Secondary structure of the SARS-CoV-2 genome is predictive of nucleotide substitution frequency

Zach Hensel[1]

[1] ITQB NOVA, Universidade NOVA de Lisboa, Lisbon, Av. da República, 2780-157, Oeiras, Portugal
Corresponding author: zach.hensel@itqb.unl.pt

Abstract

Accurate estimation of the effects of mutations on SARS-CoV-2 viral fitness can inform public-health responses such as vaccine development and predicting the impact of a new variant; it can also illuminate biological mechanisms including those underlying the emergence of variants of concern (Carabelli *et al.*, 2023). Recently, Lan *et al.* reported a model of SARS-CoV-2 secondary structure and its underlying dimethyl sulfate (DMS) reactivity data (Lan *et al.*, 2022). I investigated whether base reactivities and secondary structure models derived from them can explain some variability in the frequency of observing different nucleotide substitutions across millions of patient sequences in the SARS-CoV-2 phylogenetic tree. Nucleotide basepairing was compared to the estimated "mutational fitness" of substitutions, a measurement of the difference between a substitution's observed and expected frequency that is correlated with other estimates of viral fitness (Bloom and Neher, 2023). This comparison revealed that secondary structure is often predictive of substitution frequency, with significant decreases in substitution frequencies at basepaired positions. Focusing on the mutational fitness of C→U, the most common type of substitution, I describe C→U substitutions at basepaired positions that characterize major SARS-CoV-2 variants; such mutations may have a greater impact on fitness than appreciated when considering substitution frequency alone.

1. Introduction

While investigating the significance of the substitution C29095U, detected in a familial cluster of SARS-CoV-2 infections (Chan *et al.*, 2020), I hypothesized that this synonymous substitution reflected the high frequency of C→U substitution during the pandemic (De Maio *et al.*, 2021). Specifically, frequent C29095U substitution had previously complicated attempts to infer recombinant genomes (VanInsberghe *et al.*, 2021). Preliminary investigation revealed that C29095U was the fourth most frequent C→U substitution, occurring almost seven times as often as a typical C→U substitution (Bloom and Neher, 2023). While there was no clear reason for selection of this synonymous substitution, C29095 was found to be unpaired in a secondary structure model (Sun *et al.*, 2021). The 5′ context of C29095 is UC, and additional SARS-CoV-2 secondary structure models that I consulted (see below) agree that C29095 occurs as a single unpaired base in an asymmetrical internal loop, opposite three unpaired bases. Although one study found that exogenous expression of APOBEC3A accelerated C→U substitution in SARS-CoV-2-infected cells in the UC context (Nakata *et al.*, 2023), the most frequently mutated positions were in the context of larger unpaired regions.

I hypothesized that non-enzymatic deamination may be more frequent for unpaired cytosine residues, which was supported by a previous analysis with a resolution of approximately 300 nucleotides (Li *et al.*, 2023) and by a comparison of sequence diversity in early SARS-CoV-2 genomes to an RNA folding prediction (Simmonds, 2020). However, these studies differed in their interpretations of relatively high C→U substitution frequencies at unpaired positions, leaving it unresolved whether this was primarily driven by higher mutation rates at unpaired positions or fitness costs of mutations disrupting secondary structure. To determine whether

secondary structure was in fact correlated with mutation frequency at single-nucleotide resolution and to extend this analysis beyond C→U substitutions, the dataset reported by Lan *et al.* was compared to the mutational fitness estimates reported by Bloom and Neher (Bloom and Neher, 2023). Note that "mutational fitness" is not a measurement of viral fitness per se; rather, it is an estimate made assuming that the expected frequencies of neutral mutations are determined only by the type of substitution (with C→U being much more frequent than all other types of substitutions).

2. Methods

The data compared in this study consisted of estimated mutational fitness for the SARS-CoV-2 genome as reported by Bloom and Neher (Supplementary Data: `ntmut_fitness_all.csv`; November 2024 dataset) as well as population-averaged dimethyl sulfate (DMS) reactivities for SARS-CoV-2-infected Huh7 cells and the corresponding secondary structure model reported by Lan *et al.* (Supplementary Data 7 and 8). Two other secondary structure models based upon SHAPE reactivity data were also used (Huston *et al.*, 2021; Manfredonia *et al.*, 2020). Note that the estimated mutational fitness is logarithmically related to the ratio of the observed and expected number of occurrences of a nucleotide substitution, with large and asymmetric differences in the frequencies of different types of synonymous substitutions (De Maio *et al.*, 2021). Additionally, note that DMS data was obtained in experiments using the WA1 strain in Lineage A, which differs from the more common Lineage B at 3 positions and could have different secondary structure. Furthermore, mutational fitness is estimated from the phylogenetic tree of published sequences (the public UShER tree (Turakhia *et al.*, 2021) additionally curated to filter likely artifacts such as branches with numerous reversions) that are typically far more divergent and subsequently will have somewhat different secondary structures. Since the dataset used for mutational fitness aggregates data across viral clades, my analysis will not capture secondary structure variation between clades or indels and masked sites that were not considered in that analysis (Bloom and Neher, 2023). I focused on the most common types of nucleotide substitutions: those comprising approximately 5% or more of total substitutions.

3. Results

There was a significant increase in synonymous substitution frequencies at unpaired positions for C→U, G→U, C→A, and U→C, but not for A→G or G→A ($p < 0.05$; Tukey's range test with Bonferroni correction; A→U, G→C, and C→G were also significant in an unplanned analysis of all substitution types). For all substitution types with significant differences, unpaired substitution frequencies were higher than basepaired substitution frequencies. The largest effects were observed for C→U and G→U (Figure 1). In this secondary structure model, there is basepairing for 60% and 73% of C and G positions, respectively (limited to those covered in the mutational fitness analysis). The largest impacts on median estimated mutational fitness were observed for synonymous C→U and G→U mutations, with apparent fitness higher at unpaired positions than at paired positions by 1.44 and 1.25, respectively. Expressed in terms of substitution frequency rather than mutational fitness (Figure 1), the frequency of synonymous C→U and G→U substitutions were about four times higher at unpaired positions than at basepaired positions. Together, this demonstrates a meaningful impact of secondary structure on substitution frequencies.

Similar results were reported by two studies after this study was first preprinted. One study identified a basepairing dependence for C→U substitutions during serial passage of different SARS-CoV-2 variants and considered mutational fitness estimates from Bloom and Neher for more rare substitution types (Gout *et al.*, 2024). In another analysis that analyzed recurrent substitutions in multiple lineages, little difference was observed overall for C→U substitutions in different 5′ and 3′ contexts (Simmonds, 2024). It was observed that only the few C→U substitutions that appeared with the highest frequencies had a strong preference for 5′ U context.

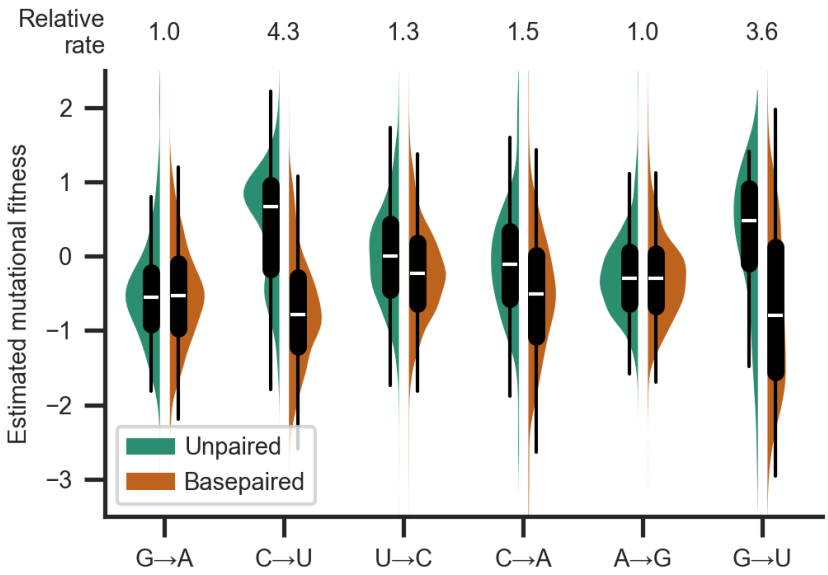

Figure 1: Basepairing is predictive of synonymous substitution frequency. Distribution of frequencies of synonymous substitutions for the most common substitutions (each approximately corresponding to 5% or more of observed substitutions), expressed as the estimated mutational fitness, which is a logarithmic comparison of the observed versus the expected number of occurrences of each type of substitution in the SARS-CoV-2 phylogenetic tree (Bloom and Neher, 2023). Distributions are grouped by substitution type and whether positions are basepaired in a full-genome secondary structure model of SARS-CoV-2 in Huh7 cells (Lan *et al.*, 2022). The median relative substitution rate (ratio of median rates of substitutions at unpaired and paired positions) is shown above each type of substitution.

Figure S1 shows that the degree of this 5′ U preference becomes more clear when limiting analysis to synonymous substitutions and separately analyzing basepaired positions (which show no 5′ preference) and unpaired positions. At unpaired positions, synonymous C→U substitutions with 5′ U context have a higher apparent fitness (0.83 [0.79, 0.89], 95% confidence interval from 1000 bootstrapped samples) than with 5′ G (0.15 [0.01, 0.50]), 5′ A (0.55 [0.45, 0.68]), or 5′ C (0.57 [0.33, 0.69]). Conversely, synonymous C→U substitution rates with 3′ G context were elevated at both unpaired and paired positions (Figure S2), suggesting some CpG suppression with little dependence on basepairing. Notably, the previously mentioned C29095 is both unpaired and appears in the context UCG.

Basepairing in the secondary structure model appeared to be more predictive of estimated mutational fitness than average DMS reactivity, with correlation coefficients of 0.59 [0.55, 0.62] and 0.45 [0.40, 0.49], respectively, for synonymous C→U substitutions (point biserial and Spearman correlation coefficients; 95% confidence intervals from 1000 bootstrapped samples). However, DMS reactivity was more strongly correlated with estimated mutational fitness than basepairing when analysis was limited to positions with detectable DMS reactivity (0.77 [0.74, 0.79], excluding the sites plotted at the minimum measured value of 0.00012). This suggests that considering structural heterogeneity captured in base reactivity datasets could improve predictions of substitution rates. Correlation coefficients remained significant, but were reduced (0.17 [0.14, 0.20] and 0.13 [0.09, 0.16]) when considering nonsynonymous C→U substitutions (Figure 2, left), consistent with larger and often negative effects of nonsynonymous mutations on viral fitness (Bloom and Neher, 2023).

Since basepairing is correlated with substitution rates, the degree of correlation between secondary structure models and substitution rates can be used to compare models that differ in underlying experimental methods and analysis. In light of this, secondary structure models derived from two DMS reactivity datasets (from SARS-CoV-2 isolate WA1 infection of Huh7 and Vero cells) (Lan *et al.*, 2022) and two SHAPE reactivity datasets (SARS-CoV-2 isolates WA1 and

Leiden0002 infection of Huh7.5 and Vero cells, respectively) (Huston *et al.*, 2021; Manfredonia *et al.*, 2020) were compared by degree of correlation with estimated mutational fitness. Correlations between basepairing and synonymous C→U substitution rates were highest for models derived from DMS data (0.59 [0.55, 0.62] and 0.63 [0.60, 0.67] for Huh7 and Vero datasets, respectively). Models derived from SHAPE data had modestly lower correlations (0.55 [0.50, 0.59] and 0.50 [0.46, 0.54] for Huh7 and Vero datasets, respectively). However, the methods used to construct each of these four models have difference beyond base reactivity probes, cell type, and viral isolate, so small differences in correlation cannot be attributed to a single factor.

Thus, the high mutation rate of C29095U could be more accurately estimated by considering DMS reactivity than basepairing alone (Figure 2) since it is in the 99th percentile of DMS activity for positions with synonymous C→U substitutions. No major difference in this trend was observed across the SARS-CoV-2 genome. As a first-order approximation, two constants were calculated to equalize median mutational fitness for synonymous substitutions at basepaired, unpaired, and all positions. An "adjusted mutational fitness" can then be calculated for C→U substitutions by incrementing mutational fitness by +0.32 at basepaired position and by −1.13 at unpaired positions (results were similar when considering fourfold degenerate positions rather than all synonymous substitutions). Scatter plots compare DMS reactivity to estimated mutational fitness at positions with nonsynonymous C→U substitutions before and after applying this coarse adjustment (Figure 2, left and right).

This adjustment equalizes median mutational fitness for synonymous C→U substitutions (Figure S3), and remaining correlation between "adjusted mutational fitness" and DMS reactivity (0.18 [0.13, 0.22]) shows that this coarse adjustment could be improved with more complex

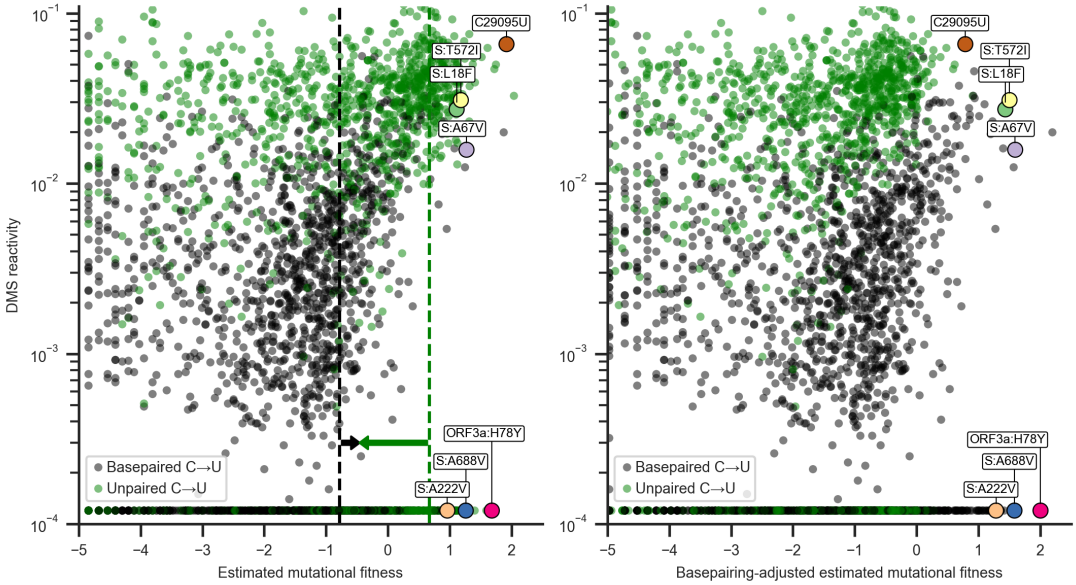

Figure 2: Estimated mutational fitness correlates with secondary structure for nonsynonymous C→U substitutions. Scatter plots compare mutational fitness to average DMS reactivity for positions with potential nonsynonymous C→U substitutions. The minimum observed DMS reactivity value is assigned to positions lacking data. Points are colored by basepairing in the full genome secondary structure model. Nonsynonymous C→U substitutions at basepaired positions are highlighted which rank highly for mutational fitness and characterize major SARS-CoV-2 lineages. Synonymous C29095U at an unpaired position is also highlighted. Left: Estimated mutational fitness based only on observed versus expected occurrences of C→U at each position. Dashed lines indicate the median estimated mutational fitness for synonymous substitutions at paired and unpaired positions. Arrows indicate the magnitudes of adjustments made to mutational fitness that result in median fitness of synonymous substitutions at paired (+0.32) and unpaired (−1.13) positions identical to the unadjusted median for all synonymous substitutions (−0.46). Right: Mutational fitness adjusted by constants derived from the medians of mutational fitness for synonymous substitutions at basepaired, unpaired, and all potential C→U positions.

analysis. However, a binary model of basepairing alone is sufficient to capture much of the variation in synonymous substitution rates and could be extended to other substitution types with rates depending on substitution types that lack reactivity data.

For a preliminary estimate of whether nonsynonymous C→U substitutions at basepaired positions are prone to underestimation of mutational fitness, I tested the hypothesis that C→U having highly ranked fitness at basepaired positions would be mutations that characterize significant SARS-CoV-2 lineages (arbitrarily defined as having 5% prevalence in the one-week average of global sequences on the CoV-Spectrum website (Chen *et al.*, 2022) at any time during the pandemic). This was the case for 6 of the top 15 C→U substitutions at basepaired positions; their encoded mutations are shown in Figure 2. Top-ranked mutations characterize Omicron lineages BA.1, one of the first recognized recombinant lineages XB, Gamma P.1, and lineage JN.1.7, which had over 20% global prevalence in spring 2024. Half of these mutations have relatively high DMS reactivity for basepaired positions and half have very low DMS reactivity. By comparison, the synonymous substitution C29095U at an unpaired position has very high estimated mutational fitness and DMS reactivity. Despite having a higher median estimated mutational fitness (1.41 vs. 1.10), only 3 of the top 15 nonsynonymous C→U at unpaired positions define major lineages (BQ.1.1, JN.1.8.1, and BA.2.86.1).

Of particular note is C22227T at a basepaired position encoding the spike A222V mutation. This was one of the mutations that characterized the B.1.177 lineage. This lineage, also known as EU1, characterized a majority of sequences in Spain in summer 2020 and eventually in several other countries in Europe prior to the emergence of the Alpha variant. However, it was unclear whether this lineage had higher fitness than other lineages or if A222V specifically conferred a fitness advantage (Hodcroft *et al.*, 2021). Further investigation as well as its recurrence in the major Delta sublineage AY.4.2 provided additional evidence of an increase in viral fitness and suggested molecular mechanisms (Ginex *et al.*, 2022). Here, I focus on C→U substitutions for comparison to DMS reactivity data, but I also note that top-ranked G→U substitutions at basepaired positions are rich in mutations to ORF3a and also include mutations that characterize variants of concern, such as nucleocapsid D377Y in Delta. Lastly, note that, following the coarse adjustment for basepairing inferred from synonymous substitutions, nonsynonymous C→U substitutions characterizing major variants now have some of the highest estimated mutational fitness for C→U substitutions (Figure 2, right).

4. Discussion

This analysis shows that it is informative to combine apparent viral fitness, inferred from millions of patient SARS-CoV-2 sequences published during the pandemic, with accurate secondary structure measurements. It is important to remember that apparent "mutational fitness" results from a combination of the rates at which diversity is generated and the subsequent selection processes. Importantly, genome secondary structure can impact both. However, even the unprecedented density of sampling SARS-CoV-2 genomes has been insufficient to reliably infer fitness impacts of single mutations more directly from dynamics subsequent to occurrences in the SARS-CoV-2 phylogenetic tree (Bloom and Neher, 2023; Richard *et al.*, 2021).

Further investigation into phenomena reported here, such as the lack of apparent secondary structure dependence for A→G and G→A substitutions, could inform investigation of underlying mutation and selection mechanisms. An analysis focused on basepairing dependence of C→U substitution frequencies considered that this may be driven by conservation of secondary structure (Gout *et al.*, 2024). However, the lack of basepairing dependence for A→G and G→A substitutions supports the hypothesis that this effect, on average, arises from differences in chemical reaction rates (Li *et al.*, 2023). Nonetheless, future analyses could consider whether substitution rates have different basepairing dependencies in regions in which synonymous mutations exhibit fitness reductions (Bloom and Neher, 2023). The extents of enzymatic and non-enzymatic mutation processes driving differences in substitution rates re-

mains unclear (Simmonds, 2024). Comparing the degree to which rates depend on secondary structure could provide insight to support different mutation mechanisms. A recent *in vitro* study suggested that oxidation of guanine can give rise to G→U substitutions (Akagawa *et al.*, 2024), which like C→U substitutions, are both very common in human SARS-CoV-2 infections and exhibit strong basepairing dependence.

I suggest that secondary structure, along with other data correlating with substitution frequency, can be used to refine estimates of mutational fitness both for future SARS-CoV-2 variants and for other viruses. More sophisticated analysis can incorporate structural heterogeneity (Lan *et al.*, 2022) as well as local sequence context (Lamb *et al.*, 2024). Figure S3 shows how the coarse adjustment described in this paper does not capture all information available in DMS reactivity data, and DMS interrogates adenine and cytosine basepairing while SHAPE reagents can be used to investigate backbone flexibility for all nucleotides. Clearly, datasets from different experiments using different chemical probes can be combined to produce more accurate estimates of apparent mutational fitness. Furthermore, additional measurements of secondary structure for genomes of new variants or modeling may reveal significant changes to secondary structure since 2020. This could be evident in *in vitro* substitution rates for different SARS-CoV-2 variants (Gout *et al.*, 2024). Lastly, secondary structures determined from *in vitro* chemical probing experiments may differ from secondary structures *in vivo*. For the spike protein, the correlation between estimated mutational fitness and pseudovirus entry quantified by deep mutational scanning serves as one metric that can be used to optimize models (Lamb *et al.*, 2024). However, it is critical to evaluate uncertainty in any model estimating fitness of a new variant. Initial estimates can be refined by rapid *in vitro* characterization and continued genomic surveillance.

Data availability

Data analyzed in this manuscript and a Python notebook to reproduce analysis are available at `https://github.com/smmlab/SARS2-fitness-secondary-structure`.

Acknowledgements

Although I did not directly access any genome sequencing databases for this work, I am grateful to the patients who volunteered samples, and to the clinicians, technicians, and teams behind the databases who have made sequencing data available so that this work is possible. I thank Erol Akcay, Alexander Crits-Christoph, Florence Débarre, Ryan Hisner, and James Yates for critical comments on the manuscript and discussions. ZH is supported by Fundação para a Ciência e a Tecnologia (FCT) through MOSTMICRO-ITQB (DOI 10.54499/UIDB/04612/2020; DOI 10.54499/UIDP/04612/2020) and LS4FUTURE Associated Laboratory (DOI 10.54499/LA/P/0087/2020).

References

Akagawa M, Sugasawa K, Ura K, Sassa A. 2024. Impact of an Oxidative RNA Lesion on in vitro Replication Catalyzed by SARS-CoV-2 RNA-dependent RNA Polymerase. p. 2024.10.13.618053.

Bloom JD, Neher RA. 2023. Fitness effects of mutations to SARS-CoV-2 proteins. Virus Evolution. 9:vead055.

Carabelli AM, Peacock TP, Thorne LG, Harvey WT, Hughes J, de Silva TI, Peacock SJ, Barclay WS, de Silva TI, Towers GJ *et al.* 2023. SARS-CoV-2 variant biology: Immune escape, transmission and fitness. Nat Rev Microbiol. 21:162–177.

Chan JFW, Yuan S, Kok KH, To KKW, Chu H, Yang J, Xing F, Liu J, Yip CCY, Poon RWS *et al.* 2020. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: A study of a family cluster. The Lancet. 395:514–523.

Chen C, Nadeau S, Yared M, Voinov P, Xie N, Roemer C, Stadler T. 2022. CoV-Spectrum: Analysis of globally shared SARS-CoV-2 data to identify and characterize new variants. Bioinformatics. 38:1735–1737.

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 Biology and Evolution. 13:evab087.

Ginex T, Marco-Marín C, Wieczór M, Mata CP, Krieger J, Ruiz-Rodriguez P, López-Redondo ML, Francés-Gómez C, Melero R, Sánchez-Sorzano CÓ *et al.* 2022. The structural role of SARS-CoV-2 genetic background in the emergence and success of spike mutations: The case of the spike A222V mutation. PLOS Pathogens. 18:e1010631.

Gout JF, Symons J, Chung C, Verheijen L, Shemtov S, de Jong D, Amatngalim G, Nijhuis M, Vermulst M. 2024. The mutational landscape of SARS-CoV-2 provides new insight into viral evolution and fitness. .

Hodcroft EB, Zuber M, Nadeau S, Vaughan TG, Crawford KHD, Althaus CL, Reichmuth ML, Bowen JE, Walls AC, Corti D *et al.* 2021. Spread of a SARS-CoV-2 variant through Europe in the summer of 2020. Nature. 595:707–712.

Huston NC, Wan H, Strine MS, Tavares RdCA, Wilen CB, Pyle AM. 2021. Comprehensive in vivo secondary structure of the SARS-CoV-2 genome reveals novel regulatory motifs and mechanisms. Molecular Cell. 81:584–598.e5.

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 Computational Biology. 20:e1011795.

Lan TCT, Allan MF, Malsick LE, Woo JZ, Zhu C, Zhang F, Khandwala S, Nyeo SSY, Sun Y, Guo JU *et al.* 2022. Secondary structural ensembles of the SARS-CoV-2 RNA genome in infected cells. Nat Commun. 13:1128.

Li Y, Hou F, Zhou M, Yang X, Yin B, Jiang W, Xu H. 2023. C-to-U RNA deamination is the driving force accelerating SARS-CoV-2 evolution. Life Science Alliance. 6.

Manfredonia I, Nithin C, Ponce-Salvatierra A, Ghosh P, Wirecki TK, Marinus T, Ogando NS, Snijder EJ, van Hemert MJ, Bujnicki JM *et al.* 2020. Genome-wide mapping of SARS-CoV-2 RNA structures identifies therapeutically-relevant elements. Nucleic Acids Research. 48:12436–12452.

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 Research. 51:783–795.

Richard D, Shaw LP, Lanfear R, Corbett-Detig R, Hinrichs A, McBroome J, Turakhia Y, Acman M, Owen CJ, Tan CC *et al.* 2021. A phylogeny-based metric for estimating changes in transmissibility from recurrent mutations in SARS-CoV-2. p. 2021.05.06.442903.

Simmonds P. 2020. Pervasive RNA Secondary Structure in the Genomes of SARS-CoV-2 and Other Coronaviruses. mBio. 11:10.1128/mbio.01661–20.

Simmonds P. 2024. C→U transition biases in SARS-CoV-2: Still rampant 4 years from the start of the COVID-19 pandemic. mBio. 0:e02493–24.

Sun L, Li P, Ju X, Rao J, Huang W, Ren L, Zhang S, Xiong T, Xu K, Zhou X *et al.* 2021. In vivo structural characterization of the SARS-CoV-2 RNA genome identifies host proteins vulnerable to repurposed drugs. Cell. 184:1865–1883.e20.

Turakhia Y, Thornlow B, Hinrichs AS, De Maio N, Gozashti L, Lanfear R, Haussler D, Corbett-Detig R. 2021. Ultrafast Sample placement on Existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. Nat Genet. 53:809–816.

VanInsberghe D, Neish AS, Lowen AC, Koelle K. 2021. Recombinant SARS-CoV-2 genomes circulated at low levels over the first year of the pandemic. Virus Evol. 7:veab059.

Supplementary figures and tables

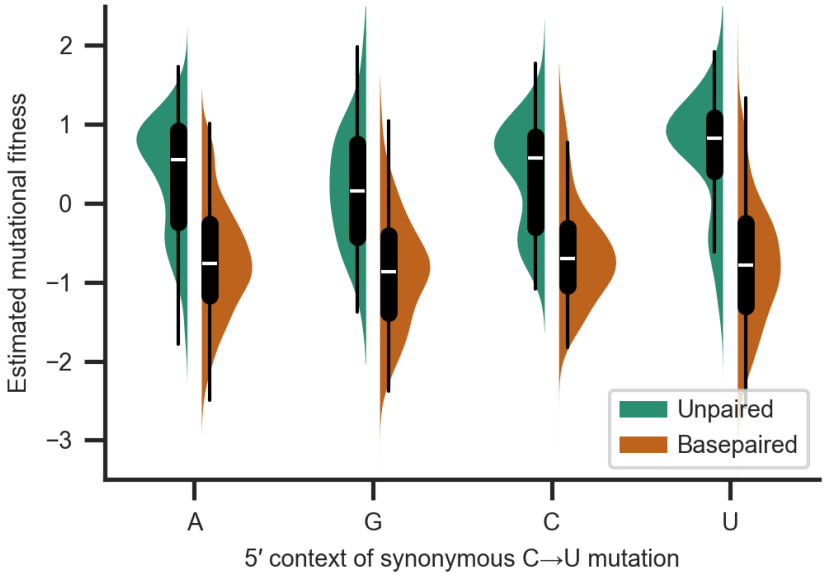

Figure S1: Basepairing dependence of effect of 5′ context on synonymous C→U substitution frequency. Distribution of frequencies of synonymous substitutions for C→U substitutions, expressed as the estimated mutational fitness, which is a logarithmic comparison of the observed versus the expected number of occurrences of each type of substitution in the SARS-CoV-2 phylogenetic tree. Distributions are grouped by the identity at the 5′ position and by whether positions are basepaired in a full-genome secondary structure model of SARS-CoV-2.

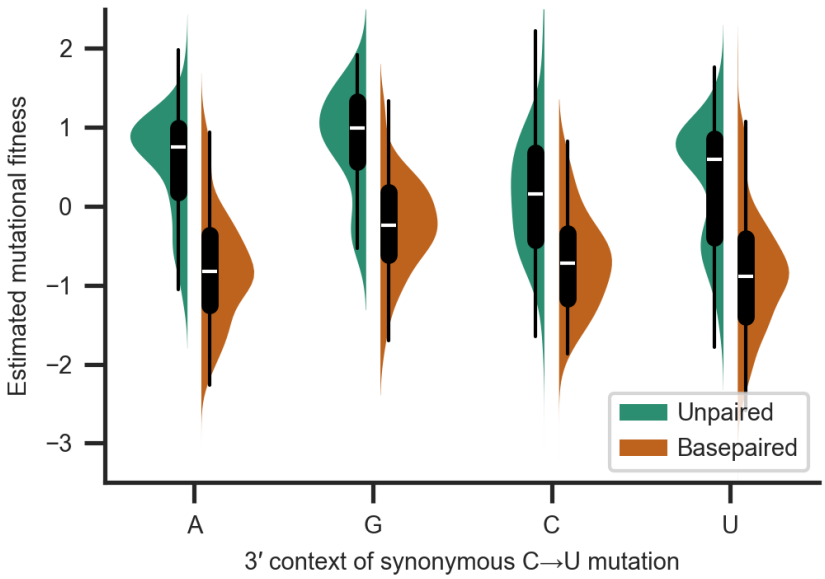

Figure S2: Basepairing dependence of effect of 3′ context on synonymous C→U substitution frequency. Distribution of frequencies of synonymous substitutions for C→U substitutions, expressed as the estimated mutational fitness, which is a logarithmic comparison of the observed versus the expected number of occurrences of each type of substitution in the SARS-CoV-2 phylogenetic tree. Distributions are grouped by the identity at the 3′ position and by whether positions are basepaired in a full-genome secondary structure model of SARS-CoV-2.

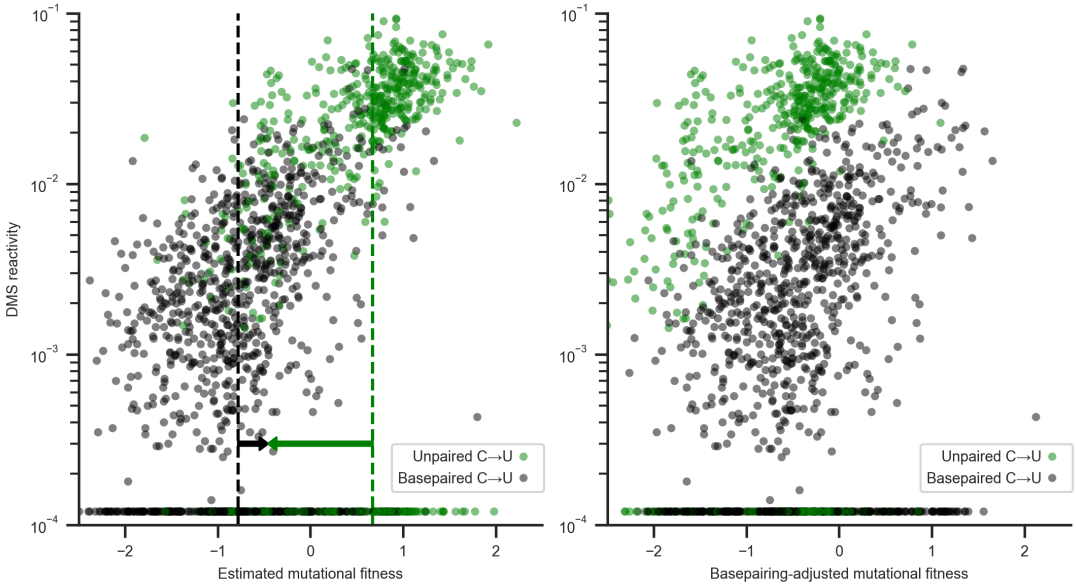

Figure S3: Estimated mutational fitness correlates with secondary structure for synonymous C→U substitutions. Scatter plots compare mutational fitness to average DMS reactivity for positions with potential synonymous C→U substitutions. The minimum observed DMS reactivity value is assigned to positions lacking data. Points are colored by basepairing in the full genome secondary structure model. Left: Estimated mutational fitness based only on observed versus expected occurrences of C→U at each position. Dashed lines indicate the median estimated mutational fitness for synonymous substitutions at paired and unpaired positions (identical to Figure 2). Arrows indicate the magnitudes of adjustments made to mutational fitness that result in median fitness of synonymous substitutions at paired (+0.32) and unpaired (−1.13) positions identical to the unadjusted median for all synonymous substitutions (−0.46). Right: Mutational fitness adjusted by constants derived from the medians of mutational fitness for synonymous substitutions at basepaired, unpaired, and all potential C→U positions.