AMERICAN SOCIETY FOR MICROBIOLOGY | mBio

Check for updates

| Virology | Research Article

C→U transition biases in SARS-CoV-2: still rampant 4 years from the start of the COVID-19 pandemic

Peter Simmonds[1]

AUTHOR AFFILIATION See affiliation list on p. 18.

ABSTRACT The evolution of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in the pandemic and post-pandemic periods has been characterized by rapid adaptive changes that confer immune escape and enhanced human-to-human transmissibility. Sequence change is additionally marked by an excess number of C→U transitions suggested as being due to host-mediated genome editing. To investigate how these influence the evolutionary trajectory of SARS-CoV-2, 2,000 high-quality, coding complete genome sequences of SARS-CoV-2 variants collected pre-September 2020 and from each subsequently appearing alpha, delta, BA.1, BA.2, BA.5, XBB, EG, HK, and JN.1 lineages were downloaded from NCBI Virus in April 2024. C→U transitions were the most common substitution during the diversification of SARS-CoV-2 lineages over the 4-year observation period. A net loss of C bases and accumulation of U's occurred at a constant rate of approximately 0.2%–0.25%/decade. C→U transitions occurred in over a quarter of all sites with a C (26.5%; range 20.0%–37.2%) around five times more than observed for the other transitions (5.3%–6.8%). In contrast to an approximately random distribution of other transitions across the genome, most C→U substitutions occurred at statistically preferred sites in each lineage. However, only the most C→U polymorphic sites showed evidence for a preferred 5′U context previously associated with APOBEC 3A editing. There was a similarly weak preference for unpaired bases suggesting much less stringent targeting of RNA than mediated by A3 deaminases in DNA editing. Future functional studies are required to determine editing preferences, impacts on replication fitness *in vivo* of SARS-CoV-2 and other RNA viruses, and impact on host tropism.

IMPORTANCE Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in the pandemic and post-pandemic periods has shown a remarkable capacity to adapt and evade human immune responses and increase its human-to-human transmissibility. The genome of SARS-CoV-2 is also increasingly scarred by the effects of multiple C→U mutations from host genome editing as a cellular defense mechanism akin to restriction factors for retroviruses. Through the analysis of large data sets of SARS-CoV-2 isolate sequences collected throughout the pandemic period and beyond, we show that C→U transitions have driven a base compositional change over time amounting to a net loss of C bases and accumulation of U's at a rate of approximately 0.2%–0.25%/decade. Most C→U substitutions occurred in the absence of the preferred upstream-base context or targeting of unpaired RNA bases previously associated with the host RNA editing protein, APOBEC 3A. The analyses provide a series of testable hypotheses that can be experimentally investigated in the future.

KEYWORDS HCV RNA secondary structure, evolution, APOBEC3, innate immunity

The COVID-19 pandemic between 2020 and 2023 followed the emergence of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in 2019. SARS-CoV-2 is a member of the genus *Betacoronavirus* in the family *Coronaviridae* (1) and likely originates

Editor Dimitrios Paraskevis, Medical School, National and Kapodistrian University of Athens, Athens, Greece

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

The authors declare no conflict of interest.

Received 19 August 2024
Accepted 24 September 2024
Published 30 October 2024

from a zoonotic spillover of a variant of sarbecoviruses widely distributed in *Rhinolophus* (horseshoe) bats in Southeast Asia (2). Considerable insight into the evolution of SARS-CoV-2 over the pandemic period and beyond has been obtained through the large international sequencing effort, with over 16 million complete genome sequences cataloged to date. SARS-CoV-2 rapidly diversified into several genetically distinct clades (Fig. 1) and multiple cycles of emergence and extinction of variants of concerns (VoC) (3). Strains similar to the prototype Wuhan-1 strain circulated in the first year of the pandemic, although a variant with the D614G mutation in the spike gene emerged in April and rapidly spread as a result of its marginally greater transmissibility (4–6). Genetically much more divergent lineages of SARS-CoV-2 subsequently appeared, such as the VoCs alpha emerging at the end of 2020 and rapidly followed by delta (Table 1). The subsequent emergence and diversification of omicron lineages at the end of 2021 led to the replacement of delta and all other SARS-CoV-2 variants worldwide (7). This selective sweep is considered to be mediated by its enhanced replication capability, higher transmissibility, and escape from pre-existing immunity to previously circulating SARS-CoV-2 strains (8–10). Omicron has since diversified into several lineages such as BA.1-5, XBB, and HK lineages and most recently JN.1 (Table 1).

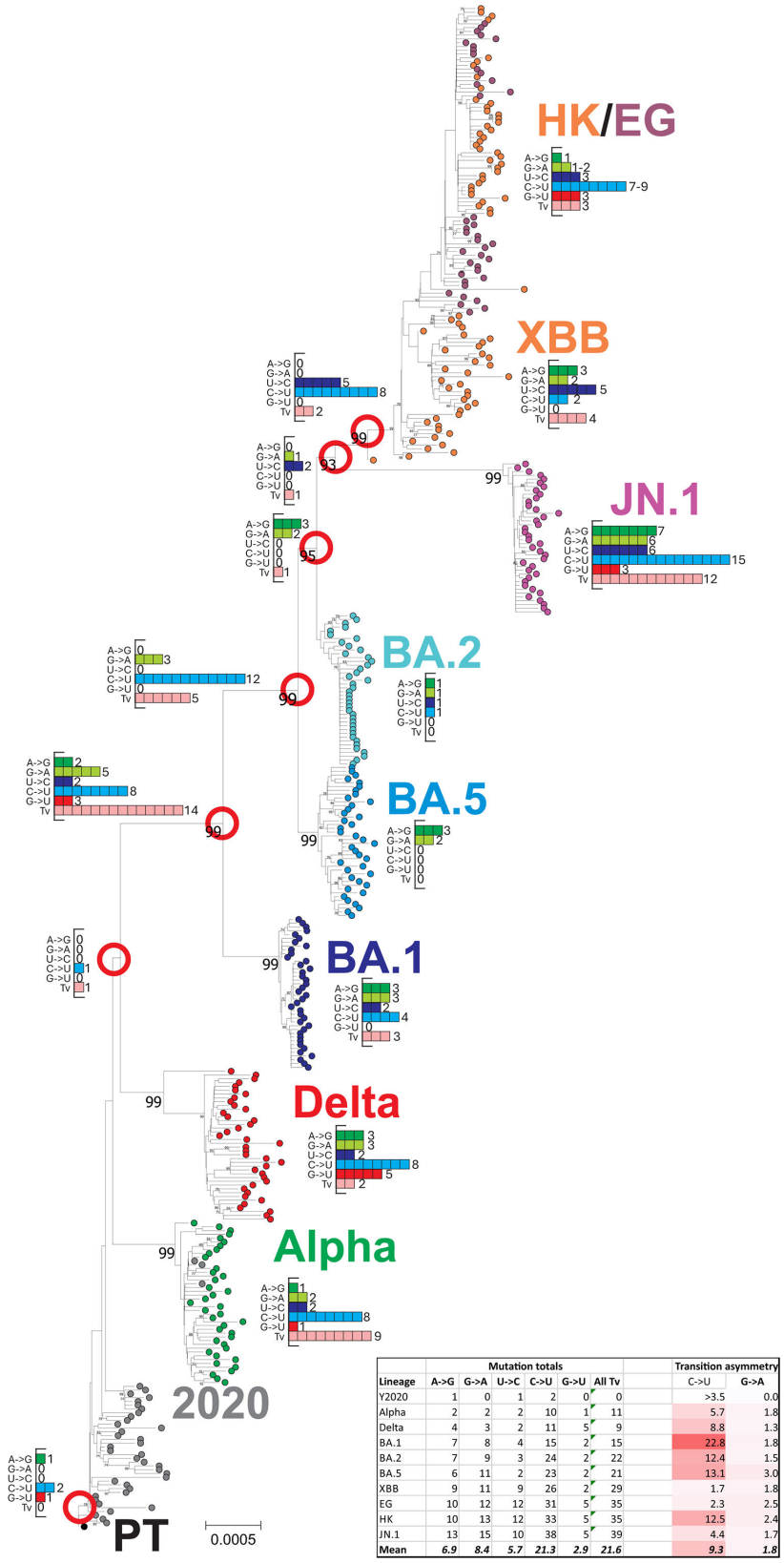
Sequence change in SARS-CoV-2 over this period has been characterized by the accumulation of neutral substitutions and rapid, phenotypically driven evolution of the spike gene. This has conferred major changes in antigenicity, neutralization susceptibility, and receptor binding, the latter potentially contributing to the marked increases in SARS-CoV-2 transmissibility over time. It also became very rapidly apparent that SARS-CoV-2 isolate consensus sequences contained a large number of often transient C→U substitutions distributed throughout the genome (11–17). Although without direct functional evidence at the time, these were frequently speculated to originate from host-mediated RNA editing pathways that constitute part of the cellular innate immune response to virus infections.

C→U substitutions are characteristic of the activity of members of the apolipoprotein B mRNA-editing enzyme, catalytic polypeptide-like (APOBEC) family (18). APOBEC typically targets single-stranded DNA templates for mutagenesis of cytidine to thymidine during reverse transcription of retroviruses and hepatitis B virus and during DNA genome replication of small DNA virus genomes such as papillomaviruses (18–20). In retroviruses, editing of the single-stranded DNA generated by reverse transcription generates a defective proviral copy whose encoded genes are unable to generate replication-competent progeny virus (21–23). APOBEC has been frequently discussed as the principal contributor to the over-representation of C→U transitions in SARS-CoV-2 (11–15, 17, 24) and potentially other viruses with RNA genomes (13, 25). However,

TABLE 1 Nomenclature, emergence, and properties of the SARS-CoV-2 variants

Label[a]	VoC[b]	PANGO[c]	Nextstrain[c]	Circulation[d]
Y2020	–	B, B.1	19A, 19B, 20A, 20B	Dec 2019–Dec 2020
Alpha	Alpha	B.1.1.7	20I	Oct 2020–Aug 2021
Delta	Delta	B.1.617	21A, 21I, 21J	Apr 2021–Dec 2021
BA.1	Omicron	BA.1	21K	Nov 2021–Mar 2022
BA.2		BA.2	21L	Dec 2021–Jun 2022
BA.5		BA.5	22B	Mar 2022–Mar 2023
XBB		XBB, XBB.1	22F, 23A, 23B, 23C, 23D	Oct 2022–Dec 2023
EG		EG.1, EG.5	23D, 23F	Mar 2023–Oct 2023
HK		HK.1–HK.34	23F	Jun 2023–Jan 2024
JN.1		JN, JN.1	24A, 24B	Oct 2023–present

[a]Abbreviated designation used in the study.
[b]WHO variant of concern assignment.
[c]PANGO and Nextstrain lineage designations (from https://cov-lineages.org/lineage_list.html; https://covariants.org).
[d]Approximate date range for global circulation (https://nextstrain.org/ncov/gisaid/global/all-time).



Lineage	Mutation totals A->G	G->A	U->C	C->U	G->U	All Tv	Transition asymmetry C->U	G->A
Y2020	1	0	1	2	0	0	>3.5	0.0
Alpha	2	2	2	10	1	11	5.7	1.8
Delta	4	3	2	11	5	9	8.8	1.3
BA.1	7	8	4	15	2	15	22.8	1.8
BA.2	7	9	3	24	2	22	12.4	1.5
BA.5	6	11	2	23	2	21	13.1	3.0
XBB	9	11	9	26	2	29	1.7	1.8
EG	10	12	12	31	5	35	2.3	2.5
HK	10	13	12	33	5	35	12.5	2.4
JN.1	13	15	10	38	5	39	4.4	1.7
Mean	*6.9*	*8.4*	*5.7*	*21.3*	*2.9*	*21.6*	*9.3*	*1.8*

FIG 1 Sequence changes in the emergence of SARS-CoV-2 lineages. Phylogenetic tree of coding region sequences of lineages of SARS-CoV-2 sequentially emerging during the course of the COVID-19 pandemic during 2020–2024. Inset histograms show the numbers of each transition, G→U, and total transversions

(Continued on next page)

Fig 1 (Continued)

that occurred between nodes (red circles) and descendant nodes or consensus lineage sequence. Substitution totals for each lineage are provided in the inset table, along with estimated transition asymmetries that account for base composition. The tree was constructed by neighbor joining of Jukes–Cantor corrected distances; robustness of grouping was determined by bootstrap re-sampling of 100 pseudo-replicas; support values of >70% are shown on the branches.

functional demonstration of its direct enzymatic role in RNA virus genome templates and a resulting impairment of replication is currently limited and controversial (26–29).

The extent to which host-induced mutations contribute to the evolutionary trajectory of SARS-CoV-2 can be better investigated several years after the start of the pandemic. In the current study, the overlay of potentially host-directed editing of SARS-CoV-2 genomes with their diversification into distinct lineages has been investigated using a large data set of sequences representing the principal SARS-CoV-2 lineages. These include alpha, delta, BA.1, BA.2, BA.5, XBB, EG, HK, and JN.1, spanning the pandemic period to the present (April 2024; Table 1). As lineages each have a clonal origin and have not inherited pre-existing population heterogeneity, the sharing of polymorphic sites (unfixed diversity) in each lineage at specific genome positions provides independent observations of mutational targeting. This provides a powerful method to investigate the potential existence of favored targets for mutation and the contexts in which they occur.

MATERIALS AND METHODS

Sequence data sets

New data sets of SARS-CoV-2 were constructed from publicly available complete genome sequences on GenBank, selected using NCBI Virus for lineage, genome sequence quality, and completeness (www.ncbi.nlm.nih.gov/labs/virus/). All available SARS-CoV-2 variants assigned as VoCs alpha and delta and the omicron lineages BA.1, BA.2, BA.5, XBB, EG, HK, and JN.1 (Table 1) were downloaded on 12 April 2024. A further data set of early SARS-CoV-2 variants not assigned to a lineage was created from sequences with sample dates pre-September 2020 (Y2020). From these downloads, 2,000 sequences from each were randomly selected (only 1,685 and 446 sequences of EG and HK were available) for analysis. GenBank annotations were used to assign sample dates to each sequence. Sequences were aligned by Nextclade v.2 with default settings (30), including the Wuhan-Hu-1 (MN908947/CN/2019) prototype strain in each alignment as a reference sequence. Available full genome sequences from SARS-CoV-2 variants infecting deer (n = 200), mink (n = 199), and panther (n = 120) were generated similarly.

Alignments of available complete genome sequences of the 10 lineages of SARS-CoV-2 and other coronaviruses analyzed in the study are provided as FASTA files in Github: https://github.com/MultipathogenGenomics/Sequence-Alignments-from-Mutation-Analysis-Study.

This includes alignments of all available seasonal coronavirus and MERS-CoV genome sequences used in a previous analysis (25).

A published compilation of substitutions observed within a collection of around 7 million SARS-CoV-2 genome sequences [including the set of all sequences in GISAID as of 29 March 2023 (31)] was used as an alternative data source. Sites were categorized by their relative frequency of occurrence, calculated as the natural logarithm of the ratio of actual to expected counts (described as estimated fitness, δf).

For both data sets, the regions spanning the start of the first open reading frame (ORF1a) to the end of the last ORF (NS10) were used for standard analysis to avoid areas of reduced coverage and potentially greater read error in the genome terminal regions.

Sequence analysis

Calculation of nucleotide composition was performed using the SSE package version 1.4 (32). Sequence changes within each full data set (up to 2,000 genome sequences) from a simple majority rule consensus sequence were compiled using the program SequenceChange; unfixed mutations were identified through the use of a <5% site variability threshold, calculated as the cumulative frequency of all non-consensus bases. The 5% threshold has been used as a means to infer the directionality of mutational change (25). Raw data on substitution frequencies at each genome position used in the analysis are provided in Table S1.

As described previously (11), normalized transition asymmetries of C→U(and comparably for U→C, G→A and A→G transitions) were calculated as f(C→U)/f(U→C) × (fU/fC), where f = frequency.

Potential biases in the identities of bases immediately 5′ and 3′ to C→U mutated sites were performed by a new modeling approach. This was based on the calculation of observed frequencies of bases at either side of C→U mutated sites in the concatenated ORF1a/ORF1b reading frame within alignments of representative sequence lineages (Y2020, delta, BA.1, and JN.1; 1,000 sequences in each) spanning the observation period of the study. These frequencies were then compared with those of an equivalent-sized data set of sequences that had been randomized in sequence. The program SequenceMutate in the SSE package performed sequence randomization under constraints: NDR, dinucleotide sequences preserved; COR, coding preserved; CDLR, coding sequence and dinucleotide frequencies preserved.

Phylogenetic analysis

Neighbor joining trees were constructed from aligned sequences using the program MEGA7 (33).

Statistical analysis

All statistical calculations and histogram constructions used SPSS version 29.

RESULTS

Sequence change during lineage evolution

Data sets of 2,000 aligned sequences of SARS-CoV-2 were selected to represent early pandemic onset strains (Y2020; pre-September 2020), VoCs alpha and delta, and the omicron lineages BA.1, BA.2, BA.5, XBB, EG, HK, and JN.1. These emerged sequentially in the 4 years since the start of the COVID-19 pandemic. Base substitutions associated with the emergence of each lineage were determined through the enumeration of sequence differences of lineage majority rule consensus sequences with the prototype Wuhan-1 strain (Fig. 1).

Substitutions from the prototype SARS-CoV-2 sequence were characterized by an excess of C→U substitutions (light blue) over other transitions and transversions throughout the tree. Major branching points were additionally associated with a higher proportion of transversions (pink bars) resulting from selected amino acid changes in the evolution of the spike gene. This was confirmed by analysis of the positions of each substitution type (Fig. 2). While there was a marked predominance of C→U changes in the ORF1a/1b gene and between ORF3a and the end of ORF10, there was a much higher proportion of other mutations in the spike gene with well-characterized involvement in antigenic change.

The effects of mutational biases in SARS-CoV-2 genomes on their overall composition were investigated by plotting C and U mononucleotide frequencies of the 10 lineage consensus sequences with mean sampling dates calculated from their component SARS-CoV-2 genome sequences (Fig. 3A and B). There was a progressive decline in the frequency of C and a corresponding rise in the frequencies of U consequent to the

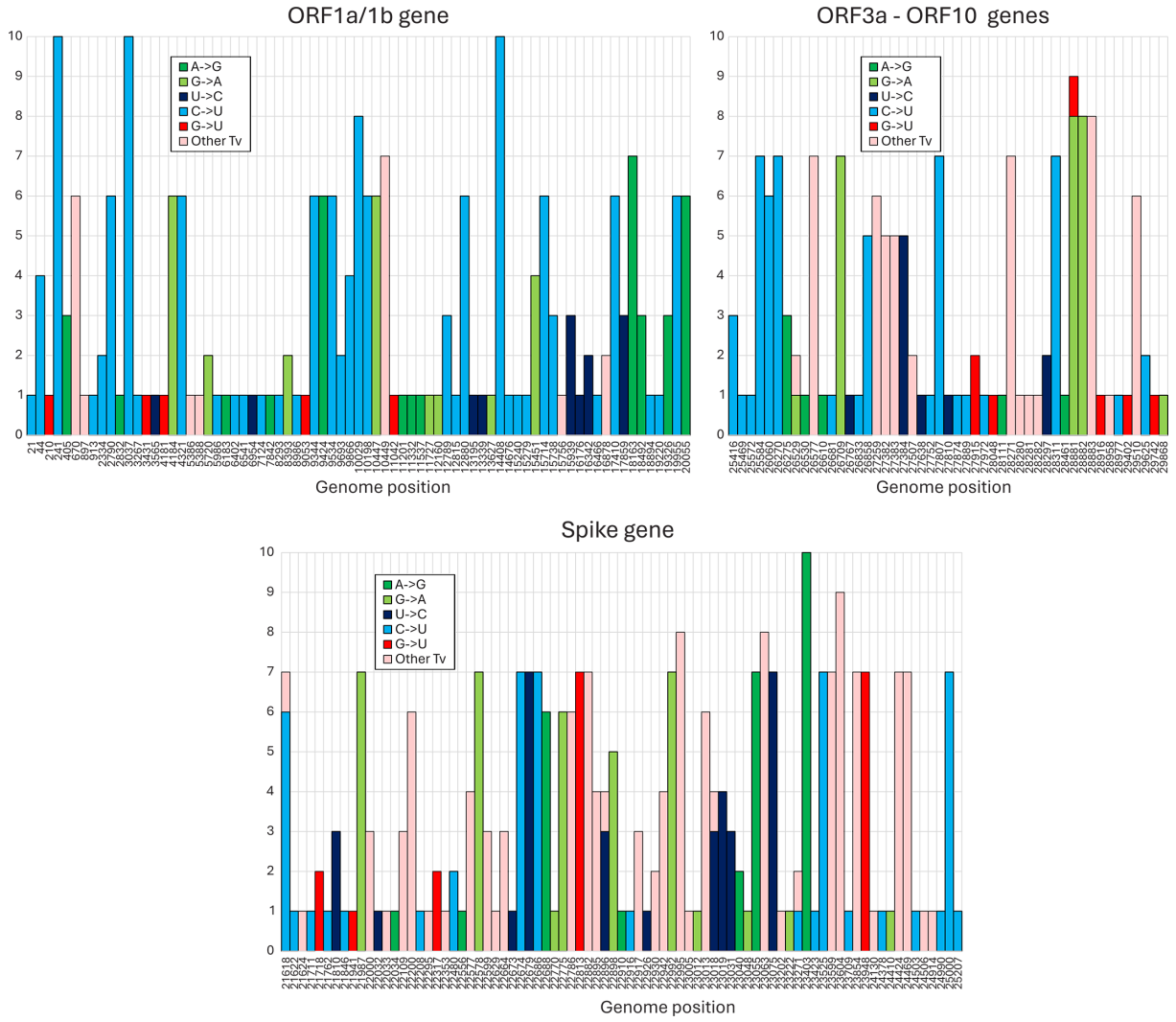


FIG 2 Genome position of substitutions in SARS-CoV-2 lineage consensus sequences. Positions of mutations in consensus sequences in the SARS-CoV-2 genome (only variable sites are plotted; positions indicated on the *x*-axis). The *y*-axis records numbers of lineages containing each indicated mutation. For clarity, the genome has been divided into ORF-1a/1b, S gene, and ORF3a-ORF10 sub-regions.

accumulated C→U transitions that were strongly coupled to the sampling date (R^2 values of 0.66 and 0.64, respectively). Linear trendlines converged on the base composition of Wuhan-1 around the end of 1989. C and U had trajectories in depletion and accumulation of 0.2%–0.25%/decade (Fig. 3C). Much smaller changes in the composition of G and A were observed (0.05%–0.06%), reflecting the much lower G→A/A→G transition asymmetry compared to C→U/U→C (Fig. 1); the depletion of G may additionally reflect the previously observed higher rate of G→U transversions (14, 16).

Compositional biases in other coronaviruses were consistent with longer-term effects of the C→U transition asymmetry (Fig. 4). All four human seasonal coronaviruses in the *Alphacoronavirus* (HCoV-229E and HCoV-NL63) and *Betacoronavirus* (HCoV-OC43 and HCoV-HKU1) genera showed far greater imbalances of C/G and U/A base frequencies (G/C: 4.9%–6.1%; U/A: 7.6%–12.8%) than found in SARS-CoV-2, SARS-CoV, or any of the sarbecoviruses infecting bats (G/C: 0.93%–2.2%; U/A: 2.1%–3.4%). The extent to which this reflects a greater accumulated C→U mutational effect on non-bat hosts is discussed below.

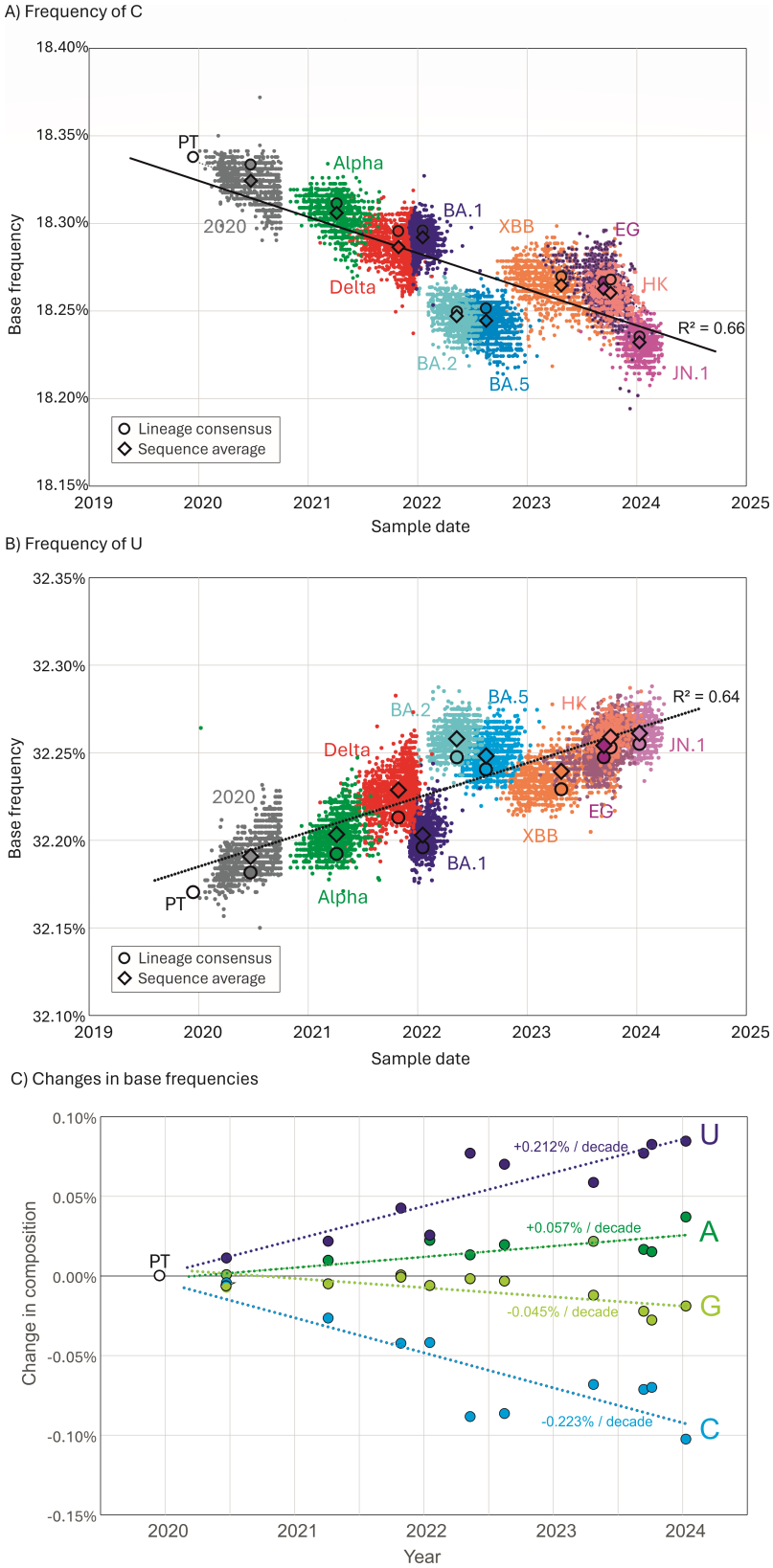


FIG 3 Changes in SARS-CoV-2 genome composition since the start of the pandemic. (A and B) Frequencies of C and U bases in individual genome sequences of SARS-CoV-2 isolates collected at different time points from the start of the pandemic. Lineage assignments are indicated. Circles

(Continued on next page)

Fig 3 (Continued)

indicate the mean sampling date and composition of consensus sequences from the 10 lineages and the prototype sequence; diamonds indicate mean compositions of individual sequences. (C) Differences in A, C, G, and U base frequencies between the genome sequence of the PT (Wuhan-1) and of lineage consensus sequences collected at different time points from the start of the pandemic.

Within-lineage base substitutions

Since the lineages analyzed in the study have a clonal origin, sequence diversity within lineages arises from evolutionarily independent events occurring within each, rather than being inherited from previous virus populations. Therefore, sequence substitutions observed within lineages have occurred *de novo* over the course of lineage diversification. Furthermore, enumerating sites where mutation frequencies are less than

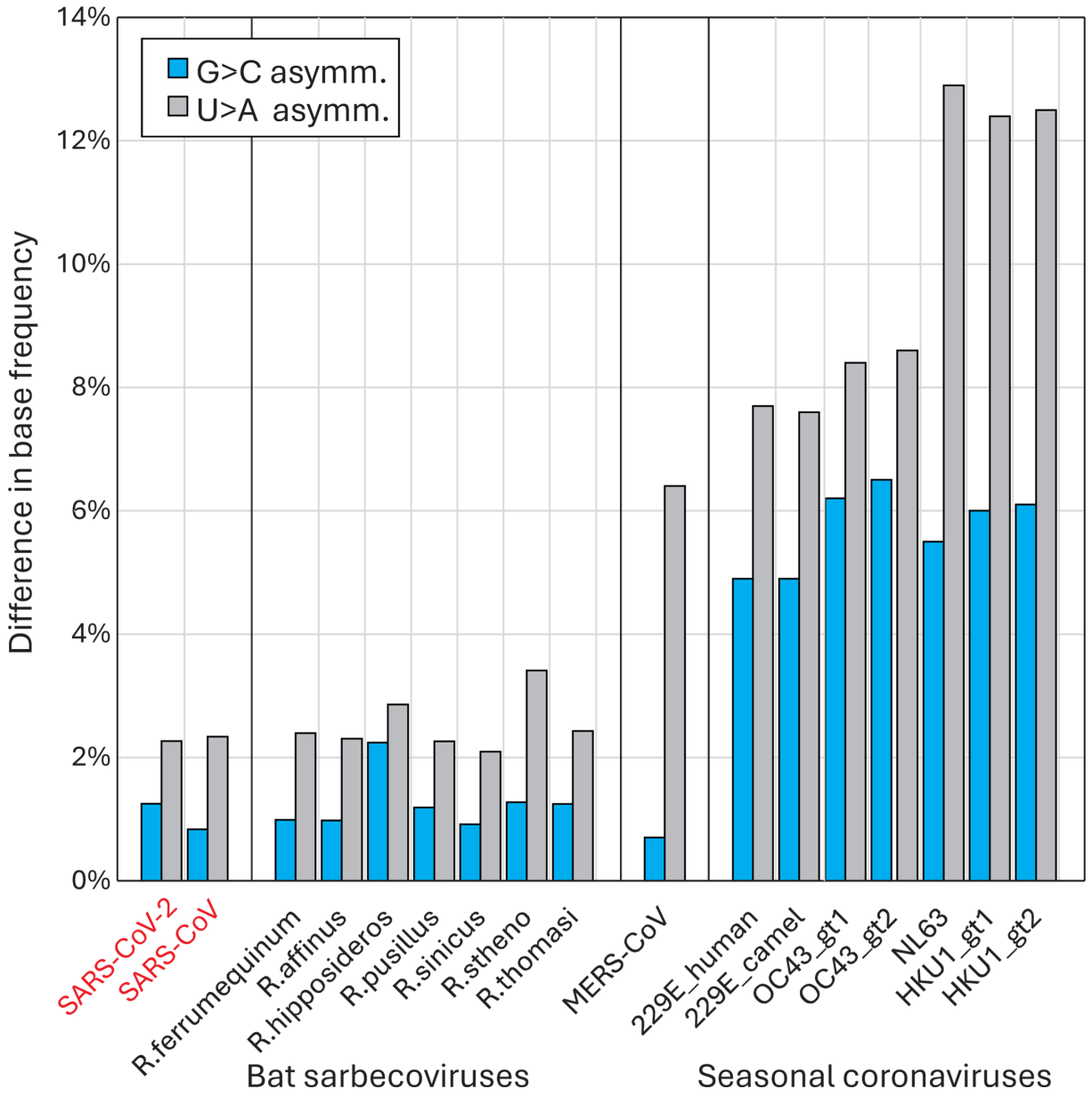


FIG 4 Compositional asymmetries in other coronaviruses. Comparison of base compositions of the recently zoonotic SARS-CoV-2 and SARS-CoV isolate sequences with sarbecoviruses infecting the *Rhinolophus* genus of bats (host species indicated on *x*-axis) and with genotypes of proposed potentially zoonotically acquired human seasonal coronaviruses. Blue bars record the excess of G over C (C depletion), and gray bars the excess of U over A (U accumulation).

5% enables an inference of directionality [as previously discussed (25)] that is independent of the evolutionary reconstructions used in the analysis of consensus sequences in the previous section. The use of large data sets of sequences for each lineage (typically 2,000 whole genome sequences) revealed around a thousand times more polymorphic sites than were apparent from parsimony-based sequence reconstructions. As they were unfixed, this also provides the means to investigate the characteristics of C→U and other mutated sites using 10 evolutionarily independent sets of observations from each lineage.

Applying the 5% heterogeneity filter to record unfixed substitutions within each lineage and their directionality, each showed similar excesses of C→U transitions compared to other substitutions (Fig. 5). The range of values (4.1–9.9) encompassed the mean transition asymmetry observed in consensus sequences (9.3). Contrastingly, there was a virtual absence of G→A/A→G transition asymmetry (ratios of 1.0–1.49) in any of the lineages. Transition asymmetries in SARS-CoV-2 variants infecting mink and panther were similar to those observed in human-derived variants, but much greater in deer (C→U/U→C: 14.4; G→A/A→G: 1.1). Transition asymmetries were characteristic of human and derived strains in animals, but not in sarbecoviruses more generally (Fig. 5B). There was no evidence for C→U/U→C (or G→A/A→G) asymmetries between the whole data set of bat sarbecoviruses with a reconstructed ancestral sequence and a subset of more closely related sequences variants including RatG13 and SARS-CoV-2.

Overall, discounting the data sets with fewer than 2,000 available sequences (HK, EG), lineages showed a mean of 1,429 sites (range 1,075–2,004) with C→U transitions, compared to sites with other transitions (A→G: 465; G→A: 392; and U→C: 534). Sites of C→U transitions occurred in over a quarter of all sites with a C (26.5%; range 20.0%–37.2%), around five times more than observed at sites with the appropriate majority base for the other transitions (5.3%–6.8%). To investigate whether certain sites were more likely to mutate than others, the distribution of sites with zero, one, two, or more lineages with C→U and other transitions was plotted and compared with the null expectation of a random distribution (Fig. 6). A→G, G→A, and U→C transitions showed frequencies among lineages that closely matched an unbiased distribution calculated from the Poisson distribution based on mean frequencies of transitions per nucleotide position. However, the distribution of C→U transitions differed strikingly, with far more transition sites shared among six or more lineages and, conversely, an over-representation of invariant sites compared to the Poisson null expectation. Sites of C→U substitutions in three or more lineages were consistently in large numerical excess to those with other transitions (total and proportions shown above the graph). Favored sites for C→U editing by A3A during the *in vitro* passage of SARS-CoV-2 were recently reported (27). Most but not all of the mutated sites in the 10 lineages were also those with high frequencies of unfixed C→U mutations among native SARS-CoV-2 variants (Fig. 6; Fig. S1), indicating some commonality in targeting (see Discussion).

Favored contexts for C→U transitions

To investigate whether particular 5′ or 3′ bases were found in association with C→U substitutions, as typically found in sites edited by APOBECs, a novel modeling approach was developed to calculate expected frequencies based on a null expectation of no context bias. A more sophisticated approach was required beyond simply predicting frequencies from mononucleotide composition, as almost all sites lie within functional protein-encoding genes that impose coding constraints from amino acid usage. Biases may additionally arise from biological selection against certain dinucleotides, such as CpG and UpA that may additionally impact upstream and downstream expected base composition.

To calculate this, expected frequencies of A, C, G, and U upstream and downstream of a C residue were calculated from in-frame alignments of the ORF1a/ORF1b concatenated gene of four representative lineages (Y202, Delta, BA.1, and JN.1 spanning the observation period) using simple mononucleotide frequencies (29.9%, 18.3%, 19.6%, and

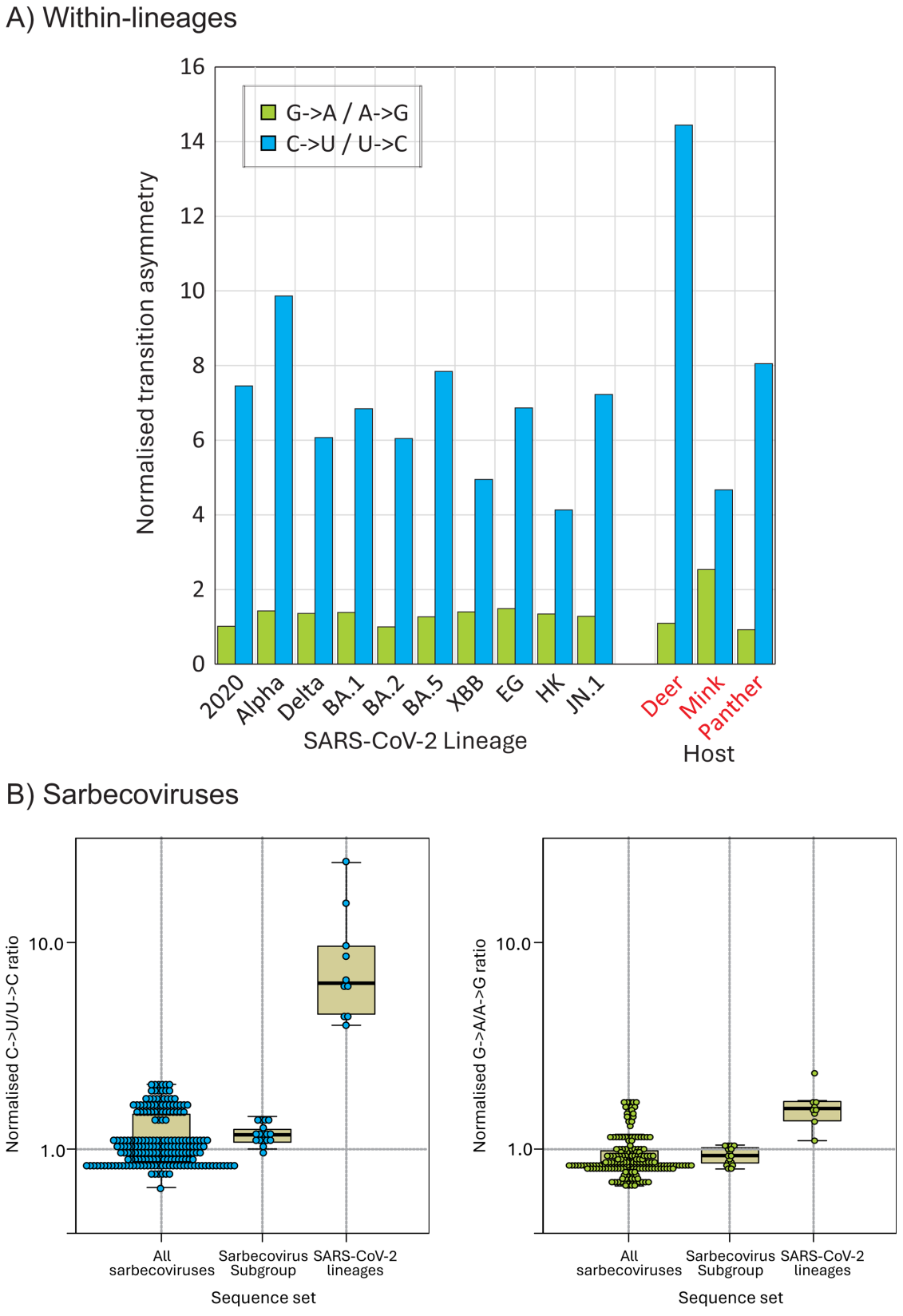


FIG 5 Transition asymmetries with SARS-CoV-2 lineages. (A) Normalized transition asymmetries of fixed C→U and G→A substitutions within the 10 SARS-CoV-2 lineages. Comparisons with SARS-CoV-2 strain infections; non-human species shown on the right panel. (B) Normalized transition asymmetries of all sarbecovirus strains infecting bats (column 1), the subset (including RatG13) most closely related to human SARS-CoV-2 (column 2) and consensus sequences of lineages within SARS-CoV-2 compared to the prototype Wuhan-1 isolate sequence (column 3).

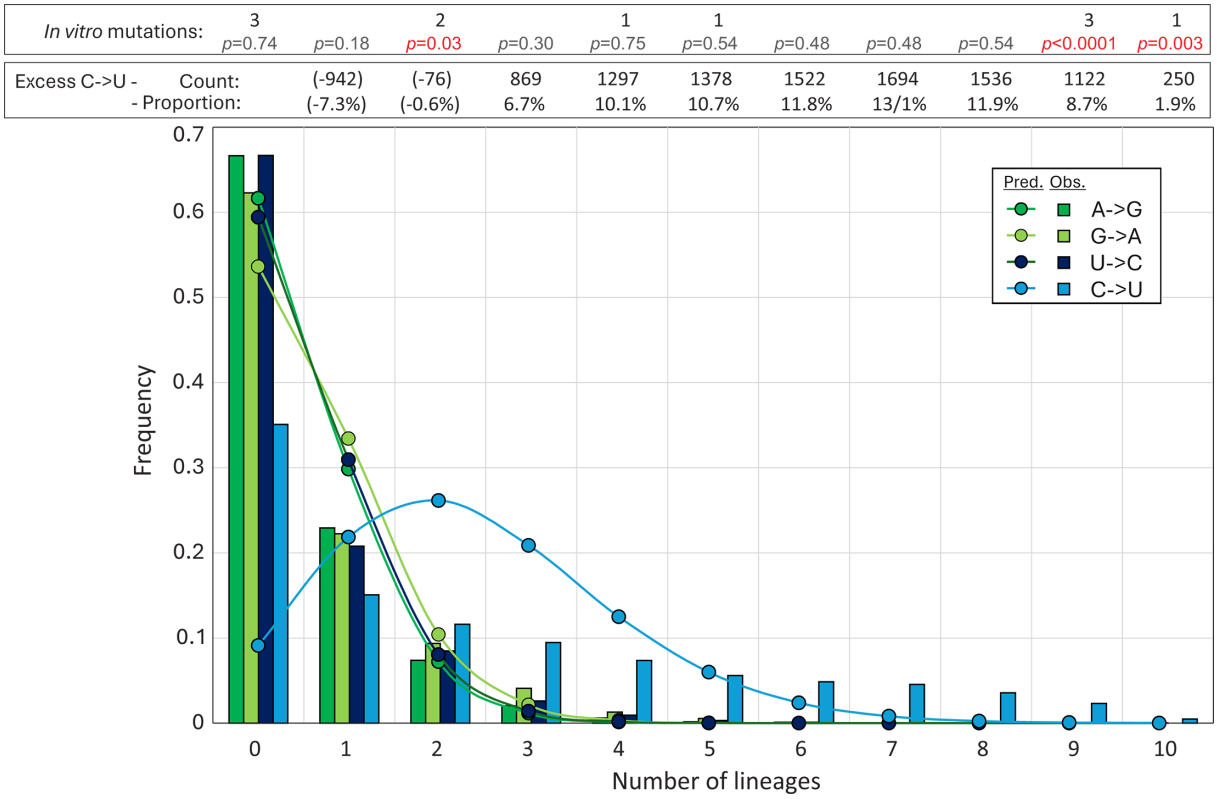


FIG 6 Distribution of mutations. The distribution of unfixed transitions (<5% frequency) between lineages (histogram) compared to expected values assuming no positional bias derived from the Poisson distribution (graph points). Excess numbers over the mean numbers of A→G, G→A, and U→C transitions (total excess 8,651 from 12,907 C→U transitions) and their proportions of total C→U substitutions in each category are shown in the upper box. The distribution of C→U transitions at sites of previously reported *in vitro* A3A-induced mutations (27) is shown in the lower box (Fig. S1); differences in lineage distributions with observed values calculated by Pearson chi-squared.

32.2% respectively; Fig. 7A and B). However, observed frequencies of the four bases in native sequences of the four lineages differed markedly, irrespective of whether the C site was non-mutated or mutated in at least one lineage or multiple lineages (red bar; Fig. 7A and B). Observed frequencies of upstream A of 44.6%, 42.4%, and 47.8% were significantly higher than the predicted 29.9%; similarly, frequencies of downstream U were substantially higher (52.1%, 49.5%, and 53.9%) than the predicted 32.2%.

To determine whether these differences arose from favored contexts for C→U mutations or whether these were expected frequencies once coding constraints and dinucleotide frequency biases in native sequences had been taken into account, a total of 1,000 ORF1a/ORF1b sequence randomizations were performed on the consensus sequences of each of the four lineages. Randomization using the algorithm CDLR (in the SSE package) created highly divergent sequences from the native sequence while preserving the protein coding of each as well as native dinucleotide frequencies. As encoded amino acid sequences were invariant, 3′ composition comparisons were most usefully performed at the downstream (+1) site of all C's at codon position (CP) 2 (all codons with a C at position 2 are four-way redundant at CP3). Similarly, upstream (−1) base frequencies with and without sequence randomization were compared for C's at CP1 at CP3 of the preceding codon. For comparability, only four-way redundant upstream codons were included in this comparison.

The analysis demonstrated virtual equivalence in both upstream and downstream base context frequencies between native SARS-CoV-2 sequences and those randomized by CDLR (Fig. 7A). As expected, the observed frequency of G downstream of C was much lower than predicted based on mononucleotide frequencies, reflecting the

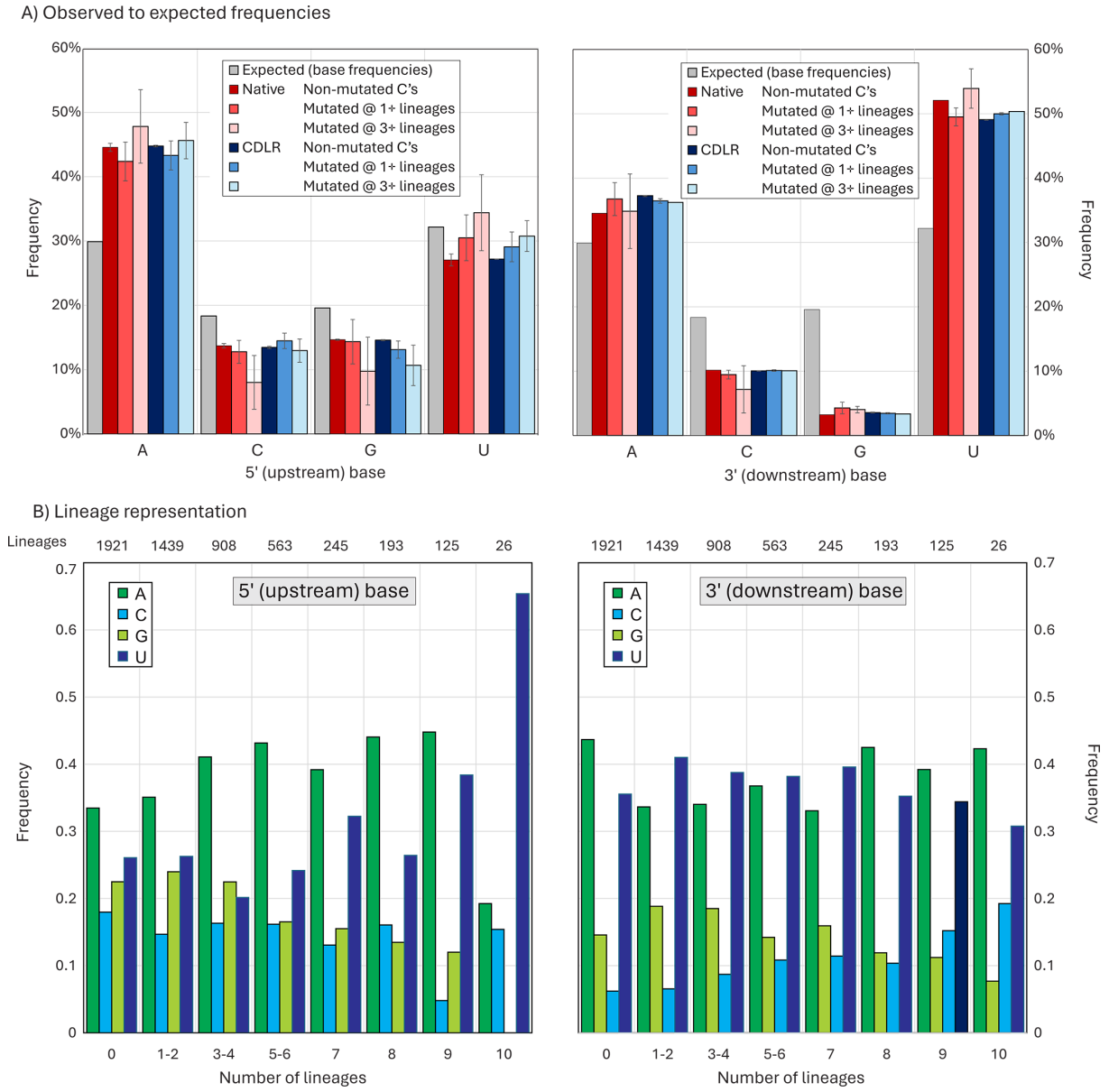


FIG 7 Observed and expected 5′ and 3′ base context frequencies of C residues in the SARS-CoV-2 genome sequences, either non-mutated or showing C→U mutations. (A) Predicted and observed frequencies of bases 5′ (upstream) and 3′ (downstream) of C's, split into sites that were invariant or mutated in one or more lineages. Observed frequencies were compared with those of sequences randomized by CDLR that preserved native coding and dinucleotide frequencies. Frequency estimations based on mononucleotide frequencies alone are indicated by gray bars. Error bars for native and CDLR sequences show one standard deviation of values from the four lineages analyzed (Y2020, delta, BA.1, and JN.1). (B) Distribution of 5′ and 3′ bases at C→U mutation sites occurring in different numbers of lineages and at invariant C sites.

suppression of CpG in vertebrate RNA viruses; however, this bias was correctly reproduced in sequences scrambled by CDLR. Furthermore, there was no 5′ or 3′ compositional difference between C's that were invariant from those that showed C→U changes in one or more lineages. All frequencies thus matched closely the null expectation once protein coding and dinucleotide biases had been taken into account.

While upstream and downstream base frequencies at C→U mutated sites were unbiased, most sites analyzed in Fig. 7A were variable in three or fewer lineages. To investigate whether more frequently mutated C bases showed a more obvious 5′ or 3′ base context bias, sites were binned into seven variability categories along with invariant

sites, and 5′ and 3′ base frequencies re-calculated (Fig. 7B). Remarkably, C→U mutation sites that occurred recurrently in 9 and 10 lineages showed an increasing and ultimately extreme 5′ context preference for U, while the 3′ context remained unchanged from less variable sites. However, there was a substantial over-representation in sites that were polymorphic in five to eight lineages, but these failed to show a site preference for a 5′U (Table S2A). This indicates that the majority of sites over-represented for C→U substitution sites did not possess the correct upstream base associated with A3A targeting.

Association of C→U mutations with RNA secondary structure formation

A consensus RNA secondary structure model of the SARS-CoV-2 genome based on three independent studies using biochemical mapping and prediction methods (34–36) was used to determine whether C→U or other transitions occurred preferentially at sites that were non-base-paired. The number of lineages with mutations occurring at unpaired or paired sites was compared with unpaired/paired ratios of invariant A, C, G, and U bases (Fig. 8A). Ratios around 1 indicated no mutational bias toward paired or unpaired sites and characterized the preferences of A→G, G→A, and U→C transitions. Contrastingly, U→C and G→U mutations were strongly conditioned by pairing status, with a remarkable relationship between occurrence in multiple lineages and increasing mutational bias toward unpaired bases. For example, of the 175 sites of C→U mutations occurring within eight or more of the 10 SARS-CoV-2 lineages, 154 were unpaired compared to 25 paired (ratio 6.2), quite different from the unpaired/ paired ratio of invariant C sites (688 and 1,311; ratio 0.53). There was consequently a >11.7-fold skew toward unpaired bases at highly mutated C→U sites. Unexpectedly, G→U transversion also showed a similar preference for unpaired bases. Frequencies of unpaired sites were similarly higher at more polymorphic sites of C→U transitions in the published analysis of 7 million SARS-CoV-2 sequences (Fig. 8B). However, as for 5′U context preferences, the majority of C→U over-represented sites (polymorphic in 4+ lineages) did not show a preference for unpaired bases (Table S2B).

DISCUSSION

Trajectory of SARS-CoV-2 evolution

The publication of an unprecedented number of accurate SARS-CoV-2 complete genome sequences 4 years on from the start of the COVID-19 pandemic provides an opportunity to re-investigate the trajectory of SARS-CoV-2 evolution. In particular, the sequence data enabled further exploration of the reported mutational biases driving the high rate of C→U transitions in consensus sequences of SARS-CoV-2 isolates (11–17). The longer timescale provides a convincing demonstration of the sustained loss of C bases and accumulation of U's resulting from this transition bias, with net rates of −0.25% and +0.25%/decade, respectively (Fig. 3 and 5). These observations provide an insight into the origin of the much greater imbalance of complementary bases (G > C, U > A) in other human seasonal coronaviruses (37, 38); if the imbalances do indeed originate from C→U hypermutation, the extreme depletion of C and accumulation of U observed in HCoV-OC43 and HCoV-HKU1 would predict prolonged periods of post-zoonotic infection compared to SARS-CoV-2. This would be considerably earlier than the proposed attribution of HCoV-OC43 to the Russian "flu" global outbreak in 1892–3 (39, 40), where human infection was proposed to have been acquired from a related coronavirus infecting cattle. However, the bovine coronavirus from which HCoV-OC43 derived may already have been compositionally altered from mutational C→U pressure in cows as an intermediate host. The extreme transition asymmetry in SARS-CoV-2 after spread into deer (Fig. 5) indeed suggests that ungulates may harbor even more potent mutational drivers than found in human (and carnivore) cells.

Although there is a lack of large data sets of nucleotide sequences from bat-derived sarbecoviruses, and existing strains are relatively more divergent from each other than between SARS-CoV-2 isolates, there was no evidence for a C→U hypermutation

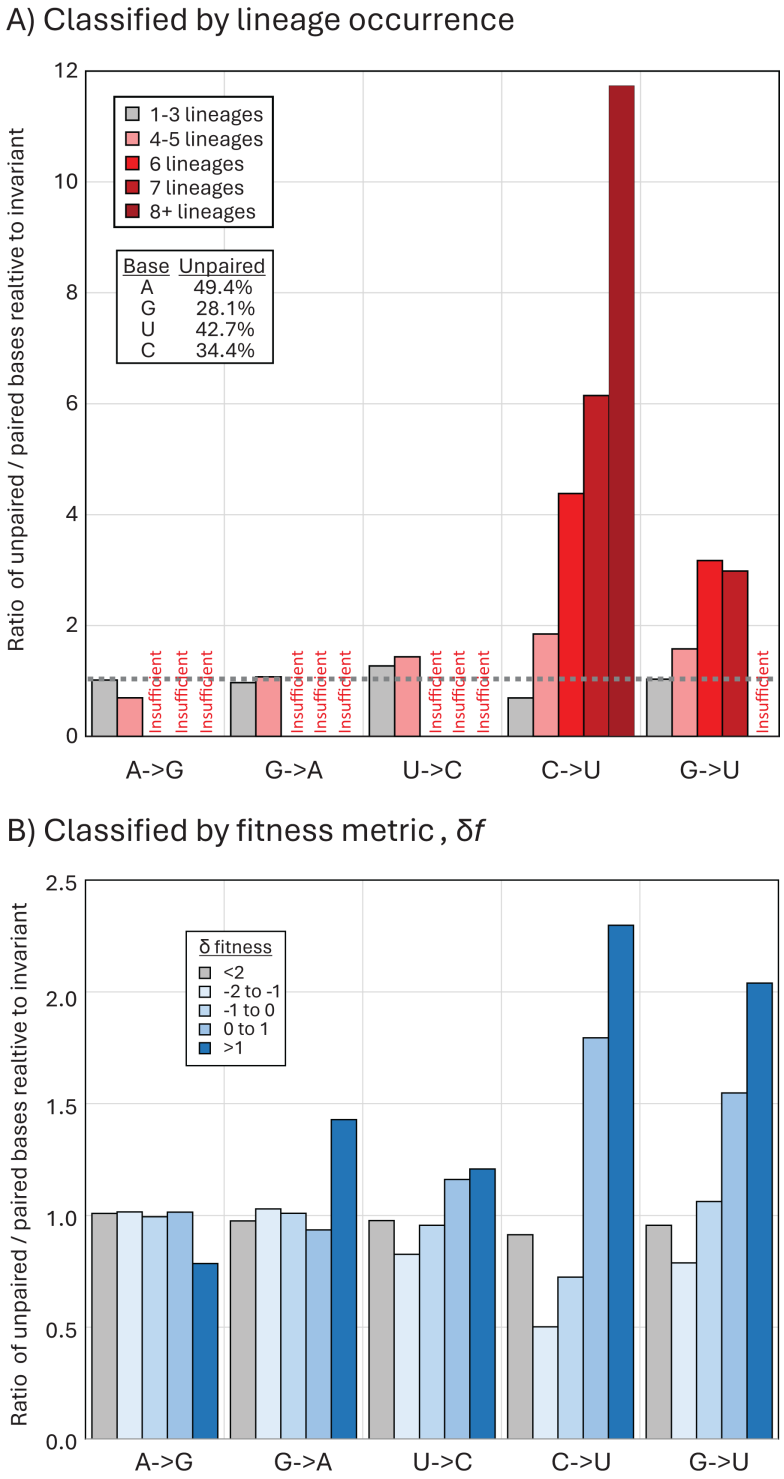


FIG 8 Transition site associations with RNA secondary structure pairing. (A) Comparison of frequencies of unpaired bases at sites with mutations in one or more lineage (gray bars) and four, six, eight, or more lineages (pink, red, and dark red bars) with those of invariant sites (proportions of unpaired bases shown in inset table). (B) Frequency of unpaired bases at sites classified by the fitness metric, *δf* (log ratio of observed substitutions/expected number) (31).

based on comparisons with reconstructed ancestral sequences and a 5% directionality threshold (Fig. 5B), nor were there marked base frequency asymmetries observed in human seasonal coronaviruses (Fig. 5A). These observations suggest that replication

in bats is not associated with the same mutational pressure as found in human or other mammalian infections. While potential mechanisms remain controversial, the host comparisons in the current study provide strong evidence that the hypermutation phenomenon is host-derived rather than being an intrinsic property of sarbecovirus and of wider coronavirus replication strategies, or misincorporation frequencies by the coronavirus RdRP. The existence of C→U/U→C transition asymmetries in the mutational spectra of a wide range of other coronaviruses (38) and mammalian +strand RNA viruses (25) similarly argues that the phenomenon represents a more general host-driven phenomenon rather than being connected to the specifics of the SARS-CoV-2 replication complex.

Mutational biases in SARS-CoV-2 genome sequences

Numerous studies document a range of mutational biases in the evolution of SARS-CoV-2. These include ADAR1-mediated editing of G bases to inosine in viral dsRNA templates that resolve to G→A substitutions (and theoretically equal frequencies of T→C mutations derived from editing of the complementary strand). Such mutations were originally described within SARS-CoV-2 populations (12, 14, 15) but are less prominent in very large data sets of consensus sequences from different SARS-CoV-2 isolates [e.g., references (31, 41) and in the original studies (11–17)]. Instead, all data sets record the dominance of the C→U transition and its asymmetry relative to other transitions. Over the 4-year study period, C→U transition asymmetries observed in the 10 lineages analyzed remained comparable (Fig. 5A), as did the trajectory of loss of C bases and gain of U's (Fig. 3). The findings do not support the previous conclusion for substantial changes in C→U editing frequencies in delta or omicron lineages (42).

There have been some countervailing views on the existence of the C→U mutational bias (26), but typically, these are based on observations of within-population polymorphisms where other mutational effects, such as sequencing errors, may contribute significantly to the observed diversity. The transition asymmetry in SARS-CoV-2 sequences and the proposed wider occurrence of the phenomenon in some but not all vertebrate RNA viruses (25) is evidently a real phenomenon that requires a mechanistic explanation.

Drivers of C→U hypermutation

The most frequently advanced hypothesis, now with some experimental evidence, is that observed C→U transition bias arises from direct editing of viral RNA sequences by one or more isoforms of the APOBEC family of nucleic acid deaminases (11–17). APOBECs A3D, A3F, A3G, and A3H are potent inhibitors of retroviruses and co-transcriptionally edit transcripts of DNA during reverse transcription of proviral sequence pre-integration [reviewed in reference (18)]. Errors so introduced create severe and permanent replication defects in the progeny virus. APOBEC activity creates an observable depletion of UpC or UpU dinucleotides in genome sequences of a range of DNA viruses, retroelements, and some RNA viruses (43, 44). The observed depletion of UpC in seasonal coronaviruses might be similarly interpreted as evidence for the activity of other APOBEC isoforms, notably human A3A, on RNA sequences. There is also, for example, evidence for extensive A3A-mediated editing of human mRNA sequences as part of a cellular stress response, with targeting of unpaired bases in RNA secondary structures (45, 46), similar to the originally described editing of the ApoB mRNA 3′UTR mRNA by A1 (47, 48). The RNA genome of rubella virus shows a marked C→U/U→C transition asymmetry and genome-wide depletion of UpC (49). The replication of the seasonal coronavirus, HCoV-NL63, was impaired in cells expressing A3C, A3F, and A3H isoforms of APOBEC although in this case it was not demonstrated to result from introduced mutations in the genome (29).

Several studies have documented extensive secondary structure formation in the SARS-CoV-2 positive sense genomic RNA, manifested through the formation of sequential extensively internally base-pair stem-loops throughout the non-coding and

coding regions of the genome (34–36, 50). On the face of it, there is, therefore, no reason to suppose that SARS-CoV-2 genomic RNA or mRNAs might not also be targeted by APOBEC in an RNA structure-dependent manner comparable to what has already been documented for cellular mRNAs and other RNA viruses with single-stranded RNA genomes. Recently, it was demonstrated that over-expression of A3A (but not other isoforms of human APOBEC) induced mutations in the genome of SARS-CoV-2 during *in vitro* passage in 293T cells at several genomic sites (27). These occurred in the favored 5′U context and occurred in unpaired bases in the terminal loops of RNA secondary structures. Induced mutations did not impair the overall replication fitness of SARS-CoV-2, consistent with a reported lack of correlation between the extent of C→U editing and SARS-CoV-2 titer on *in vitro* passaging (51). Indeed, another study showed that a functional A3A appeared to enhance replication, perhaps by providing greater opportunities for the fixation of phenotypically advantageous mutations, such as C241U in the 5′UTR (28). However, high multiplicity passage would rapidly drive out even marginally deleterious C→U mutations through natural selection and might hide the actual impact of APOBEC-mediated editing on replication fitness.

While the *in vitro* studies provide an important functional characterization of A3A-mediated RNA editing, it is not clear whether A3A drives the bulk of C→U transitions *in vivo*. None of the published data sets analyzing consensus sequences of SARS-CoV-2 variants demonstrate a pronounced preference for a 5′U context at sites of C→U transitions that could not be ottherwise accounted for by high genome content of U (and A) bases in the SARS-CoV-2 genome. This includes the set of C→U substitutions fixed during lineage diversification (Fig. 1), and unfixed mutations within lineages (Fig. 5)—for the latter, a modeling approach in which randomization was applied under the same constraints that might operate *in vivo* (coding and preservation of dinucleotide frequency biases) resulted in 5′ and 3′ base frequencies identical to those surrounding C→U edited sites (Fig. 7A). In this bioinformatic analysis (Fig. 7B) and in the functional studies, it appears that only the most frequently mutated sites *in vivo* or *in vitro* conform to the expected targeting preferences of A3A for a 5′U and positioning inside unpaired regions of stem-loops.

These, however, represent an extremely small fraction of the total number of *in vitro* edited sites of C→U transitions (27) and those showing elevated C→U transition frequencies in SARS-Cov-2 isolates (Fig. 7B). For example, there was little or no 5′ base preference among the unfixed C→U substitutions within eight or fewer lineages in the current study data (Fig. 7B), corresponding to around 80% of the excess C→U substitutions over other transitions. Similarly, the specific targeting of unpaired bases by A3A recorded in *in vitro* studies was not reproduced in site heterogeneity of SARS-CoV-2 isolates (Fig. 8A and B). While a higher proportion of highly polymorphic C→U sites were unpaired, a substantial proportion of the excess C→U substitutions also occurred in base-paired regions.

The absence of a clear context for C→U edited sites within SARS-CoV-2 populations is consistent with the hypothesis for alternative, currently functionally uncharacterized mutational mechanisms in edited sites (26). One potential source of mutations in the study population is the activity of the SARS-CoV-2 antiviral, molnupiravir, whose inhibitory effect on replication is mediated through inducing G→A and C→U mutations (52, 53). However, this explanation can be ruled out on several grounds; most obviously, the C→U mutation phenomenon was repeatedly observed in sequence data sets (11–15, 17, 24) long before the clinical trials and subsequent restricted use of molnupiravir for COVID-19 treatment from 2022 (54); it has also never been widely used enough to have any impact on the global data sets of SARS-CoV-2 strains assembled for the current study. Secondly, despite its documented mechanism of action, the expected signature of increased G→A and C→U transitions induced by molnupiravir was not observed in treated patients (55), perhaps attributable to the fact that mutated viruses would not propagate and might only be observed in very restricted lineages of SARS-CoV-2, rather than in bulk populations (56). Finally, the expected mutational signature of an increased

frequency of G→U and C→U mutations was not the primary observation in SARS-CoV-2 data sets (11–15, 17, 24), where only C→U transitions were obviously over-represented. Indeed, analysis of SARS-CoV-2 lineages circulating before and after the introduction of molnupiravir for therapeutic use showed no increase in G→A transition asymmetry (Fig. 5A).

An alternative possibility for the lack of defined contexts in all but the most heavily mutated sites in SARS-CoV-2 might simply be that the activity of APOBECs on RNA does not possess the target specificity associated with DNA editing (46). Characterization of mutations induced by A3A in human mRNA sequences indeed provided evidence for extensive off-target activity—while C bases in human mRNA were preferentially edited within unpaired regions of stem-loops, around a third of mutated sites occurred in other RNA structure contexts. There was similar variability in 5′ contexts and pairing in sites edited in an *in vitro* RNA expression system of a short section of the SARS-CoV-2 genome (28). Therefore, unlike APOBEC-mediated editing of DNA sequences, specific contexts for RNA mutagenesis may be preferred but not absolutely required (46).

Biological impact of C→U mutations in COVID-19

APOBECs have evolved as a potent and essential defense against many virus infections; the expansion of the A3 locus in primates is thought to have been a host response to the spread of retroviruses early in their diversification (57, 58). Consequently, APOBEC repertoires are highly diverse among different orders of mammals, and the existence and activity of precise homologs of A3A (that appears to mediate editing of SARS-CoV-2 in human cells) in other species are poorly understood. Speculatively, the remarkable editing of SARS-CoV-2 in deer (Fig. 4) potentially mediated by one of the much more restricted number of A3 genes in artiodactyls (one with the A3Z1 domain organization corresponding to human A3A (59)) perhaps represents an example of a more active restriction mechanism for RNA viruses than found in the human or other primate genome.

RNA viruses with single-stranded genomes typically follow the first of Chargaff's rules that genome frequencies of C should match those of G and A ≈ T/U (60); these reflect the symmetry of base incorporation biases on copying plus and minus strands of genomic RNAs. However, genome compositions of seasonal coronaviruses violate this principle [Fig. 4 (38)], and it might be speculated that their composition represents a much longer-term endpoint of the loss of C bases and accumulation of U's apparent in the SARS-CoV-2 genome (Fig. 3). Contrastingly, bat sarbecovirus show minimal imbalances (Fig. 4) and no evidence for excess C→U substitutions in their diversification (Fig. 5B). These observations suggest that either bat APOBECs are incapable of editing sarbecovirus sequences or, more likely, sarbecoviruses have evolved antagonists to bat APOBECs that prevent genome editing as part of a longer process of virus/host co-adaptation. This might be analogous to the evolution of Vif and other antagonists of APOBECs that target post-entry reverse transcription of retroviral genomic RNA (23, 61). An inability of SARS-CoV-2 to antagonize human A3A is indeed a plausible outcome of the broad genetic and functional diversity of the APOBEC locus in different mammalian orders. As another possible analogy, monkeypox virus infection in humans was associated with the appearance of a clade bearing multiple (40+) and in this case symmetric C→T/G→A mutations attributed to genome editing of the dsDNA genome associated with a failure to antagonize DNA-targeting A3s such as A3G (62). Similarly, an inability of Vif in simian and feline lentiviruses to antagonize APOBECs in heterologous species represented a key factor limiting their host range (63, 64). These examples support a more general hypothesis that the failure of viruses in the "wrong" host to antagonize the APOBEC editing pathway may be key determinants in their host range (65).

This background provides some context for understanding the burning question of whether the observed C→U mutations introduced into the SARS-CoV-2 genome during the pandemic have measurably affected its replication and ability to transmit. Related to that, are seasonal coronaviruses irreversibly damaged by their current skewed

genome composition (Fig. 4)? Kim et al. proposed that C→U editing may be beneficial by providing a source of novel mutations that might enhance the ability of SARS-CoV-2 to adapt to human transmission and escape from host immune responses (28, 66). These studies discuss the potential of RNA editing to generate novel mutations that may confer fitness benefits to the virus, such as immune escape, enhancement of receptor binding, and escape for host immunity. It was notable that C→U substitutions have been fixed in 68 different sites within the 10 lineages of SARS-CoV-2 analyzed in the current study (Fig. 1), consistent with the potential contribution of some of these to replication fitness and increased transmissibility of SARS-CoV-2 variants emerging during the pandemic.

On the other hand, the vast majority of C→U substitutions are likely unobserved because of their adverse effect on replication phenotypes; alternatively, they may impart a marginal fitness loss that contributes to observed cycles of mutation and reversion at favored sites for C→U mutations (15, 25, 28, 67). Unfortunately, while extensive analysis of C→U substitution dynamics using human-derived SARS-CoV-2 strains in this and previous studies provides some insights, there is no real "negative control" to address the question of whether SARS-CoV-2 replication fitness, infectivity, and evolutionary rates in natural chains of human-to-human transmission would be enhanced in the absence of C→U hypermutation. The human A3A gene is insufficiently polymorphic in humans to enable such comparisons to be made observationally, and mouse or hamster models with their much more limited host A3 repertoires and functionally quite different APOBEC restriction systems are unlikely to provide an accurate model of human antiviral responses. However, determining the importance of APOBEC-mediated restriction of SARS-CoV-2 and indeed other RNA viruses showing C→U hypermutation is an important future path of investigation that will illuminate the role of this pathway in RNA virus defense and determining host range and zoonotic potential.

AUTHOR AFFILIATION

[1]Nuffield Department of Medicine, Peter Medawar Building for Pathogen Research, University of Oxford, Oxford, United Kingdom

AUTHOR ORCIDs

Peter Simmonds http://orcid.org/0000-0002-7964-4700

AUTHOR CONTRIBUTIONS

Peter Simmonds, Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Supervision, Validation, Visualization, Writing – original draft

ADDITIONAL FILES

The following material is available online.

Supplemental Material

Supplemental material (mBio02493-24-s0001.docx). Table S2 and Fig. S1.
Table S1 (mBio02493-24-s0002.xlsx). Listing of polymorphic sites within the 10 lineages.

REFERENCES

1. Woo PCY, de Groot RJ, Haagmans B, Lau SKP, Neuman BW, Perlman S, Sola I, van der Hoek L, Wong ACP, Yeh S-H. 2023. ICTV virus taxonomy profile: coronaviridae 2023. J Gen Virol 104. https://doi.org/10.1099/jgv.0.001843
2. Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. 2020. The proximal origin of SARS-CoV-2. Nat Med 26:450–452. https://doi.org/10.1038/s41591-020-0820-9
3. Markov PV, Ghafari M, Beer M, Lythgoe K, Simmonds P, Stilianakis NI, Katzourakis A. 2023. The evolution of SARS-CoV-2. Nat Rev Microbiol 21:361–379. https://doi.org/10.1038/s41579-023-00878-2
4. Plante JA, Liu Y, Liu J, Xia H, Johnson BA, Lokugamage KG, Zhang X, Muruato AE, Zou J, Fontes-Garfias CR, Mirchandani D, Scharton D, Bilello JP, Ku Z, An Z, Kalveram B, Freiberg AN, Menachery VD, Xie X, Plante KS, Weaver SC, Shi PY. 2021. Spike mutation D614G alters SARS-CoV-2 fitness. Nature 592:116–121. https://doi.org/10.1038/s41586-020-2895-3

5. Hou YJ, Chiba S, Halfmann P, Ehre C, Kuroda M, Dinnon KH III, Leist SR, Schäfer A, Nakajima N, Takahashi K, et al. 2020. SARS-CoV-2 D614G variant exhibits efficient replication *ex vivo* and transmission *in vivo*. Science 370:1464–1468. https://doi.org/10.1126/science.abe8499
6. Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, Hengartner N, Giorgi EE, Bhattacharya T, Foley B, Hastie KM, Parker MD, Partridge DG, Evans CM, Freeman TM, de Silva TI, McDanal C, Perez LG, Tang H, Moon-Walker A, Whelan SP, LaBranche CC, Saphire EO, Montefiori DC, Sheffield COVID-19 Genomics Group. 2020. Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus. Cell 182:812–827. https://doi.org/10.1016/j.cell.2020.06.043
7. Viana R, Moyo S, Amoako DG, Tegally H, Scheepers C, Althaus CL, Anyaneji UJ, Bester PA, Boni MF, Chand M, et al. 2022. Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa. Nature 603:679–686. https://doi.org/10.1038/s41586-022-04411-y
8. Tuekprakhon A, Nutalai R, Dijokaite-Guraliuc A, Zhou D, Ginn HM, Selvaraj M, Liu C, Mentzer AJ, Supasa P, Duyvesteyn HME, et al. 2022. Antibody escape of SARS-CoV-2 Omicron BA.4 and BA.5 from vaccine and BA.1 serum. Cell 185:2422–2433. https://doi.org/10.1016/j.cell.2022.06.005
9. Backer JA, Eggink D, Andeweg SP, Veldhuijzen IK, van Maarseveen N, Vermaas K, Vlaemynck B, Schepers R, van den Hof S, Reusken CB, Wallinga J. 2022. Shorter serial intervals in SARS-CoV-2 cases with Omicron BA.1 variant compared with Delta variant, the Netherlands, 13 to 26 December 2021. Euro Surveill 27:2200042. https://doi.org/10.2807/1560-7917.ES.2022.27.6.2200042
10. Hui KPY, Ho JCW, Cheung MC, Ng KC, Ching RHH, Lai KL, Kam TT, Gu H, Sit KY, Hsin MKY, Au TWK, Poon LLM, Peiris M, Nicholls JM, Chan MCW. 2022. SARS-CoV-2 Omicron variant replication in human bronchus and lung *ex vivo*. Nature 603:715–720. https://doi.org/10.1038/s41586-022-04479-6
11. 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
12. Di Giorgio S, Martignano F, Torcia MG, Mattiuz G, Conticello SG. 2020. Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2. Sci Adv 6:eabb5813. https://doi.org/10.1126/sciadv.abb5813
13. Klimczak LJ, Randall TA, Saini N, Li JL, Gordenin DA. 2020. Similarity between mutation spectra in hypermutated genomes of rubella virus and in SARS-CoV-2 genomes accumulated during the COVID-19 pandemic. PLoS One 15:e0237689. https://doi.org/10.1371/journal.pone.0237689
14. Mourier T, Sadykov M, Carr MJ, Gonzalez G, Hall WW, Pain A. 2021. Host-directed editing of the SARS-CoV-2 genome. Biochem Biophys Res Commun 538:35–39. https://doi.org/10.1016/j.bbrc.2020.10.092
15. Graudenzi A, Maspero D, Angaroni F, Piazza R, Ramazzotti D. 2021. Mutational signatures and heterogeneous host response revealed via large-scale characterization of SARS-CoV-2 genomic diversity. iScience 24:102116. https://doi.org/10.1016/j.isci.2021.102116
16. De Maio N, Walker CR, Turakhia Y, Lanfear R, Corbett-Detig R, Goldman N. 2021. Mutation rates and selection on synonymous mutations in SARS-CoV-2. Genome Biol Evol 13:evab087. https://doi.org/10.1093/gbe/evab087
17. Ratcliff J, Simmonds P. 2021. Potential APOBEC-mediated RNA editing of the genomes of SARS-CoV-2 and other coronaviruses and its impact on their longer term evolution. Virology (Auckl) 556:62–72. https://doi.org/10.1016/j.virol.2020.12.018
18. Harris RS, Dudley JP. 2015. APOBECs and virus restriction. Virology (Auckl) 479–480:131–145. https://doi.org/10.1016/j.virol.2015.03.012
19. Suspène R, Guétard D, Henry M, Sommer P, Wain-Hobson S, Vartanian JP. 2005. Extensive editing of both hepatitis B virus DNA strands by APOBEC3 cytidine deaminases *in vitro* and *in vivo*. Proc Natl Acad Sci U S A 102:8321–8326. https://doi.org/10.1073/pnas.0408223102
20. Vartanian JP, Guétard D, Henry M, Wain-Hobson S. 2008. Evidence for editing of human papillomavirus DNA by APOBEC3 in benign and precancerous lesions. Science 320:230–233. https://doi.org/10.1126/science.1153201
21. 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
22. 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
23. 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
24. De Maio N, Walker CR, Turakhia Y, Lanfear R, Corbett-Detig R, Goldman N. 2021. Mutation rates and selection on synonymous mutations in SARS-CoV-2. bioRxiv:2021.01.14.426705. https://doi.org/10.1101/2021.01.14.426705
25. Simmonds P, Ansari MA. 2021. Extensive C→U transition biases in the genomes of a wide range of mammalian RNA viruses; potential associations with transcriptional mutations, damage- or host-mediated editing of viral RNA. PLoS Pathog 17:e1009596. https://doi.org/10.1371/journal.ppat.1009596
26. Bradley CC, Wang C, Gordon AJE, Wen AX, Luna PN, Cooke MB, Kohrn BF, Kennedy SR, Avadhanula V, Piedra PA, Lichtarge O, Shaw CA, Ronca SE, Herman C. 2024. Targeted accurate RNA consensus sequencing (tARC-seq) reveals mechanisms of replication error affecting SARS-CoV-2 divergence. Nat Microbiol 9:1382–1392. https://doi.org/10.1038/s41564-024-01655-4
27. Nakata Y, Ode H, Kubota M, Kasahara T, Matsuoka K, Sugimoto A, Imahashi M, Yokomaku Y, Iwatani Y. 2023. Cellular APOBEC3A deaminase drives mutations in the SARS-CoV-2 genome. Nucleic Acids Res 51:783–795. https://doi.org/10.1093/nar/gkac1238
28. Kim K, Calabrese P, Wang S, Qin C, Rao Y, Feng P, Chen XS. 2022. The roles of APOBEC-mediated RNA editing in SARS-CoV-2 mutations, replication and fitness. Res Sq:rs.3.rs-1524060. https://doi.org/10.21203/rs.3.rs-1524060/v1
29. 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
30. Ahmad W, Ahmad S, Basha R. 2022. Analysis of the mutation dynamics of SARS-CoV-2 genome in the samples from Georgia State of the United States. Gene 841:146774. https://doi.org/10.1016/j.gene.2022.146774
31. Bloom JD, Neher RA. 2023. Fitness effects of mutations to SARS-CoV-2 proteins. Virus Evol 9:vead055. https://doi.org/10.1093/ve/vead055
32. 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
33. Kumar S, Stecher G, Tamura K. 2016. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol 33:1870–1874. https://doi.org/10.1093/molbev/msw054
34. Huston NC, Wan H, Strine MS, de Cesaris Araujo Tavares R, Wilen CB, Pyle AM. 2021. Comprehensive *in vivo* secondary structure of the SARS-CoV-2 genome reveals novel regulatory motifs and mechanisms. Mol Cell 81:584–598. https://doi.org/10.1016/j.molcel.2020.12.041
35. Manfredonia I, Nithin C, Ponce-Salvatierra A, Ghosh P, Wirecki TK, Marinus T, Ogando NS, Snijder EJ, van Hemert MJ, Bujnicki JM, Incarnato D. 2020. Genome-wide mapping of SARS-CoV-2 RNA structures identifies therapeutically-relevant elements. Nucleic Acids Res 48:12436–12452. https://doi.org/10.1093/nar/gkaa1053
36. Lan TCT, Allan MF, Malsick LE, Woo JZ, Zhu C, Zhang F, Khandwala S, Nyeo SSY, Sun Y, Guo JU, Bathe M, Näär A, Griffiths A, Rouskin S. 2022. Secondary structural ensembles of the SARS-CoV-2 RNA genome in infected cells. Nat Commun 13:1128. https://doi.org/10.1038/s41467-022-28603-2
37. Woo PCY, Wong BHL, Huang Y, Lau SKP, Yuen K-Y. 2007. Cytosine deamination and selection of CpG suppressed clones are the two major independent biological forces that shape codon usage bias in coronaviruses. Virology (Auckl) 369:431–442. https://doi.org/10.1016/j.virol.2007.08.010
38. 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
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. Vijgen L, Keyaerts E, Moës E, Thoelen I, Wollants E, Lemey P, Vandamme A-M, Van Ranst M. 2005. Complete genomic sequence of human coronavirus OC43: molecular clock analysis suggests a relatively recent zoonotic coronavirus transmission event. J Virol 79:1595–1604. https://doi.org/10.1128/JVI.79.3.1595-1604.2005
41. Lamb KD, Luka MM, Saathoff M, Orton RJ, Phan MVT, Cotten M, Yuan K, Robertson DL. 2024. Mutational signature dynamics indicate SARS-CoV-2's evolutionary capacity is driven by host antiviral molecules. PLoS Comput Biol 20:e1011795. https://doi.org/10.1371/journal.pcbi.1011795
42. Bloom JD, Beichman AC, Neher RA, Harris K. 2023. Evolution of the SARS-CoV-2 mutational spectrum. Mol Biol Evol 40:msad085. https://doi.org/10.1093/molbev/msad085
43. Poulain F, Lejeune N, Willemart K, Gillet NA. 2020. Footprint of the host restriction factors APOBEC3 on the genome of human viruses. PLoS Pathog 16:e1008718. https://doi.org/10.1371/journal.ppat.1008718
44. Anwar F, Davenport MP, Ebrahimi D. 2013. Footprint of APOBEC3 on the genome of human retroelements. J Virol 87:8195–8204. https://doi.org/10.1128/JVI.00298-13
45. 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
46. Sharma S, Baysal BE. 2017. Stem-loop structure preference for site-specific RNA editing by APOBEC3A and APOBEC3G. PeerJ 5:e4136. https://doi.org/10.7717/peerj.4136
47. Powell LM, Wallis SC, Pease RJ, Edwards YH, Knott TJ, Scott J. 1987. A novel form of tissue-specific RNA processing produces apolipoprotein-B48 in intestine. Cell 50:831–840. https://doi.org/10.1016/0092-8674(87)90510-1
48. Chen S-H, Habib G, Yang C-Y, Gu Z-W, Lee BR, Weng S-A, Silberman SR, Cai S-J, Deslypere JP, Rosseneu M, Gotto AM, Li W-H, Chan L. 1987. Apolipoprotein B-48 is the product of a messenger RNA with an organ-specific in-frame stop codon. Science 238:363–366. https://doi.org/10.1126/science.3659919
49. Perelygina L, Chen MH, Suppiah S, Adebayo A, Abernathy E, Dorsey M, Bercovitch L, Paris K, White KP, Krol A, Dhossche J, Torshin IY, Saini N, Klimczak LJ, Gordenin DA, Zharkikh A, Plotkin S, Sullivan KE, Icenogle J. 2019. Infectious vaccine-derived rubella viruses emerge, persist, and evolve in cutaneous granulomas of children with primary immunodeficiencies. PLoS Pathog 15:e1008080. https://doi.org/10.1371/journal.ppat.1008080
50. Simmonds P. 2020. Pervasive RNA secondary structure in the genomes of SARS-CoV-2 and other coronaviruses. mBio 11:e01661-20. https://doi.org/10.1128/mBio.01661-20
51. Kurkowiak M, Fletcher S, Daniels A, Mozolewski P, Silvestris DA, Król E, Marek-Trzonkowska N, Hupp T, Tait-Burkard C. 2023. Differential RNA editing landscapes in host cell versus the SARS-CoV-2 genome. iScience 26:108031. https://doi.org/10.1016/j.isci.2023.108031
52. Gordon CJ, Tchesnokov EP, Schinazi RF, Götte M. 2021. Molnupiravir promotes SARS-CoV-2 mutagenesis via the RNA template. J Biol Chem 297:100770. https://doi.org/10.1016/j.jbc.2021.100770
53. Kabinger F, Stiller C, Schmitzová J, Dienemann C, Kokic G, Hillen HS, Höbartner C, Cramer P. 2021. Mechanism of molnupiravir-induced SARS-CoV-2 mutagenesis. Nat Struct Mol Biol 28:740–746. https://doi.org/10.1038/s41594-021-00651-0
54. Butler CC, Hobbs FDR, Gbinigie OA, Rahman NM, Hayward G, Richards DB, Dorward J, Lowe DM, Standing JF, Breuer J, et al. 2023. Molnupiravir plus usual care versus usual care alone as early treatment for adults with COVID-19 at increased risk of adverse outcomes (PANORAMIC): an open-label, platform-adaptive randomised controlled trial. Lancet 401:281–293. https://doi.org/10.1016/S0140-6736(22)02597-1
55. Donovan-Banfield I, Penrice-Randal R, Goldswain H, Rzeszutek AM, Pilgrim J, Bullock K, Saunders G, Northey J, Dong X, Ryan Y, et al. 2022. Characterisation of SARS-CoV-2 genomic variation in response to molnupiravir treatment in the AGILE Phase IIa clinical trial. Nat Commun 13:7284. https://doi.org/10.1038/s41467-022-34839-9
56. Sanderson T, Hisner R, Donovan-Banfield I, Hartman H, Løchen A, Peacock TP, Ruis C. 2023. A molnupiravir-associated mutational signature in global SARS-CoV-2 genomes. Nature 623:594–600. https://doi.org/10.1038/s41586-023-06649-6
57. McLaughlin RN, Gable JT, Wittkopp CJ, Emerman M, Malik HS. 2016. Conservation and innovation of APOBEC3A restriction functions during primate evolution. Mol Biol Evol 33:1889–1901. https://doi.org/10.1093/molbev/msw070
58. 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
59. LaRue RS, Jónsson SR, Silverstein KA, Lajoie M, Bertrand D, El-Mabrouk N, Hötzel I, Andrésdóttir V, Smith TP, Harris RS. 2008. The artiodactyl *APOBEC3* innate immune repertoire shows evidence for a multi-functional domain organization that existed in the ancestor of placental mammals. BMC Mol Biol 9:104. https://doi.org/10.1186/1471-2199-9-104
60. Chargaff E, Lipshitz R, Green C. 1952. Composition of the desoxypentose nucleic acids of four genera of sea-urchin. J Biol Chem 195:155–160.
61. Malim MH. 2009. APOBEC proteins and intrinsic resistance to HIV-1 infection. Philos Trans R Soc Lond B Biol Sci 364:675–687. https://doi.org/10.1098/rstb.2008.0185
62. O'Toole Á, Neher RA, Ndodo N, Borges V, Gannon B, Gomes JP, Groves N, King DJ, Maloney D, Lemey P, Lewandowski K, Loman N, Myers R, Omah IF, Suchard MA, Worobey M, Chand M, Ihekweazu C, Ulaeto D, Adetifa I, Rambaut A. 2023. APOBEC3 deaminase editing in mpox virus as evidence for sustained human transmission since at least 2016. Science 382:595–600. https://doi.org/10.1126/science.adg8116
63. Gaba A, Flath B, Chelico L. 2021. Examination of the APOBEC3 barrier to cross species transmission of primate lentiviruses. Viruses 13:1084. https://doi.org/10.3390/v13061084
64. Kosugi Y, Uriu K, Suzuki N, Yamamoto K, Nagaoka S, Kimura I, Konno Y, Aso H, Willett BJ, Kobayashi T, Koyanagi Y, Ueda MT, Ito J, Sato K. 2021. Comprehensive investigation on the interplay between feline APOBEC3Z3 proteins and feline immunodeficiency virus Vif proteins. J Virol 95:e0017821. https://doi.org/10.1128/JVI.00178-21
65. Ratcliff J, Simmonds P. 2023. The roles of nucleic acid editing in adaptation of zoonotic viruses to humans. Curr Opin Virol 60:101326. https://doi.org/10.1016/j.coviro.2023.101326
66. Wang J, Wu L, Pu X, Liu B, Cao M. 2023. Evidence supporting that C-to-U RNA editing is the major force that drives SARS-CoV-2 evolution. J Mol Evol 91:214–224. https://doi.org/10.1007/s00239-023-10097-1
67. van Dorp L, Acman M, Richard D, Shaw LP, Ford CE, Ormond L, Owen CJ, Pang J, Tan CCS, Boshier FAT, Ortiz AT, Balloux F. 2020. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infect Genet Evol 83:104351. https://doi.org/10.1016/j.meegid.2020.104351