AMERICAN SOCIETY FOR MICROBIOLOGY | mSphere®

RESEARCH ARTICLE
Ecological and Evolutionary Science

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

P. Simmonds[a]

[a]Nuffield Department of Medicine, University of Oxford, Oxford, United Kingdom

ABSTRACT The pandemic of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has motivated an intensive analysis of its molecular epidemiology following its worldwide spread. To understand the early evolutionary events following its emergence, a data set of 985 complete SARS-CoV-2 sequences was assembled. Variants showed a mean of 5.5 to 9.5 nucleotide differences from each other, consistent with a midrange coronavirus substitution rate of 3×10^{-4} substitutions/site/year. Almost one-half of sequence changes were C→U transitions, with an 8-fold base frequency normalized directional asymmetry between C→U and U→C substitutions. Elevated ratios were observed in other recently emerged coronaviruses (SARS-CoV, Middle East respiratory syndrome [MERS]-CoV), and decreasing ratios were observed in other human coronaviruses (HCoV-NL63, -OC43, -229E, and -HKU1) proportionate to their increasing divergence. C→U transitions underpinned almost one-half of the amino acid differences between SARS-CoV-2 variants and occurred preferentially in both 5′ U/A and 3′ U/A flanking sequence contexts comparable to favored motifs of human APOBEC3 proteins. Marked base asymmetries observed in nonpandemic human coronaviruses (U ≫ A > G ≫ C) and low G+C contents may represent long-term effects of prolonged C→U hypermutation in their hosts. The evidence that much of sequence change in SARS-CoV-2 and other coronaviruses may be driven by a host APOBEC-like editing process has profound implications for understanding their short- and long-term evolution. Repeated cycles of mutation and reversion in favored mutational hot spots and the widespread occurrence of amino acid changes with no adaptive value for the virus represent a quite different paradigm of virus sequence change from neutral and Darwinian evolutionary frameworks and are not incorporated by standard models used in molecular epidemiology investigations.

IMPORTANCE The wealth of accurately curated sequence data for severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), its long genome, and its low substitution rate provides a relatively blank canvas with which to investigate effects of mutational and editing processes imposed by the host cell. The finding that a large proportion of sequence change in SARS-CoV-2 in the initial months of the pandemic comprised C→U mutations in a host APOBEC-like context provides evidence for a potent host-driven antiviral editing mechanism against coronaviruses more often associated with antiretroviral defense. In evolutionary terms, the contribution of biased, convergent, and context-dependent mutations to sequence change in SARS-CoV-2 is substantial, and these processes are not incorporated by standard models used in molecular epidemiology investigations.

KEYWORDS APOBEC, COVID-19, SARS, SARS coronavirus 2, coronavirus, hypermutation, SARS-CoV-2

Citation Simmonds P. 2020. 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:e00408-20. https://doi.org/10.1128/mSphere.00408-20.

Editor Martin Schwemmle, University Medical Center Freiburg

Address correspondence to Peter.Simmonds@ndm.ox.ac.uk.

Received 1 May 2020
Accepted 11 June 2020
Published 24 June 2020

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) emerged late in 2019 in the Hubei province, China, as a cause of respiratory disease occasionally leading to acute respiratory distress syndrome and death (COVID-19) (1–4). Since the first reports in December 2019, infections with SARS-CoV-2 were reported from a rapidly increasing number of countries worldwide and led to its declaration as a pandemic by the World Health Organization in March 2020. To understand the origins and transmission dynamics of SARS-CoV-2, sequencing of SARS-CoV-2 directly from samples of infected individuals worldwide has been performed on an unprecedented scale. These efforts have generated many thousands of high-quality consensus sequences spanning the length of the genome and have defined a series of geographically defined clusters that recapitulate the early routes of international spread. However, as commented elsewhere (5), there is remarkably little virus diversity at this early stage of the pandemic, and analyses of its evolutionary dynamics remain at an early stage.

The relative infrequency of substitutions is the consequence of a much lower error rate on genome copying by the viral RNA polymerase of the larger nidovirales, including coronaviruses. This is achieved through the development of a proofreading capability through mismatch detection and excision by a viral encoded exonuclease, Nsp14-ExoN (6–8). Consequently, coronaviruses show a low substitution rate over time, typically in the range of 1.5×10^{-4} to 10×10^{-4} substitutions per site per year (SSY) (9–14). Applying a midrange estimate to the 3- to 5-month timescale of the SARS-CoV-2 pandemic indicates that epidemiologically unrelated strains might show around 6 to 10 nucleotide differences from each other over the 30,000-base length of their genomes.

In the present study, we have analyzed the nature of the sequence diversity generated within the SARS-CoV-2 virus populations revealed by current and ongoing virus sequencing studies. We obtained evidence for a preponderance of driven mutational events within the short evolutionary period following the zoonotic transmission of SARS-CoV-2 into humans. Sequence substitutions were characterized by a preponderance of cytidine-to-uridine (C→U) transitions. The possibility that the initial diversity within a viral population was largely host induced would have major implications for evolutionary reconstruction of SARS-CoV-2 variants in the current pandemic as well as in our understanding both of host antiviral pathways against coronaviruses and of the longer-term shaping effects on their genome composition.

RESULTS

Sequence changes in SARS-CoV-2. Four separate data sets of full-length (near-) complete genome sequences of SARS-CoV-2 collected from the start of the pandemic to those most recently deposited on 24 April 2020 were aligned and analyzed (accession numbers listed in Table S1A in the supplemental material). Each data set showed minimal levels of sequence divergence, with mean pairwise distances ranging from 5.5 to 9.5 nucleotide differences between each sequence. However, several aspects of the frequencies and sequence contexts of the observed changes were unexpected. First, the ratio of nonsynonymous (amino acid changing) to synonymous substitutions (*dN*/*dS*) was high, in the range of 0.57 to 0.73 among the different SARS-CoV-2 data sets. This contrasts with a much lower ratio (consistently below <0.22) in sequence data sets assembled for the other human coronaviruses (Table 1). Including a range of coronaviruses in the analysis, there was a consistent association between *dN*/*dS* ratios and the degree of sequence divergence (Fig. 1).

We next estimated the frequencies of individual transitions and transversions occurring during the short-term evolution of SARS-CoV-2. Sequence differences between each SARS-CoV-2 full-genome sequence and a majority rule consensus sequence generated for each of the four SARS-CoV-2 data sets were calculated. The directionality of sequence change underlying the observed substitutions was inferred by restricting the analysis to polymorphic sites with a minimal number of variable bases (typically singletons). In practice, because of the scarcity of substitutions, variability thresholds of 10%, 5%, 2%, and 1% yielded similar numbers and relative frequencies of each transition and transversion. Equivalent evidence for directionality was obtained through

TABLE 1 Coronavirus sequence data sets used for the study

Virus	No.	Length (bp)	MPD[a]	*dN/dS*[b]
Zoonotic coronaviruses				
SARS-CoV-2_Charite	115	29748	0.000187	0.728
SARS-CoV-2_Repl1	300	29409	0.000267	0.569
SARS-CoV-2_Repl2	300	29408	0.000306	0.630
SARS-CoV-2_Repl3	286	29404	0.000322	0.650
SARS-CoV-1-like (bat)	40	29480	0.048414	0.121
SARS-CoV-1	22	29443	0.000381	0.428
MERS-CoV	26	30043	0.005065	0.228
Other human and related coronaviruses				
OC43-human	178	30135	0.0081	0.219
OC43-bovine	113	30485	0.0104	0.139
HKU1-gt1	27	29613	0.0023	0.155
HKU1-gt2	12	29610	0.0121	0.096
NL63	61	27453	0.0086	0.131
229E-human	26	26846	0.0071	0.203
229E-camel	33	27051	0.002	0.167

[a]Mean nucleotide *p* distances between complete genome sequences.
[b]Frequency of nonsynonymous (*dN*) to synonymous (*dS*) *p* distances.

comparison of each sequence in the data set with the first outbreak sequence (MN908947; Wuhan-Hu-1), approximately ancestral to the currently circulating SARS-CoV-2 strains (data not shown). For the purposes of the analysis presented here, a consensus-based 5% threshold was used.

A listing of the sequence changes revealed a striking (approximately 4-fold) excess of sites where C→U substitutions occurred in SARS-CoV-2 sequences compared to the other three transitions (Fig. 2A). This excess was the more remarkable given there was an almost 2-fold greater number of U bases in the SARS-CoV-2 genome than Cs (32.1% compared to 18.4%, respectively). To formally analyze the excess of C→U transitions, we calculated an index of asymmetry (frequency [f][C→U]/f[U→C]) × (fU/fC) and compared this with degrees of sequence divergence and *dN/dS* ratios in SARS-CoV-2 and other coronavirus data sets (Fig. 2B and C). This comparison showed that the excess

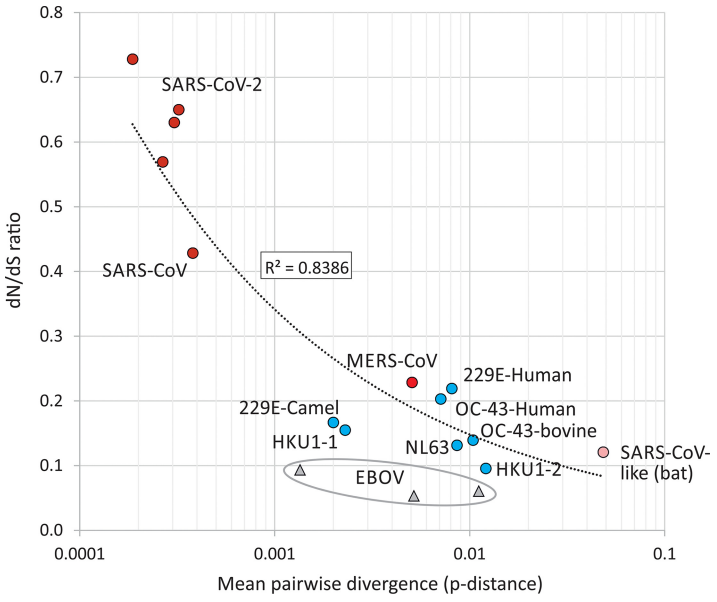

FIG 1 Association between sequence divergence and *dN/dS* ratio. A comparison of *dN/dS* ratios in recently emerged coronaviruses (red circles), other human coronaviruses and relatives infecting other species (blue circles), and a collection of bat sarbecoviruses (SARS-like) (pink circle). A power law line of best fit showed a significant correlation between divergence and *dN/dS* ratio ($P = 0.000006$). Sequences of the three data sets of EBOV control sequences were included (gray triangles).

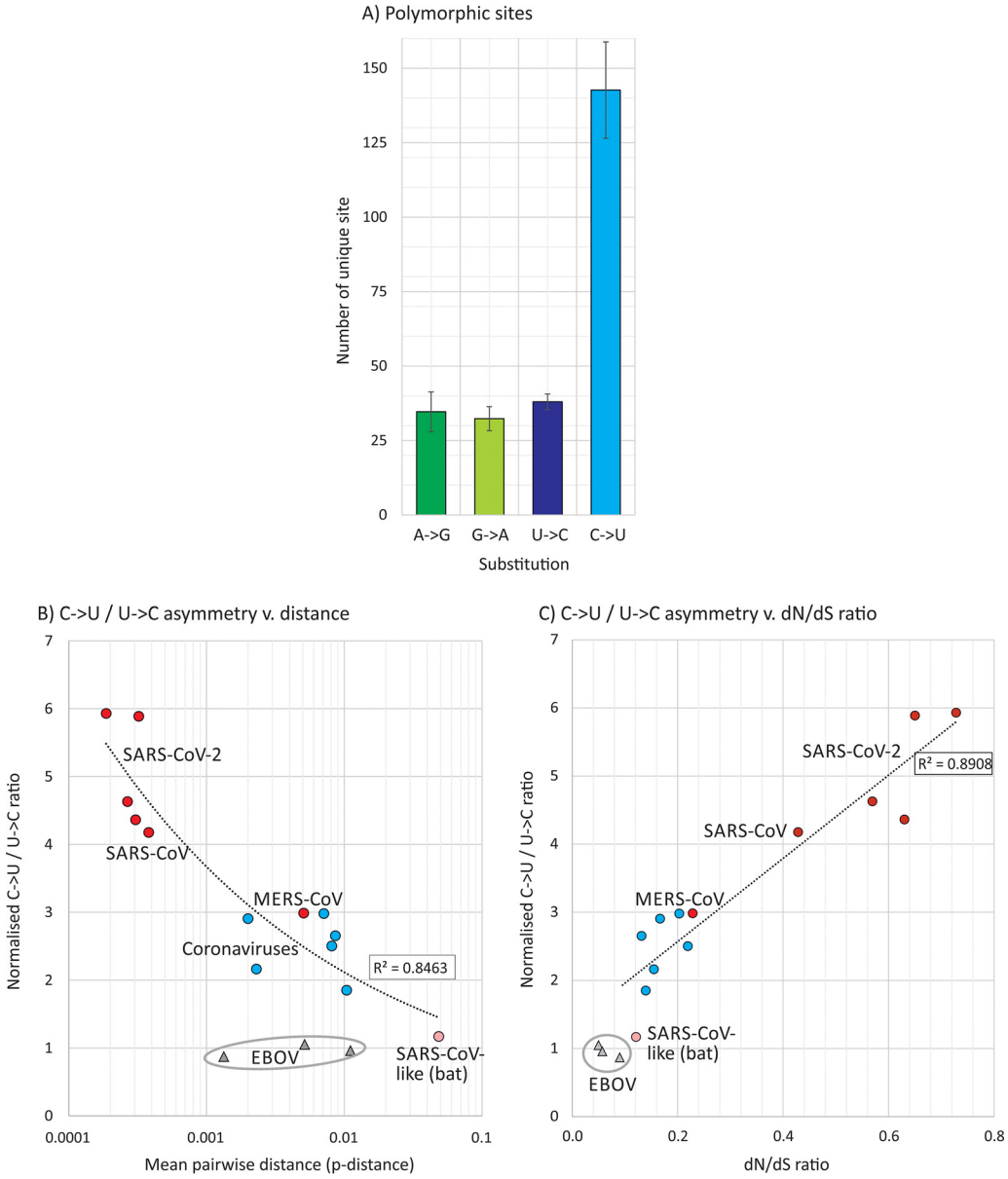

FIG 2 Association of excess C→U transitions with divergence. (A) Numbers of sites in the SARS-CoV-2 genome with each of the four transitions. Bar heights represent the means from the three sequence samples; error bars show one standard deviation (SD). (B) Relationship between sequence diversity and a normalized metric of asymmetry between the numbers of C→U and U→C transitions (where 1.0 is the expected number). Power law regression line was significant at a *P* value of <0.0001. (C) Association of *dN/dS* ratio with C→U/U→C asymmetry. The power law regression lines were significant at *P* values of 0.001 and 0.0004, respectively. Points are colored as in Fig. 1.

of C→U substitutions was most marked among very recently diverged sequences associated with the SARS-CoV-2 and SARS-CoV outbreaks and was reduced significantly in sequence data sets of the more divergent human coronaviruses (NL63, OC43, 229E, and OC43) as sequences accumulated substitutions.

A parallel analysis of the full-genome sequences of Ebola virus (EBOV) was performed to determine whether the compositional abnormalities observed in SARS-CoV-2 arose as artifacts of the next-generation sequencing (NGS) methods used to generate the data or indeed occurred in a different RNA virus with distinct entry, replication, and packaging strategies. Available sequences of EBOV on GenBank were divided into three groups, corresponding to those associated with the most recent outbreak in the Congo, in Sierra Leone, and elsewhere in West Africa in the 2014 outbreak, and finally a

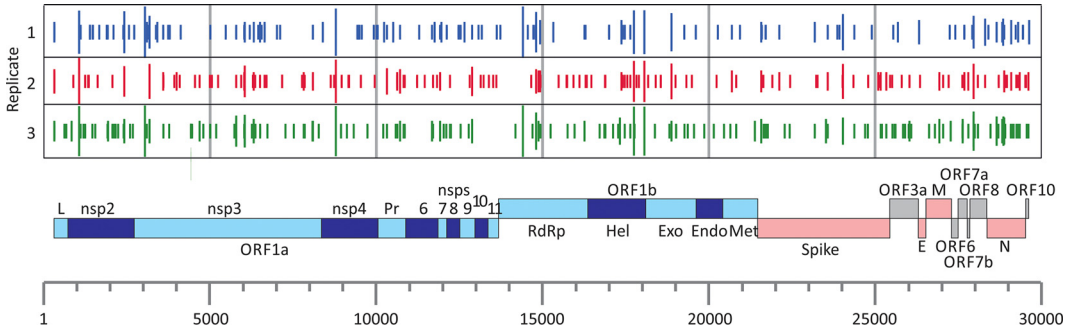

FIG 3 Positions of C→U transitions in the SARS-CoV-2 genome in each of the three replicate SARS-CoV-2 sequence data sets were matched to a genome diagram of SARS-CoV-2 (using the annotation from the prototype sequence MN908947). The numbers of transitions at each site are shown on a log scale, with the shortest bars indicating individual substitutions.

collection of older strains (see Table S1B). These showed mean levels of within-group sequence divergence of 0.1%, 1.1%, and 0.5%, respectively, spanning the range of divergences in the analyzed SARS-CoV-2 and other coronavirus data sets. In marked contrast to that of SARS-CoV-2, sequences consistently showed *dN/dS* ratios of <0.1 (Fig. 1) and no mutational asymmetry of C→U/U→C (Fig. 2B), irrespective of their sequence divergence.

C→U substitutions were scattered throughout the SARS-CoV-2 genome (Fig. 3). Long bars representing more polymorphic sites were frequently shared between replicate data sets, but unique substitutions (occurring once in the data set [short bars]) showed largely separate distributions. Substitutions were not focused toward any particular gene or intergenic region, although all three data sets showed marginally higher frequencies of substitutions in the N gene. A selection of sequences showing C→U changes in different genome regions was plotted in a phylogenetic tree containing sequences from the SARS-CoV-2 data set (Fig. 4). With the resolution possible in the tree generated from such a minimally divergent data set, many sequences with shared C→U changes were not monophyletic (e.g., those with substitutions at positions 5784, 10319, 21575, 28657, and 28887). This lack of grouping is consistent with multiple *de novo* occurrences of the same mutation in different SARS-CoV-2 lineages.

The abnormally high *dN/dS* ratios of 0.6 to 0.7 in SARS-CoV-2 sequences (Table 1; Fig. 1) indicated that around 50% of nucleotide substitutions would produce amino acid changes (if approximately 75% of nucleotide changes are nonsynonymous). On analysis of amino acid sequence changes, a remarkable 52% of nonsynonymous transitions in the SARS-CoV-2 sequence data set were the consequence of C→U transitions (Fig. 5), compared to 26%, 10%, and 7% for G→A, U→C, and A→C transitions, respectively. These ratios are comparable to those at all sites (Fig. 1) apart from the greater proportion of nonsynonymous G→A changes. Some variability might be expected given the potential fitness effects of specific amino acid changes and their likelihood of fixation. Notwithstanding this, the underlying mechanism that leads to C→U hypermutation therefore also drives much of the amino acid sequence diversity observed in SARS-CoV-2.

The context of cytidines within a sequence strongly influenced the likelihood of it mutating to a U (Fig. 6). The greatest numbers of mutations were observed if the upstream (5′) base was an A or U. There was also a similar approximately 4-fold increase in transitions if these bases were located on the downstream (3′) side. The effects of the 5′ and 3′ contexts were additive: C residues surrounded by an A or U at both 5′ and 3′ sides were 10-fold more likely to mutate than those flanked by C or G residues (mean of 31.9 transitions compared to 3.6). Splitting the data down into the 16 combinations of 5′ and 3′ contexts, a 5′ U far more potently restricted non-C→U substitutions than a 5′ A (see Fig. S1), while 5′ G or 5′ C almost eliminated substitutions irrespective of the 3′ context. No context created any substantial asymmetry in G→A compared to A→G transitions.

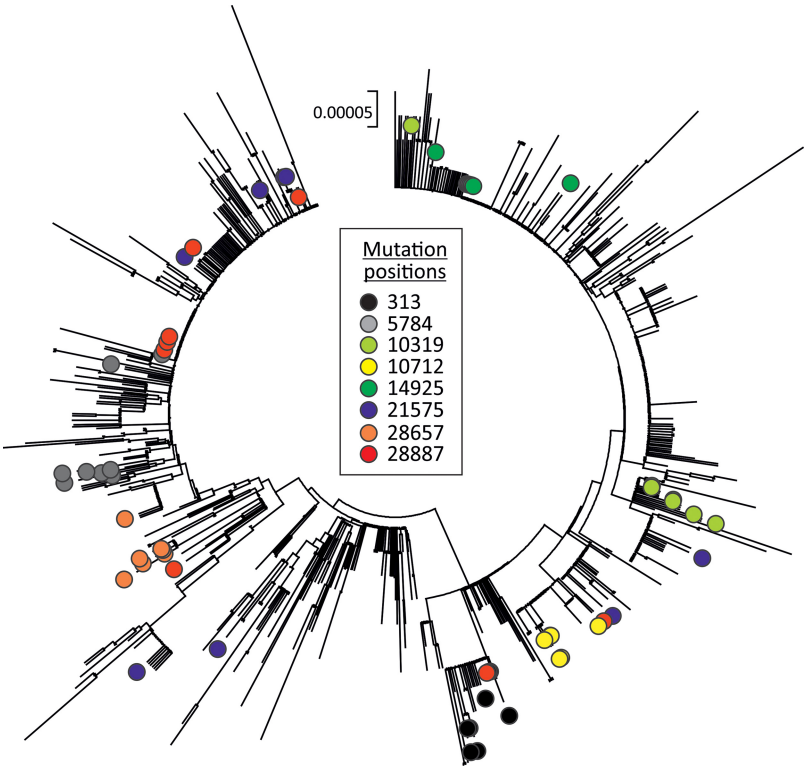

FIG 4 Phylogeny of SARS-CoV-2 and positions of sequences with C→U changes. A neighbor-joining tree of 865 SARS-CoV-2 complete genome sequences was constructed in MEGA6 (41). Labels show the position of sequences containing a selection of C→U transitions at the genome positions indicated in the key.

The G+C content of coronaviruses varied substantially between species, with highest frequencies in the recently emerged zoonotic coronaviruses (Middle East respiratory syndrome [MERS]-CoV, 41%; SARS-CoV, 41%; and SARS-CoV-2, 38%) and lowest in HKU1 (32%). Collectively, there was a significant relationship between C depletion and U enrichment with G+C content (Fig. 7). The difference in G+C content was indeed almost entirely attributable to changes in the frequencies of C and U bases: the 9% difference in G+C content between MERS-CoV and HKU1 arose primarily from the 20% to 13% reduction in frequencies of C. There was a comparable 8% increase in the frequency of U. Their combined effects left frequencies of G and A relatively unchanged. It has been proposed that the asymmetry in C and U frequencies may originate in part through the selective loss of CpG dinucleotides in the genome (15). To investigate this, the degree of suppression in SARS-CoV-2, other sarbecoviruses, and other coronaviruses was compared with representative sequences of each currently classified mammalian RNA virus species (excluding double-stranded RNA [dsRNA] viruses). Mammalian RNA viruses (Fig. 8, gray circles) demonstrate the previously described relationship between G+C content and CpG suppression (16). The data points for the separately labeled SARS-CoV-2, other SARS-like viruses in bats (sarbecoviruses; red), and the remainder of the coronaviruses (blue) and arteriviruses (green) overlap these values (Fig. 8). Overall, SARS-CoV-2 and other coronaviruses actually show less suppression of CpG for a given G+C content than is typical for other RNA viruses. SARS-CoV-2 and other coronaviruses are therefore not compositionally unusual by these metrics, providing no evidence that CpG suppression *per se* is associated with their mutational and compositional abnormalities.

DISCUSSION

The wealth of sequence data generated since the outset of the SARS-CoV-2 pandemic, the accuracy of the sequences obtained by a range of NGS technologies, and the

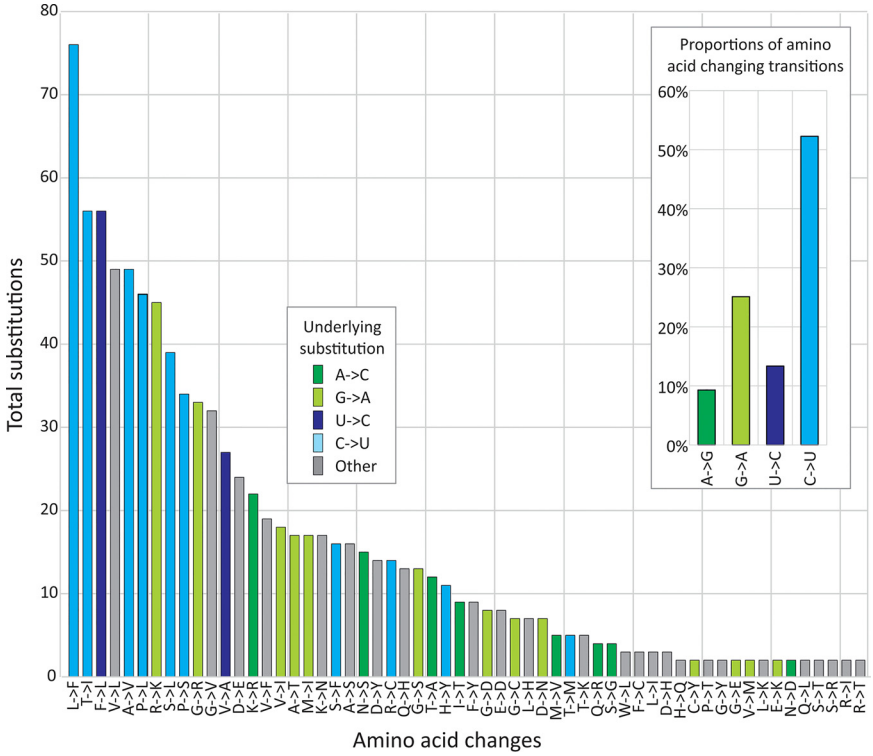

FIG 5 Amino acid changes induced by different nucleotide substitutions. Numbers of individual amino acids changes observed in the combined SARS-CoV-2 data set (864 sequences) at a 5% variability threshold. Bars are colored based on the underlying nucleotide changes. Inset graph shows the relative proportions of transitions leading to amino acid changes.

long genomes and very low substitution rate of coronaviruses provided a unique opportunity to investigate sequence diversification at very high resolution. The findings additionally provide insights into the mutational mechanisms and contexts where sequence changes occur. Third, it informs us about the longer-term evolution of viruses and potential effects of the cell in molding virus composition.

The mechanism of SARS-CoV-2 hypermutation. The most striking finding that emerged from the analysis of more than 1,000 SARS-CoV-2 genomes was the preponderance of C→U transitions compared to other substitutions in the initial 4 to 5 months of its evolution. These accounted for 38% to 42% of all changes in the four SARS-CoV-2 data sets. In seeking alternative nonbiological explanations for this observation, they cannot have arisen through misincorporation errors in the nex-generation sequence methods used to produce the data set because the analysis in the present study was restricted to consensus sequences. These are generally assembled from libraries that typically possess reasonable coverage and read depth; error frequencies of $<10^{-4}$ per site (17) would therefore improbably create a consensus change in a sequence library. There was, furthermore, no comparable increase in G→A mutations (Fig. 2), and the sequence context in which sequencing errors occur (a 5′ or 3′ C or G [17]) did not match the favored context for mutation observed in our data set (Fig. 6). Finally, the parallel analysis of a similarly large collection of consensus sequences of EBOV demonstrated very low *dN/dS* values and no excess of C→U mutations over U→C (Fig. 1 and 2). As EBOV and SARS-CoV-2 data were generated by comparable NGS methods, it is therefore extremely unlikely that the mutational and compositional abnormalities of SARS-CoV-2 described in the study were the results of methodological artifacts.

This asymmetric mutation, furthermore, cannot have arisen through a mutational effect of the coronavirus RNA-dependent RNA polymerase during virus replication. By definition, a coronavirus RNA genome descends from any other through an equal number of copies of the positive and negative strands—any tendency to misincorpo-

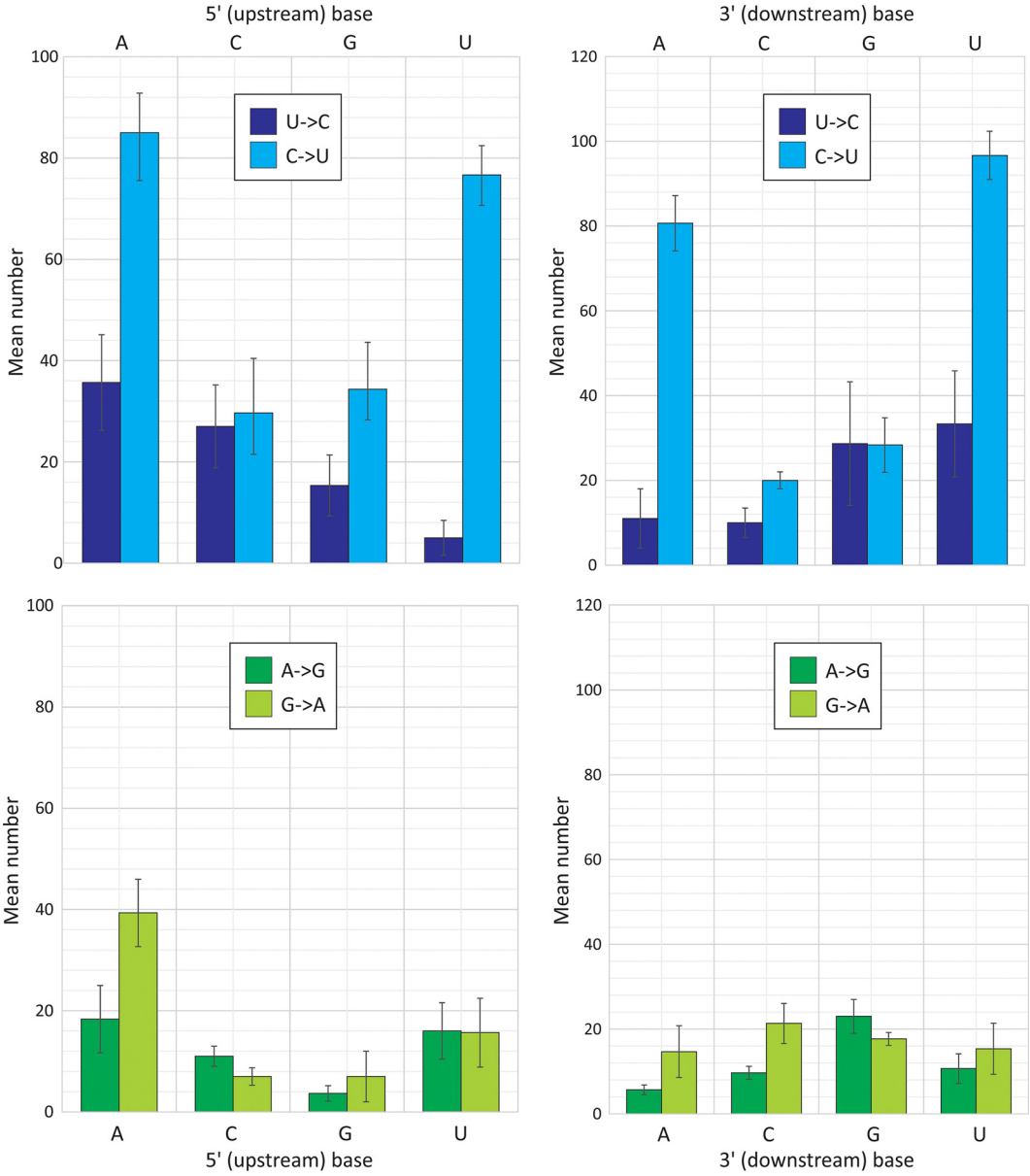

FIG 6 Influence of 5′ and 3′ base contexts on C↔U and G↔A transition frequencies. Totals of each transition in the SARS-CoV-2 sequence data set split into subtotals based on the identity of the 5′ (left) and 3′ (right) base. Bar heights represent the means from the three sequence samples; error bars show standard deviations. A further division into the 16 combinations of 5′ and 3′ base contexts is provided in Fig. S1 in the supplemental material.

rate a U instead of a C would be reflected in a parallel number of G→A mutations where this error occurred on the minus strand (or vice versa). As demonstrated, however, G→A mutations occurred at a much lower frequency than C→U mutations and similarly to A→G (Fig. 2A and 6).

The most cogent explanation for C→U hypermutation is the action of RNA editing processes within the infected cell. A well-characterized antiviral pathway involves the interferon-inducible isoform of adenosine deaminase acting on RNA type 1 (ADAR1) (18). This edits A to inosine in regions of viral double-stranded RNA, which is subsequently copied as a G. Irrespective of its widely demonstrated antiviral role in a range of typically minus-stranded RNA viruses, the mutations it creates do not match those observed in SARS-CoV-2 or other coronaviruses. First, ADAR1 targets dsRNA, and so editing effects tend to be symmetric with A→G substitutions being matched by U→C

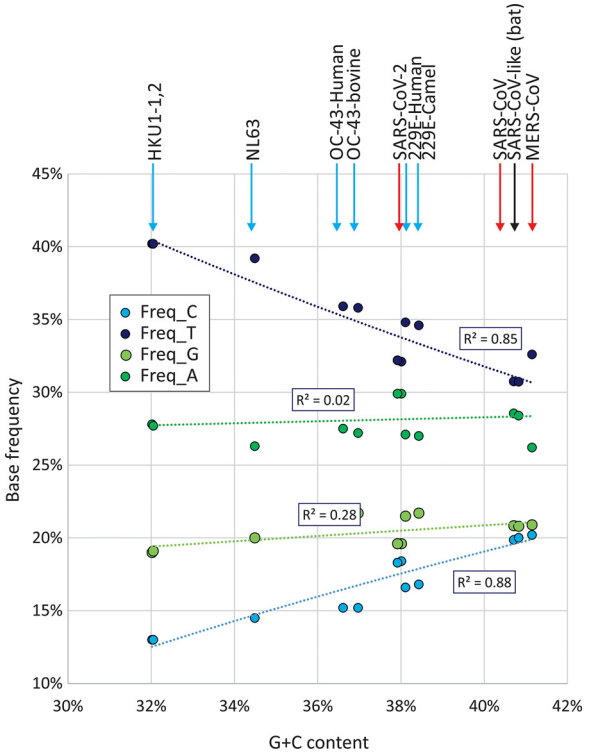

FIG 7 Base frequencies in different coronaviruses. Relationship between G+C content and frequencies of individual bases in coronaviruses. The associations between C depletion and U enrichment with G+C content were both significant by linear regression at $P = 5 \times 10^{-7}$ and $P = 5 \times 10^{-6}$, respectively. No significant associations were observed between G+C content and G ($P = 0.05$) or A ($P = 0.62$) frequencies. Arrows are color coded as for Fig. 1.

mutations. Second, the direction of mutation is wrong. The focus of the analysis in the present study was on infrequent or unique polymorphisms where ancestral and mutant bases can be inferred. The excess of C→U transitions is the opposite of those induced by ADAR1.

A second interferon-inducible pathway edits retroviral DNA during transcription and is strand specific; its typical antiviral activity is to mutate single-stranded proviral DNA formed after first-strand synthesis from genomic RNA (19–21). The deamination of Cs to Ts leads to the observed excess of G→A changes in the complementary positive-stranded RNA virus genome (22). This editing function is performed by members of the apolipoprotein B mRNA-editing enzyme, catalytic polypeptide-like (APOBEC) family, many of which possess defined antiviral functions against retroviruses, hepatitis B viruses, small DNA viruses, and intracellular mobile retroelements (reviewed in reference 23). The APOBEC3 gene family members that are primarily involved in antiviral defense show evidence of extensive positive selection and expansion over the course of mammalian evolution, particularly in the primate lineage. Humans possess 7 active antiviral proteins (A3A, A3B, A3C, A3D, A3F, A3G, and A3H) that contrast with the single A1 gene in rodents and a range of other mammals (24–26). Other mammals possess a diverse range of largely independently duplicated APOBEC3 genes, with four paralogs in cats, three in cows and sheep, six in horses, and often more than 10 in different bat species. However, their comparative activities and editing capabilities for different DNA and RNA substrates remain functionally largely unexplored.

While deamination of cytidines in single-stranded DNA sequences is a hallmark of APOBEC function, APOBECs show binding affinities for single-stranded RNA templates that may mediate antiviral functions. A3B and A3F have been shown to block retrotransposition of a LINE-1 transposon mRNA through a nondeamination pathway (27), potentially through binding to single-stranded RNA. Direct editing of HIV-1 RNA by the

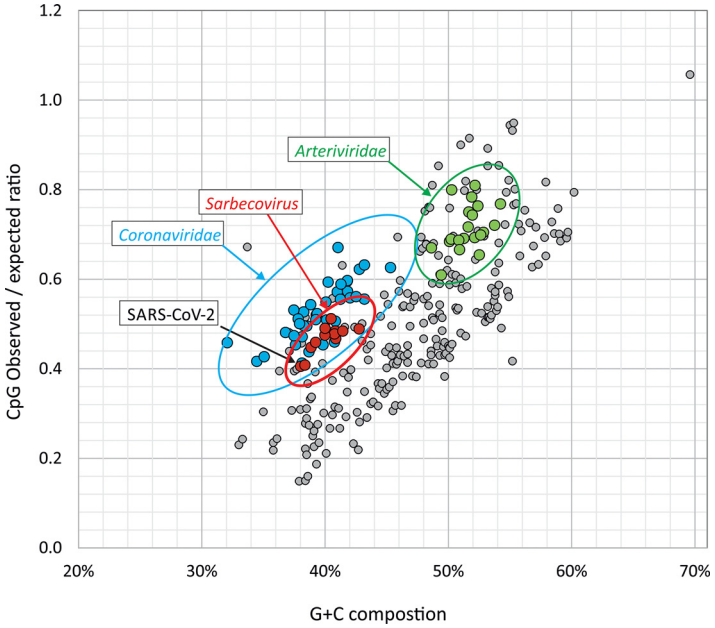

FIG 8 Suppression of CpG dinucleotides in SARS-CoV-2 and other coronaviruses. Comparison of CpG frequencies of SARS-CoV-2, other coronaviruses, and a set of other mammalian RNA viruses; each data point represents an individual currently classified species; accession numbers are listed in Table S1. CpG frequencies were expressed as the ratio of their observed frequency to the expected frequency based on their G+C content (*y* axis).

rat A1 APOBEC and the accumulation of C→U hypermutation verified that RNA could also be used as a substrate for deamination (28). This suggested to the authors at that time that APOBEC-mediated RNA editing was a potential antiviral activity mechanism against RNA viruses as well as retroviruses.

Since then, evidence supporting this conjecture has been difficult to obtain; the virus inhibitory effect of APOBECs against enterovirus A71, measles, mumps, and respiratory syncytial viruses were not shown to be associated with the development of virus mutations (29, 30). Similarly, A3C, A3F, or A3H, but not A3A, A3D, and A3G, were shown to inhibit the replication of the human coronavirus, HCoV-NL63, but their expression did not lead to *de novo* C→U (or G→A) mutations on virus passaging (31). On the other hand, it has been demonstrated that A3A and A3G possess potent RNA editing capability on mRNA expressed in hypoxic macrophages (32), natural killer cells (33), and transfected A3G-overexpressing HEK 293T cells (34). These latter findings verify that APOBECs possess RNA editing capabilities but do not provide any mechanistic context for the potential inhibition of RNA virus replication by this mechanism. Nevertheless, the pronounced asymmetry in C→U transitions in SARS-CoV-2 and the preferential substitution of Cs flanked by U and A bases on both 5′ and 3′ sides (Fig. 6) that broadly matches what is known about the favored contexts of A3A, A3F, and A3H (35) provides strong circumstantial grounds for suspecting a role of one or more APOBEC proteins in coronavirus mutagenesis.

The findings of C→U mutations at the consensus genome sequence level have also been observed within virus populations in a recent analysis of intrahost sequence diversity in lung-derived COVID-19 samples (36). Mutations showed the 5′ and 3′ A/U contexts as observed in the present study and were proposed by the authors as representing editing sites for APOBEC1. Intrahost diversity in this study was, however, dominated by minor populations generated from G→A and U→C substitutions; their symmetry and lack of 5′ or 3′ context led the authors to propose the editing effects of ADAR in viral dsRNA. These and other mutations such as A→U and U→A transversions mediated through an as-yet-uncharacterized mechanism hint at the complexity of host effects on virus sequence change. The combination of exceptionally long genomes

(≈30,000 bases), an otherwise low mutation rate, and the unprecedented size of the present data set of accurate minimally divergent SARS-CoV-2 sequences assembled postpandemic has enabled these mutational signatures to be so clearly observed. RNA editing may indeed represent a powerful antiviral mechanism with potentially lethal effects of even single mutations introduced into the genomic sequence. These make APOBEC-mediated anticoronaviral activity plausible in virological terms.

Evolutionary implications. The key findings in the study were the combined evidence for an APOBEC-like editing process driving initial sequences changes in SARS-CoV-2 and that the observed substitutions have not arisen through a typical pattern on random mutation and fixation that is assumed in evolutionary models. A specific problem for evolutionary reconstructions would be the existence of highly uneven substitution rates at different sites; APOBEC-mediated editing (and indeed the pattern of C→U transition in SARS-CoV-2 sequences) is strongly dependent on sequence context and, for at least two APOBECs, additionally influenced by their proximity to RNA secondary structure elements in the target sequence (32, 35). Sequence changes in SARS-CoV-2 and other coronavirus genomes may therefore be partially or largely restricted to a number of mutational hot spots that may promote convergent changes between otherwise genetically unlinked strains. As demonstrated in Fig. 4, these can conflict with relationships reconstructed from phylogenetically informative sites. Furthermore, the substitution rate reconstructed for SARS-CoV-2 and potentially other coronaviruses may represent an uncomfortable amalgam of both the accumulation of neutral changes and forced changes induced by APOBEC-like editing processes that may obscure temporal reconstructions. A recent analysis of SARS-CoV-2 genomes illustrates these problems (5); only a tiny fraction of variable sites (0.34%) were found to phylogenetically informative, while a high frequency of unresolved quartets demonstrates further the lack of phylogenetic signal in SARS-CoV-2 evolution reconstructions. The occurrence of multiple driven changes under host-induced selection is consistent with these cautionary observations.

The other important consequence of C→U hypermutation is that most of the amino acid sequence diversity observed in SARS-CoV-2 strains originates directly from forced mutations and therefore cannot be regarded in any way as adaptive for the virus (Fig. 5). An RNA editing mechanism of the type discussed above evidently places a huge mutational load on SARS-CoV-2 that may underpin the abnormally high *dN/dS* ratios recorded in SARS-CoV-2 and SARS-CoV sequence data sets (Fig. 1). It is likely that many or most amino acid changes are mildly deleterious and transient; repeated rounds of mutation at favored editing sites followed by reversion may therefore contribute to the large numbers of scattered substitutions in SARS-CoV-2 sequences that conflict with their phylogeny.

Finally, it is intriguing to speculate on the long-term effects of the C→U/U→C asymmetry and the extent to which this may contribute to the previously described compositional abnormalities of coronaviruses (15, 37). As described above in connection with mutation frequencies, the compositional asymmetries cannot directly arise through viral RdRp mutational biases, because any resulting base frequency differences would be symmetric (i.e., G ≈ C and A ≈ U). Instead, it appears that the observed imbalances in frequencies of complementary bases reflect the progressive depletion of C residues and accumulation of Us by the APOBEC-like mutational process on the genomic (+) strand of coronaviruses. Culminating in the compositionally highly abnormal HKU1 sequences (15), this appears to have driven the G+C content of coronaviruses as low as 32% while remarkably leaving G and A frequencies more or less unaltered (Fig. 7). Intriguingly, the bat-derived coronaviruses along with the recently zoonotically transferred viruses into humans show the least degree of compositional asymmetry.

The expansions in APOBEC gene numbers, extensive positive selection, and the consequent variability in APOBEC nucleic acid targeting (23) may indeed create distinct selection pressures on coronaviruses in different hosts. The immediate appearance of

C→U hypermutation in SARS-CoV-2 and SARS-CoV genomes in humans may therefore represent the initial effects of replication in a more hostile internal cellular environment than that found in what might be a better coadapted virus-tolerized immune system of a bat (38). Zoonotic origins are suspected for other human coronaviruses but at more remote times (39); perhaps they have taken their mutational and adaptive journeys already.

MATERIALS AND METHODS

SARS-CoV-2 and other coronavirus data sets. The 1,000 closest matched sequences to the prototype strain of SARS-CoV-2, NC_045512, were downloaded on 24 April 2020. Sequences with large internal gaps, ambiguous bases, and other markers of poor sequence quality were excluded, leaving a total of 865 sequences for analysis. These were divided into three data samples, corresponding to sequences 1 to 300, 301 to 600, and 601 to 865 (sequences listed in Table S1A in the supplemental material). An additional data set of 117 well-curated SARS-CoV-2 sequences was downloaded from Konsiliarlabor für Coronaviren (https://civnb.info/sequences/) on 13 April 2020 and represents a further independent sample set. A listing of further data sets of SARS-CoV, MERS-CoV, and other human coronaviruses is provided in Table 1. All available complete genome sequences of EBOV were downloaded from GenBank on 3 May 2020, of which 1,193 were used for mutational analysis after removal of incomplete, poor quality, and synthetic sequences (Table S1B).

Sequence analysis. Calculation of pairwise distances, nucleotide composition, and listing of sequence changes were performed using the SSE package version 1.4 (http://www.virus-evolution.org/Downloads/Software/) (40).

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

FIG S1, DOCX file, 0.3 MB.

TABLE S1, DOCX file, 0.1 MB.

TABLE S2, DOCX file, 0.1 MB.

ACKNOWLEDGMENT

The work was supported by a Wellcome Investigator Award Grant (WT103767MA).

REFERENCES

1. Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, Ren R, Leung KSM, Lau EHY, Wong JY, Xing X, Xiang N, Wu Y, Li C, Chen Q, Li D, Liu T, Zhao J, Liu M, Tu W, Chen C, Jin L, Yang R, Wang Q, Zhou S, Wang R, Liu H, Luo Y, Liu Y, Shao G, Li H, Tao Z, Yang Y, Deng Z, Liu B, Ma Z, Zhang Y, Shi G, Lam TTY, Wu JT, Gao GF, Cowling BJ, Yang B, Leung GM, Feng Z. 2020. Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. N Engl J Med 382:1199–1207. https://doi.org/10.1056/NEJMoa2001316.
2. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, Si HR, Zhu Y, Li B, Huang CL, Chen HD, Chen J, Luo Y, Guo H, Jiang RD, Liu MQ, Chen Y, Shen XR, Wang X, Zheng XS, Zhao K, Chen QJ, Deng F, Liu LL, Yan B, Zhan FX, Wang YY, Xiao GF, Shi ZL. 2020. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579:270–273. https://doi.org/10.1038/s41586-020-2012-7.
3. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, Zhao X, Huang B, Shi W, Lu R, Niu P, Zhan F, Ma X, Wang D, Xu W, Wu G, Gao GF, Tan W, China Novel Coronavirus Investigating and Research Team. 2020. A novel coronavirus from patients with pneumonia in China, 2019. N Engl J Med 382:727–733. https://doi.org/10.1056/NEJMoa2001017.
4. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, Hu Y, Tao ZW, Tian JH, Pei YY, Yuan ML, Zhang YL, Dai FH, Liu Y, Wang QM, Zheng JJ, Xu L, Holmes EC, Zhang YZ. 2020. A new coronavirus associated with human respiratory disease in China. Nature 579:265–269. https://doi.org/10.1038/s41586-020-2008-3.
5. Mavian C, Marini S, Manes C, Capua I, Prosperi M, Salemi M. 20 March 2020. Regaining perspective on SARS-CoV-2 molecular tracing and its implications. medRxiv https://www.medrxiv.org/content/10.1101/2020.03.16.20034470v1.
6. Smith EC, Blanc H, Surdel MC, Vignuzzi M, Denison MR. 2013. Coronaviruses lacking exoribonuclease activity are susceptible to lethal mutagenesis: evidence for proofreading and potential therapeutics. PLoS Pathog 9:e1003565. https://doi.org/10.1371/journal.ppat.1003565.
7. Eckerle LD, Becker MM, Halpin RA, Li K, Venter E, Lu X, Scherbakova S, Graham RL, Baric RS, Stockwell TB, Spiro DJ, Denison MR. 2010. Infidelity of SARS-CoV Nsp14-exonuclease mutant virus replication is revealed by complete genome sequencing. PLoS Pathog 6:e1000896. https://doi.org/10.1371/journal.ppat.1000896.
8. Eckerle LD, Lu X, Sperry SM, Choi L, Denison MR. 2007. High fidelity of murine hepatitis virus replication is decreased in nsp14 exoribonuclease mutants. J Virol 81:12135–12144. https://doi.org/10.1128/JVI.01296-07.
9. Salemi M, Fitch WM, Ciccozzi M, Ruiz-Alvarez MJ, Rezza G, Lewis MJ. 2004. Severe acute respiratory syndrome coronavirus sequence characteristics and evolutionary rate estimate from maximum likelihood analysis. J Virol 78:1602–1603. https://doi.org/10.1128/jvi.78.3.1602-1603.2004.
10. Sanchez CM, Gebauer F, Sune C, Mendez A, Dopazo J, Enjuanes L. 1992. Genetic evolution and tropism of transmissible gastroenteritis coronaviruses. Virology 190:92–105. https://doi.org/10.1016/0042-6822(92)91195-z.
11. Vijgen L, Lemey P, Keyaerts E, Van Ranst M. 2005. Genetic variability of human respiratory coronavirus OC43. J Virol 79:3223–3224. https://doi.org/10.1128/JVI.79.5.3223-3225.2005.
12. Fu X, Fang B, Liu Y, Cai M, Jun J, Ma J, Bu D, Wang L, Zhou P, Wang H, Zhang G. 2018. Newly emerged porcine enteric alphacoronavirus in southern China: identification, origin and evolutionary history analysis. Infect Genet Evol 62:179–187. https://doi.org/10.1016/j.meegid.2018.04.031.
13. Homwong N, Jarvis MC, Lam HC, Diaz A, Rovira A, Nelson M, Marthaler D. 2016. Characterization and evolution of porcine deltacoronavirus in the United States. Prev Vet Med 123:168–174. https://doi.org/10.1016/j.prevetmed.2015.11.001.
14. Cotten M, Watson SJ, Zumla AI, Makhdoom HQ, Palser AL, Ong SH, Al Rabeeah AA, Alhakeem RF, Assiri A, Al-Tawfiq JA, Albarrak A, Barry M, Shibl A, Alrabiah FA, Hajjar S, Balkhy HH, Flemban H, Rambaut A, Kellam P, Memish ZA. 2014. Spread, circulation, and evolution of the Middle East respiratory syndrome coronavirus. mBio 5:e01062-13. https://doi.org/10.1128/mBio.01062-13.
15. Woo PC, Wong BH, Huang Y, Lau SK, Yuen KY. 2007. Cytosine deamina-

tion and selection of CpG suppressed clones are the two major independent biological forces that shape codon usage bias in coronaviruses. Virology 369:431–442. https://doi.org/10.1016/j.virol.2007.08.010.
16. Simmonds P, Xia W, Baillie JK, McKinnon K. 2013. Modelling mutational and selection pressures on dinucleotides in eukaryotic phyla–selection against CpG and UpA in cytoplasmically expressed RNA and in RNA viruses. BMC Genomics 14:610. https://doi.org/10.1186/1471-2164-14-610.
17. Ma X, Shao Y, Tian L, Flasch DA, Mulder HL, Edmonson MN, Liu Y, Chen X, Newman S, Nakitandwe J, Li Y, Li B, Shen S, Wang Z, Shurtleff S, Robison LL, Levy S, Easton J, Zhang J. 2019. Analysis of error profiles in deep next-generation sequencing data. Genome Biol 20:50. https://doi.org/10.1186/s13059-019-1659-6.
18. Samuel CE. 2011. Adenosine deaminases acting on RNA (ADARs) are both antiviral and proviral. Virology 411:180–193. https://doi.org/10.1016/j.virol.2010.12.004.
19. Zhang H, Yang B, Pomerantz RJ, Zhang C, Arunachalam SC, Gao L. 2003. The cytidine deaminase CEM15 induces hypermutation in newly synthesized HIV-1 DNA. Nature 424:94–98. https://doi.org/10.1038/nature01707.
20. Mangeat B, Turelli P, Caron G, Friedli M, Perrin L, Trono D. 2003. Broad antiretroviral defence by human APOBEC3G through lethal editing of nascent reverse transcripts. Nature 424:99–103. https://doi.org/10.1038/nature01709.
21. Harris RS, Bishop KN, Sheehy AM, Craig HM, Petersen-Mahrt SK, Watt IN, Neuberger MS, Malim MH. 2003. DNA deamination mediates innate immunity to retroviral infection. Cell 113:803–809. https://doi.org/10.1016/s0092-8674(03)00423-9.
22. Vartanian JP, Meyerhans A, Sala M, Wain-Hobson S. 1994. G→A hypermutation of the human immunodeficiency virus type 1 genome: evidence for dCTP pool imbalance during reverse transcription. Proc Natl Acad Sci U S A 91:3092–3096. https://doi.org/10.1073/pnas.91.8.3092.
23. Harris RS, Dudley JP. 2015. APOBECs and virus restriction. Virology 479–480:131–145. https://doi.org/10.1016/j.virol.2015.03.012.
24. Münk C, Willemsen A, Bravo IG. 2012. An ancient history of gene duplications, fusions and losses in the evolution of APOBEC3 mutators in mammals. BMC Evol Biol 12:71. https://doi.org/10.1186/1471-2148-12-71.
25. Henry M, Terzian C, Peeters M, Wain-Hobson S, Vartanian JP. 2012. Evolution of the primate APOBEC3A cytidine deaminase gene and identification of related coding regions. PLoS One 7:e30036. https://doi.org/10.1371/journal.pone.0030036.
26. Sawyer SL, Emerman M, Malik HS. 2004. Ancient adaptive evolution of the primate antiviral DNA-editing enzyme APOBEC3G. PLoS Biol 2:E275. https://doi.org/10.1371/journal.pbio.0020275.
27. Stenglein MD, Harris RS. 2006. APOBEC3B and APOBEC3F inhibit L1 retrotransposition by a DNA deamination-independent mechanism. J Biol Chem 281:16837–16841. https://doi.org/10.1074/jbc.M602367200.
28. Bishop KN, Holmes RK, Sheehy AM, Malim MH. 2004. APOBEC-mediated editing of viral RNA. Science 305:645. https://doi.org/10.1126/science.1100658.
29. Fehrholz M, Kendl S, Prifert C, Weissbrich B, Lemon K, Rennick L, Duprex PW, Rima BK, Koning FA, Holmes RK, Malim MH, Schneider-Schaulies J. 2012. The innate antiviral factor APOBEC3G targets replication of measles, mumps and respiratory syncytial viruses. J Gen Virol 93:565–576. https://doi.org/10.1099/vir.0.038919-0.
30. Wang H, Zhong M, Li Y, Li K, Wu S, Guo T, Cen S, Jiang J, Li Z, Li Y. 2019. APOBEC3G is a restriction factor of EV71 and mediator of IMB-Z antiviral activity. Antiviral Res 165:23–33. https://doi.org/10.1016/j.antiviral.2019.03.005.
31. Milewska A, Kindler E, Vkovski P, Zeglen S, Ochman M, Thiel V, Rajfur Z, Pyrc K. 2018. APOBEC3-mediated restriction of RNA virus replication. Sci Rep 8:5960. https://doi.org/10.1038/s41598-018-24448-2.
32. Sharma S, Patnaik SK, Taggart RT, Kannisto ED, Enriquez SM, Gollnick P, Baysal BE. 2015. APOBEC3A cytidine deaminase induces RNA editing in monocytes and macrophages. Nat Commun 6:6881. https://doi.org/10.1038/ncomms7881.
33. Sharma S, Wang J, Alqassim E, Portwood S, Cortes Gomez E, Maguire O, Basse PH, Wang ES, Segal BH, Baysal BE. 2019. Mitochondrial hypoxic stress induces widespread RNA editing by APOBEC3G in natural killer cells. Genome Biol 20:37. https://doi.org/10.1186/s13059-019-1651-1.
34. Sharma S, Patnaik SK, Taggart RT, Baysal BE. 2016. The double-domain cytidine deaminase APOBEC3G is a cellular site-specific RNA editing enzyme. Sci Rep 6:39100. https://doi.org/10.1038/srep39100.
35. McDaniel YZ, Wang D, Love RP, Adolph MB, Mohammadzadeh N, Chelico L, Mansky LM. 2020. Deamination hotspots among APOBEC3 family members are defined by both target site sequence context and ssDNA secondary structure. Nucleic Acids Res 20:1353–1371. https://doi.org/10.1093/nar/gkz1164.
36. Di Giorgio S, Martignano F, Torcia MG, Mattiuz G, Conticello SG. 18 May 2020. Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2. Sci Adv https://doi.org/10.1126/sciadv.abb5813.
37. Berkhout B, van Hemert F. 2015. On the biased nucleotide composition of the human coronavirus RNA genome. Virus Res 202:41–47. https://doi.org/10.1016/j.virusres.2014.11.031.
38. Baker ML, Schountz T, Wang LF. 2013. Antiviral immune responses of bats: a review. Zoonoses Public Health 60:104–116. https://doi.org/10.1111/j.1863-2378.2012.01528.x.
39. Corman VM, Muth D, Niemeyer D, Drosten C. 2018. Hosts and sources of endemic human coronaviruses. Adv Virus Res 100:163–188. https://doi.org/10.1016/bs.aivir.2018.01.001.
40. Simmonds P. 2012. SSE: a nucleotide and amino acid sequence analysis platform. BMC Res Notes 5:50. https://doi.org/10.1186/1756-0500-5-50.
41. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol Biol Evol 30:2725–2729. https://doi.org/10.1093/molbev/mst197.