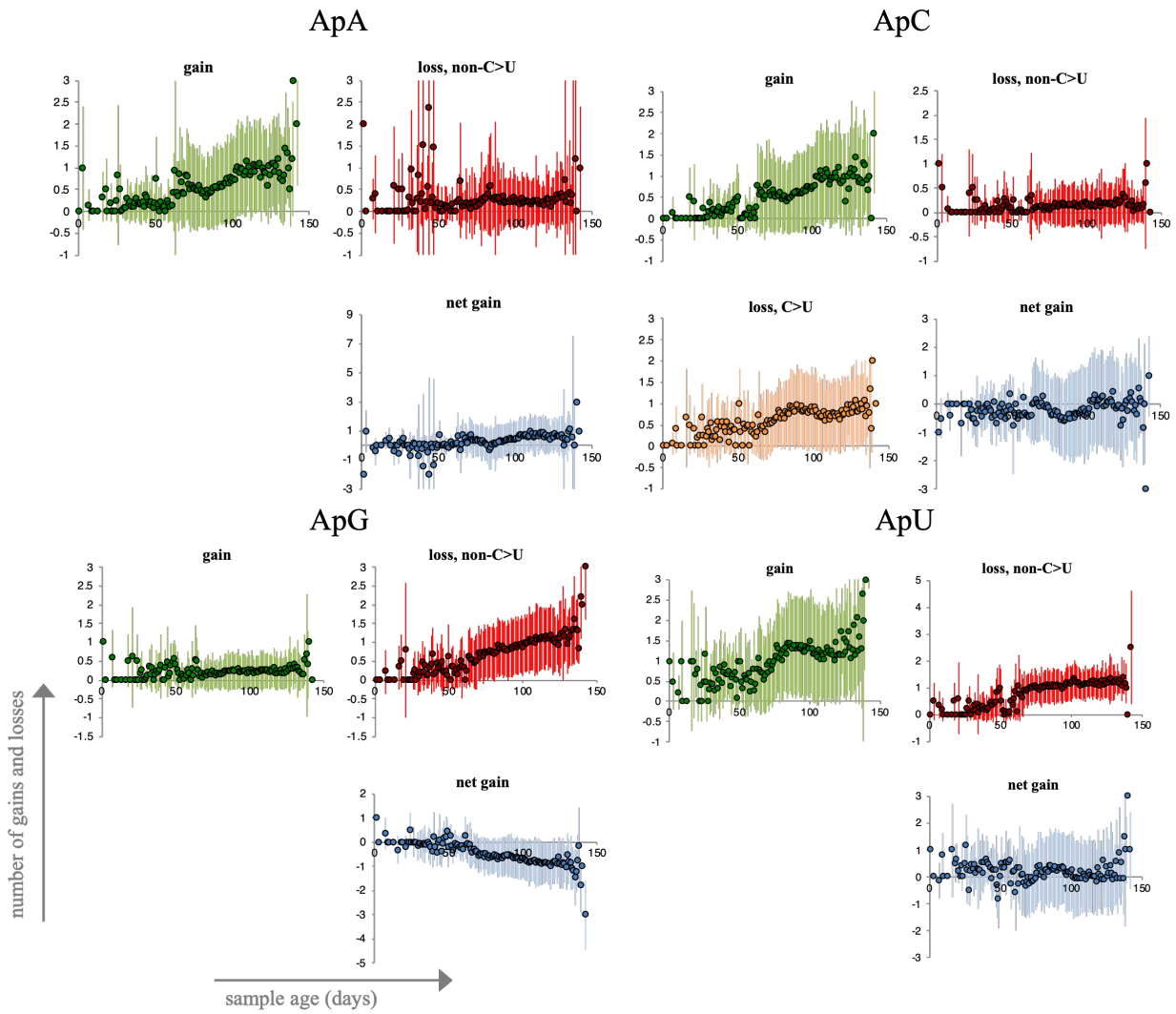
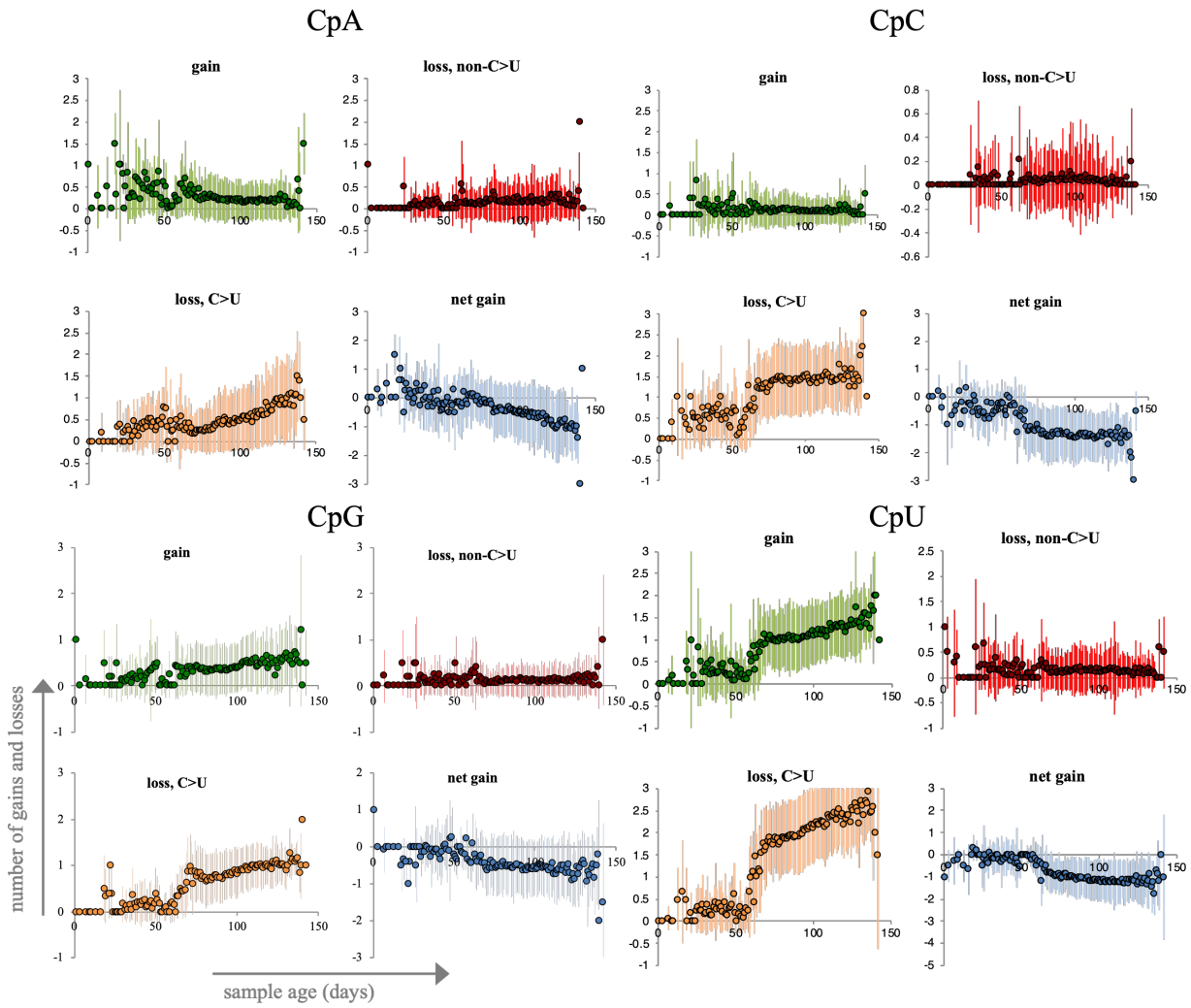


Figure S4. Number of dinucleotide gains and losses over time.



247

248 **Figure S4 (cont.).** Number of dinucleotide gains and losses over time.

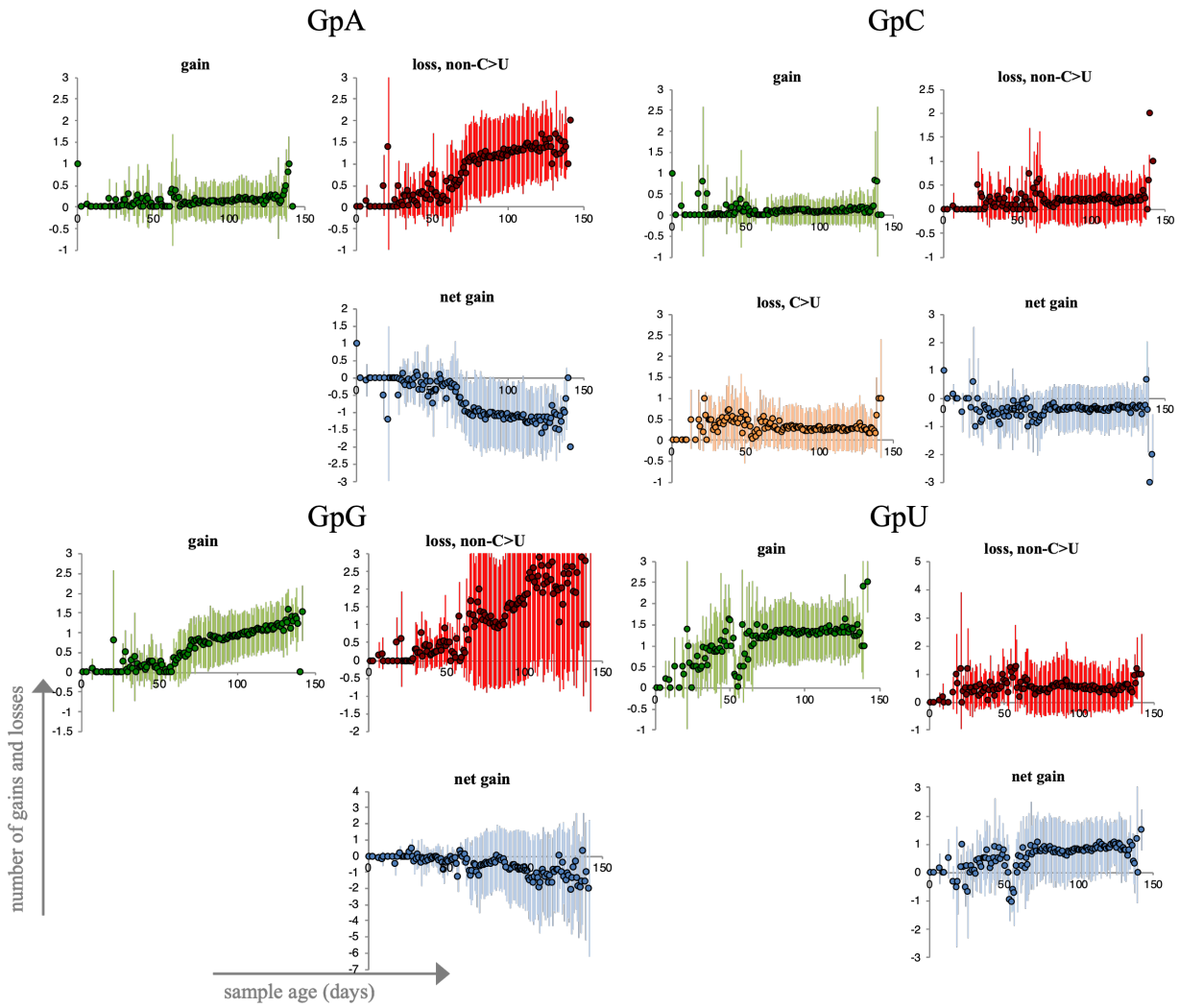


Figure S4 (cont.). Number of dinucleotide gains and losses over time.

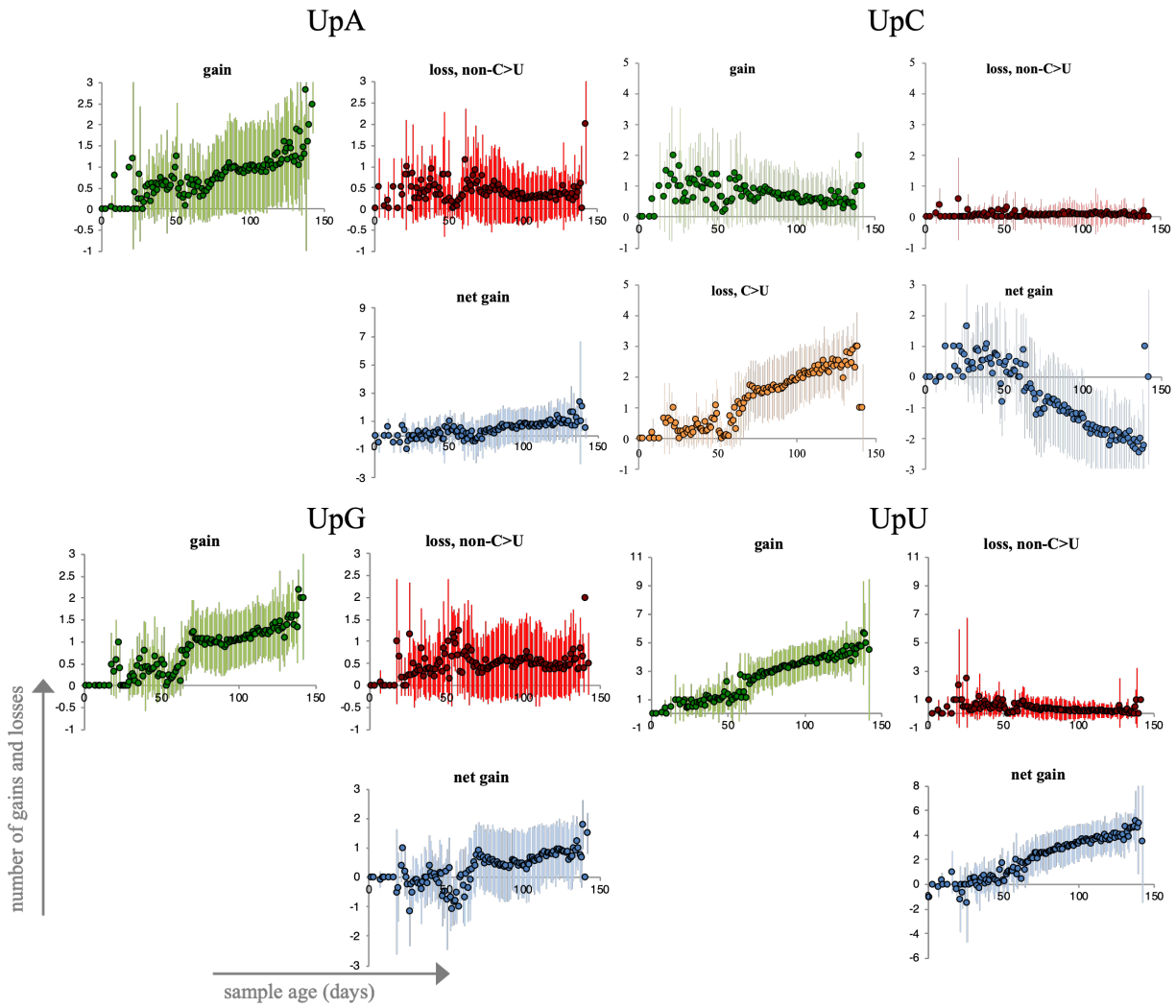


Figure S4 (cont.). Number of dinucleotide gains and losses over time.

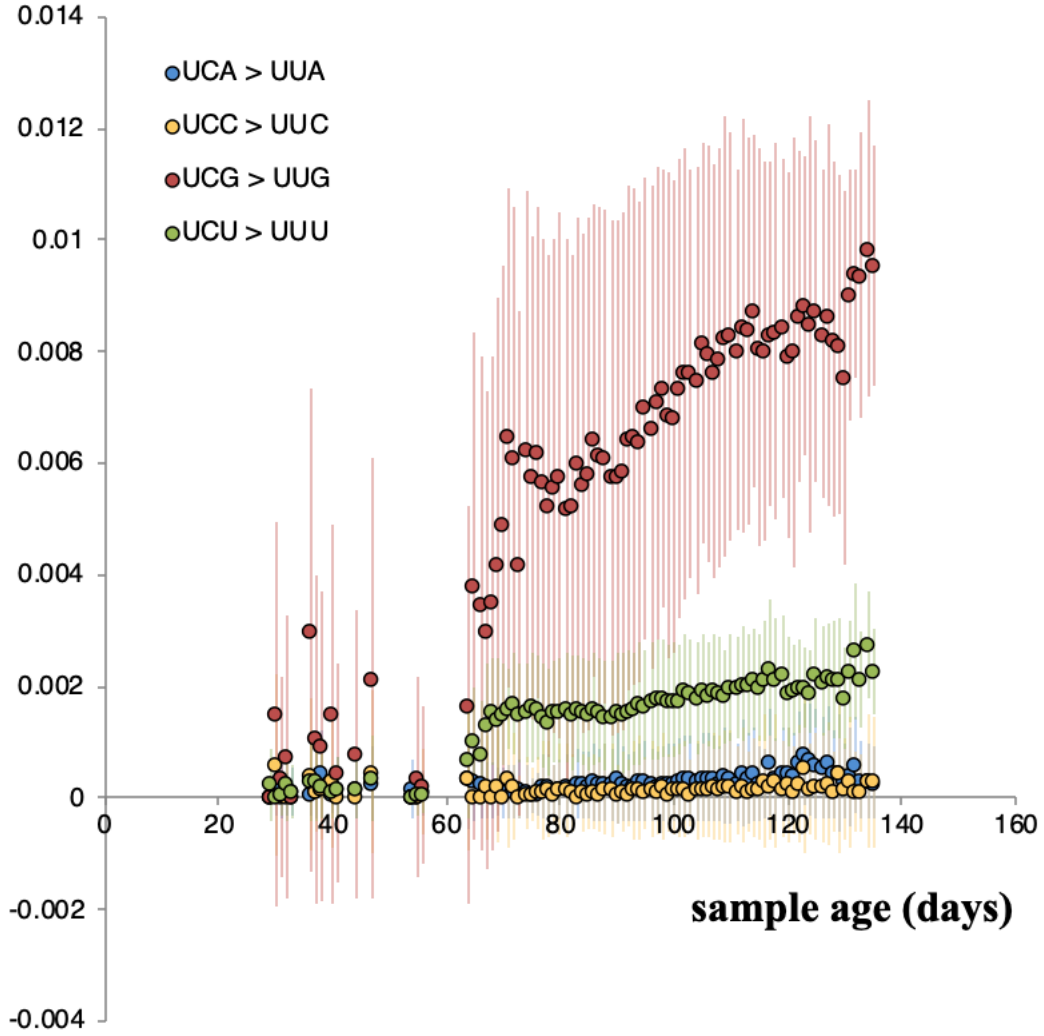
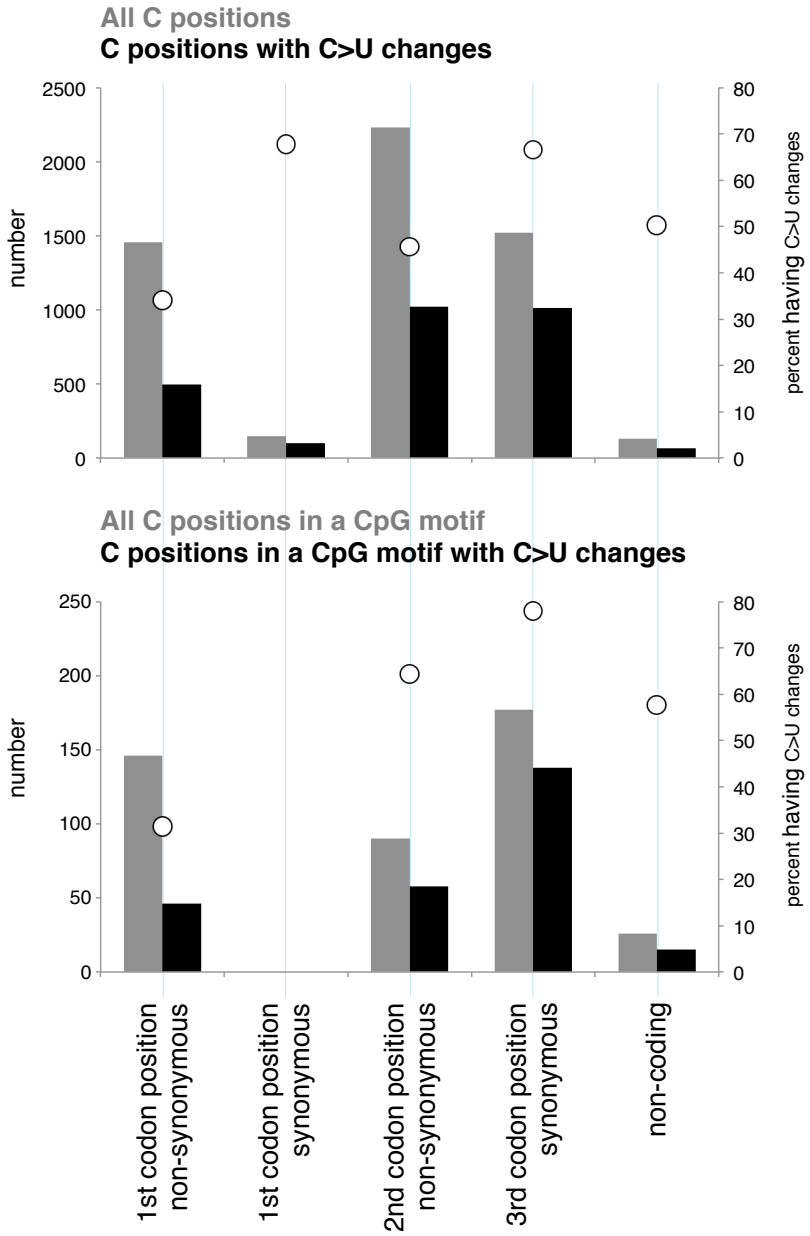


Figure S5. The fraction of UC[A/C/G/U] to UU[A/C/G/U] changes over time. The fraction of trinucleotide changes relative to the reference genome are shown as circles for each sampling day (x-axis). Error bars denote plus/minus one standard deviation.

257



258

259 **Figure S6.** Plots showing the number C positions (top) and CpG positions (bottom) in

260 the reference genome (black bars) and the number of these that have C>U changes

261 (grey bars). Numbers are shown for different codon positions as indicated below the

262 plots. CpG positions are defined as the position of the C in the motif, and positions with

263 no changes are defined as either synonymous or non-synonymous based on the effect

264 of a potential C>U change. The percentage of bases and motifs that have C>U changes

265 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.

Availability of Data

The data underlying this article are available in GISAID, at https://gisaid.org. The ID numbers of genomes used are provided in Table S1.

285 **References**

286 Atkinson NJ, Witteveldt J, Evans DJ, Simmonds P. 2014. The influence of CpG and
287 UpA dinucleotide frequencies on RNA virus replication and characterization of
288 the innate cellular pathways underlying virus attenuation and enhanced
289 replication. Nucleic Acids Res. doi:10.1093/nar/gku075.

290 Blanco-Melo D, Nilsson-Payant BE, Liu WC, Uhl S, Hoagland D, Møller R, Jordan TX,
291 Oishi K, Panis M, Sachs D, et al. 2020. Imbalanced Host Response to SARS-
292 CoV-2 Drives Development of COVID-19. Cell. doi:10.1016/j.cell.2020.04.026.

293 Booth DS, Szmidt-Middleton H, King N, Westbrook MJ, Young SL, Kuo A, Abedin M,
294 Chapman J, Fairclough SR, Hellsten U, et al. 2018. RStudio: Integrated
295 Development for R. Nature. doi:10.1108/eb003648.

296 Chemudupati M, Kenney AD, Bonifati S, Zani A, McMichael TM, Wu L, Yount JS. 2019.
297 From APOBEC to ZAP: Diverse mechanisms used by cellular restriction factors
298 to inhibit virus infections. Biochim Biophys Acta - Mol Cell Res.
299 doi:10.1016/j.bbamcr.2018.09.012.

300 Elbe S, Buckland-Merrett G. 2017. Data, disease and diplomacy: GISAID's innovative
301 contribution to global health. Glob Challenges. doi:10.1002/gch2.1018.

302 Fehr AR, Perlman S. 2015. Coronaviruses: An overview of their replication and
303 pathogenesis. In: Coronaviruses: Methods and Protocols.

304 Ficarelli M, Antzin-Anduetza I, Hugh-White R, Firth AE, Sertkaya H, Wilson H, Neil SJD,
305 Schulz R, Swanson CM. 2019. CpG Dinucleotides Inhibit HIV-1 Replication
306 through Zinc Finger Antiviral Protein (ZAP)-Dependent and -Independent

307 Mechanisms. J Virol. doi:10.1128/jvi.01337-19.
308 Di Gioacchino A, Sulc P, Komarova A V., Greenbaum BD, Monasson R, Cocco S. 2020.
309 The Heterogeneous Landscape and Early Evolution of Pathogen-Associated
310 CpG Dinucleotides in SARS-CoV-2. SSRN Electron J.
311 doi:10.2139/ssrn.3611280.
312 Giorgio S Di, Martignano F, Torcia MG, Mattiuz G, Conticello SG. 2020. Evidence for
313 host-dependent RNA editing in the transcriptome of SARS-CoV-2. Sci Adv.
314 doi:10.1126/sciadv.abb5813.
315 Greenbaum BD, Levine AJ, Bhanot G, Rabadan R. 2008. Patterns of evolution and host
316 gene mimicry in influenza and other RNA viruses. PLoS Pathog.
317 doi:10.1371/journal.ppat.1000079.
318 Gu H, Fan RLY, Wang D, Poon LLM. 2019. Dinucleotide evolutionary dynamics in
319 influenza A virus. Virus Evol. doi:10.1093/ve/vez038.
320 Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, Sagulenko P,
321 Bedford T, Neher RA. 2018. NextStrain: Real-time tracking of pathogen
322 evolution. Bioinformatics. doi:10.1093/bioinformatics/bty407.
323 Harris RS, Dudley JP. 2015. APOBECs and virus restriction. Virology.
324 doi:10.1016/j.virol.2015.03.012.
325 Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7:
326 Improvements in performance and usability. Mol Biol Evol.
327 doi:10.1093/molbev/mst010.
328 Kim EY, Lorenzo-Redondo R, Little SJ, Chung YS, Phalora PK, Maljkovic Berry I,

329 Archer J, Penugonda S, Fischer W, Richman DD, et al. 2014. Human APOBEC3
330 Induced Mutation of Human Immunodeficiency Virus Type-1 Contributes to
331 Adaptation and Evolution in Natural Infection. PLoS Pathog.
332 doi:10.1371/journal.ppat.1004281.
333 Kluge SF, Sauter D, Kirchhoff F. 2015. SnapShot: Antiviral Restriction Factors. Cell.
334 doi:10.1016/j.cell.2015.10.019.
335 Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF,
336 Hofacker IL. 2011. ViennaRNA Package 2.0. Algorithms Mol Biol.
337 doi:10.1186/1748-7188-6-26.
338 Luo X, Wang X, Gao Y, Zhu J, Liu S, Gao G, Gao P. 2020. Molecular Mechanism of
339 RNA Recognition by Zinc-Finger Antiviral Protein. Cell Rep.
340 doi:10.1016/j.celrep.2019.11.116.
341 McDaniel YZ, Wang D, Love RP, Adolph MB, Mohammadzadeh N, Chelico L, Mansky
342 LM. 2020. Deamination hotspots among APOBEC3 family members are defined
343 by both target site sequence context and ssDNA secondary structure. Nucleic
344 Acids Res. doi:10.1093/nar/gkz1164.
345 Meagher JL, Takata M, Gonçalves-Carneiro D, Keane SC, Rebendenne A, Ong H, Orr
346 VK, MacDonald MR, Stuckey JA, Bieniasz PD, et al. 2019. Structure of the zinc-
347 finger antiviral protein in complex with RNA reveals a mechanism for selective
348 targeting of CG-rich viral sequences. Proc Natl Acad Sci U S A.
349 doi:10.1073/pnas.1913232116.
350 Milewska A, Kindler E, Vkovski P, Zeglen S, Ochman M, Thiel V, Rajfur Z, Pyrc K. 2018.

351 APOBEC3-mediated restriction of RNA virus replication. Sci Rep.
352 doi:10.1038/s41598-018-24448-2.
353 Nchioua R, Kmiec D, Mueller J, Conzelmann C, Gross R, Swanson CM, Neil S, Stenger
354 S, Sauter D, Muench J, et al. 2020. The Zinc Finger Antiviral Protein restricts
355 SARS-CoV-2. bioRxiv. doi:10.1101/2020.06.04.134379.
356 Rice AM, Morales AC, Ho AT, Mordstein C, Mühlhausen S, Watson S, Cano L, Young
357 B, Kudla G, Hurst LD. 2020. Evidence for strong mutation bias towards, and
358 selection against, U content in SARS-CoV-2: implications for vaccine design. Mol
359 Biol Evol. doi:10.1093/molbev/msaa188.
360 Salemi M, Fitch WM, Ciccozzi M, Ruiz-Alvarez MJ, Rezza G, Lewis MJ. 2004. Severe
361 Acute Respiratory Syndrome Coronavirus Sequence Characteristics and
362 Evolutionary Rate Estimate from Maximum Likelihood Analysis. J Virol.
363 doi:10.1128/jvi.78.3.1602-1603.2004.
364 Salter JD, Smith HC. 2018. Modeling the Embrace of a Mutator: APOBEC Selection of
365 Nucleic Acid Ligands. Trends Biochem Sci. doi:10.1016/j.tibs.2018.04.013.
366 Sharma S, Patnaik SK, Thomas Taggart R, Kannisto ED, Enriquez SM, Gollnick P,
367 Baysal BE. 2015. APOBEC3A cytidine deaminase induces RNA editing in
368 monocytes and macrophages. Nat Commun. doi:10.1038/ncomms7881.
369 Sharma S, Wang J, Alqassim E, Portwood S, Cortes Gomez E, Maguire O, Basse PH,
370 Wang ES, Segal BH, Baysal BE. 2019. Mitochondrial hypoxic stress induces
371 widespread RNA editing by APOBEC3G in natural killer cells. Genome Biol.
372 doi:10.1186/s13059-019-1651-1.

373 Simmonds P. 2020. Rampant C->U hypermutation in the genomes of SARS-CoV-2 and
374 other coronaviruses - causes and consequences for their short and long
375 evolutionary trajectories. biorxiv. doi:10.1101/2020.05.01.072330.
376 Stern A, Yeh M Te, Zinger T, Smith M, Wright C, Ling G, Nielsen R, Macadam A,
377 Andino R. 2017. The Evolutionary Pathway to Virulence of an RNA Virus. Cell.
378 doi:10.1016/j.cell.2017.03.013.
379 Takata MA, Gonçalves-Carneiro D, Zang TM, Soll SJ, York A, Blanco-Melo D, Bieniasz
380 PD. 2017. CG dinucleotide suppression enables antiviral defence targeting non-
381 self RNA. Nature. doi:10.1038/nature24039.
382 Thielen BK, McNevin JP, McElrath MJ, Hunt BVS, Klein KC, Lingappa JR. 2010. Innate
383 immune signaling induces high levels of TC-specific deaminase activity in
384 primary monocyte-derived cells through expression of APOBEC3A isoforms. J
385 Biol Chem. doi:10.1074/jbc.M110.102822.
386 Wang Y, Mao J-M, Wang G-D, Luo Z-P, Yang L, Yao Q, Chen K-P. 2020. Human
387 SARS-CoV-2 has evolved to reduce CG dinucleotide in its open reading frames.
388 Sci Rep. 10(1):12331. doi:10.1038/s41598-020-69342-y.
389 https://doi.org/10.1038/s41598-020-69342-y.
390 Wei Y, Silke JR, Aris P, Xia X. 2020. Coronavirus genomes carry the signatures of their
391 habitats. bioRxiv. doi:10.1101/2020.06.13.149591.
392 Woo PCY, Wong BHL, Huang Y, Lau SKP, Yuen KY. 2007. Cytosine deamination and
393 selection of CpG suppressed clones are the two major independent biological
394 forces that shape codon usage bias in coronaviruses. Virology.

395 doi:10.1016/j.virol.2007.08.010.

396 Xia X. 2020. Extreme genomic CpG deficiency in SARS-CoV-2 and evasion of host

397 antiviral defense. Mol Biol Evol. doi:10.1093/molbev/msaa094.

398 Zhang J, Kang J, Liu M, Han B, Li L, He Y, Yi Z, Chen L. 2020 Jan 1. Multi-site co-

399 mutations and 5'UTR CpG immunity escape drive the evolution of SARS-CoV-2.

400 bioRxiv.:2020.07.21.213405. doi:10.1101/2020.07.21.213405.

401 http://biorxiv.org/content/early/2020/07/27/2020.07.21.213405.abstract.

402