# 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

* Correspondence: 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[1]. Recently, Lan et al reported a high-quality model of SARS-CoV-2 secondary structure and its underlying dimethyl sulfate (DMS) reactivity data[2]. I investigated whether secondary structure can explain some variability in the frequency of observing different nucleotide substitutions across millions of patient sequences in the SARS-CoV-2 phylogenetic tree[3]. 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[4]. 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→T, the most common type of substitution, I describe C→T 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.

## Introduction

While investigating the significance of the substitution C29095T, detected in a familial cluster of SARS-CoV-2 infections[5], I hypothesized that this synonymous substitution reflected the high frequency of C→T substitution during the pandemic[6]. Specifically, frequent C29095T substitution had previously complicated attempts to infer recombinant genomes[7]. Preliminary investigation of C29095T revealed that it was the fourth most frequent C→T substitution; C29095T occurrs almost seven times as often as a typical C→T substitution[4]. While there was no clear reason for the selection of this synonymous substitution, C29095 was found to be unpaired in a secondary structure model[8]. I hypothesized that deamination may be more frequent for unpaired cytosine residues. This was supported by previous analysis with a resolution of ~300 nucleotides[9]. To determine whether secondary structure was in fact correlated with mutation frequency at single-nucleotide resolution, the data set reported by Lan et al[2] was compared to the mutational fitness estimates reported by Bloom and Neher[4]. 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→T being much more frequent than all other types of substitutions).

The data compared in this study consisted of estimated mutational fitness for the SARS-CoV-2 genome as reported by Bloom and Neher[4] (Supplementary Data ntmut_fitness_all.csv and nt_fitness.csv) 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[2] (Supplementary Data 7 and 8). 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[6]. 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. I focused on the most common types of nucleotide substitutions: those comprising approximately 5% or more of total substitutions.

## Results

There was a significant increase in synonymous substitution frequencies at unpaired positions for C→T, G→T, C→A, and T→C, but not for A→G or G→A ($p < 0.05$; Tukey's range test with Bonferroni correction; A→T, 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→T and G→T (**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). Median estimated mutational fitness for synonymous C→T and G→T at unpaired positions are higher than at basepaired positions by 1.46 and 1.36, respectively. Expressed in terms of substitution frequency rather than mutational fitness, the frequency of synonymous C→T and G→T substitution is about four times higher at unpaired positions than at basepaired positions. Together, this demonstrates a meaningful impact of secondary structure on substitution frequencies.

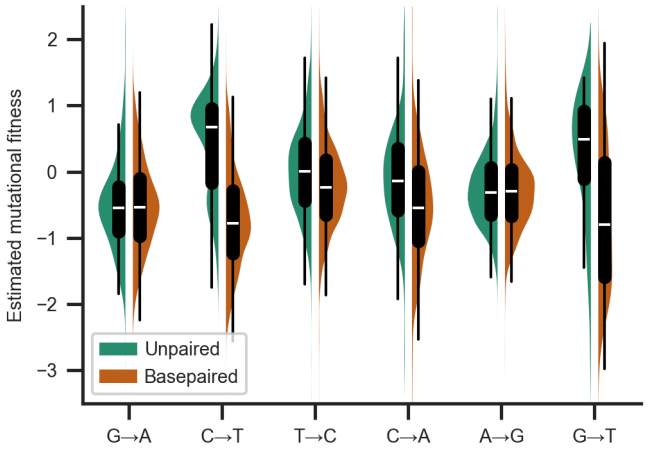


**Figure 1. Basepairing is predictive of synonymous substitution frequency.** Distribution of frequencies of synonymous substitutions for the most common substitutions (each approximately corresponding to 5% or more of observed substitutions), expressed as the estimated mutational fitness, which is a logarithmic comparison of the observed versus the expected number of occurrences of each type of substitution in the SARS-CoV-2 phylogenetic tree[4]. Distributions are grouped by substitution type and whether or not positions are basepaired in a full-genome secondary structure of SARS-CoV-2 in Huh7 cells[2].

Basepairing in the secondary structure model appears to be more predictive of estimated mutational fitness than average DMS reactivity, with correlation coefficients of 0.59 and 0.45, respectively, for C→T substitutions (point biserial and Spearman correlation coefficients). Correlation coefficients remain significant, but are reduced (0.18 and 0.13) when considering nonsynonymous mutations (**Figure 2**, left), consistent with larger and often negative effects of nonsynonymous mutations on viral fitness[4]. However, DMS reactivity is more correlated with estimated mutational fitness than basepairing when analysis is limited to positions with detectable DMS reactivity (excluding the sites plotted at the minimum measured value of 0.00012). 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→T substitutions by incrementing mutational fitness by +0.32 at basepaired position and by –1.14 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→T substitutions before and after applying this coarse adjustment (**Figure 2**, left and right).

For a preliminary estimate of whether nonsynonymous C→T substitutions at basepaired positions are prone to underestimation of mutational fitness, I tested the hypothesis that C→T 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[10] at any time during the pandemic). This was the case for 6 of the top 15 C→T substitutions at basepaired positions; their encoded mutations are shown in **Figure 2**. Top-ranked mutations characterize Omicron BA.1, one of the first recognized recombinant lineages XB, Gamma P.1, and the current fast-growing lineage JN.1.7. Half of these mutations have relatively high DMS reactivity for basepaired positions and half have very low DMS reactivity. By comparison, the synonymous substitution C29095T 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→T at unpaired positions define major lineages (BQ.1.1, JN.1.8.1, and BA.2.86.1).

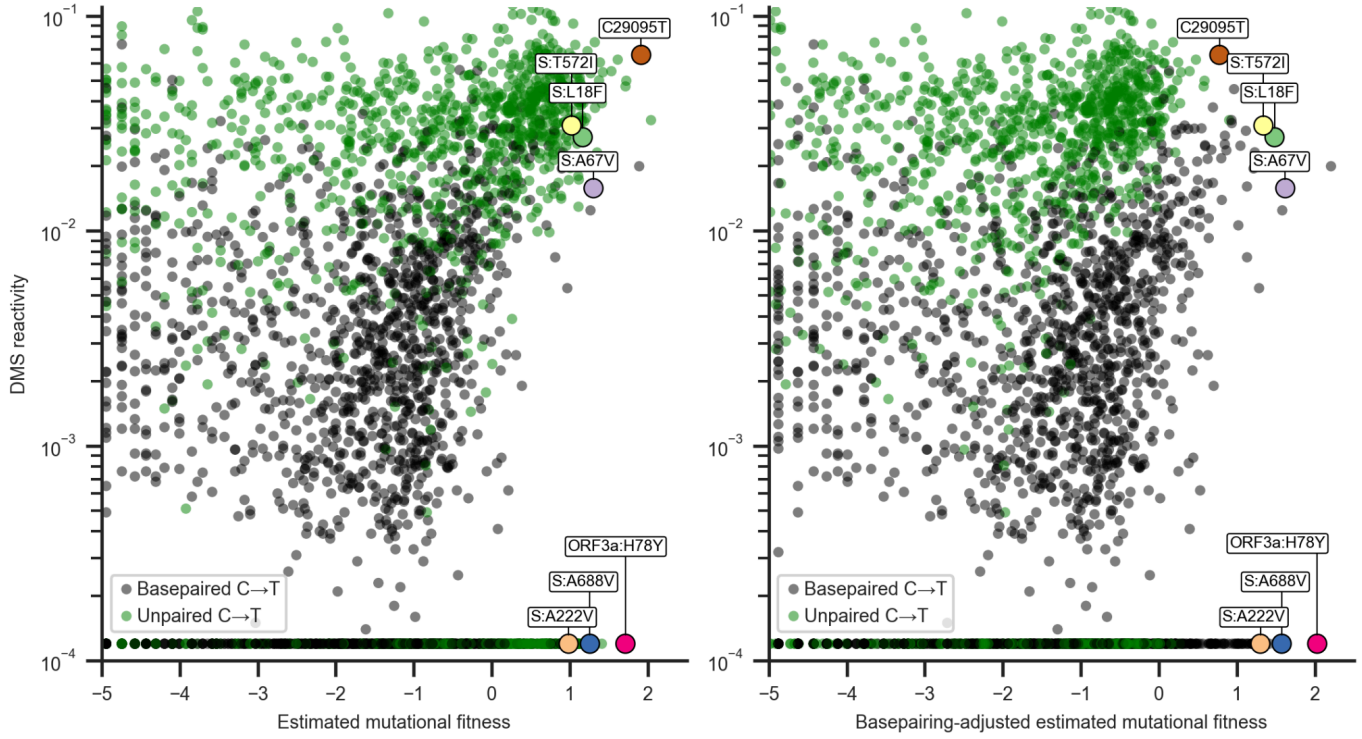


**Figure 2. Estimated mutational fitness correlates with secondary structure for nonsynonymous C→T substitutions.** Scatter plots compare mutational fitness to average DMS reactivity for positions with potential nonsynonymous C→T substitutions. The minimum observed DMS reactivity value is assigned to positions lacking data. Points are colored by basepairing in the full genome secondary structure model. Nonsynonymous C→T substitutions at basepaired positions are highlighted which rank highly for mutational fitness and characterize major SARS-CoV-2 lineages. Synonymous C29095T at an unpaired position is also highlighted. **Left**: Estimated mutational fitness based only on observed versus expected occurrences of C→T at each position. **Right**: Mutational fitness adjusted by constants derived from the medians of mutational fitness for synonymous substitutions at basepaired, unpaired, and all potential C→T positions.

Of particular note is C22227T at a basepaired position encoding the spike A222V mutation. This was one mutation that characterized the B.1.177 lineage, and it was unclear whether it conferred any fitness advantage[11]. 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[12]. Here, I focus on C→T substitutions for comparison to DMS reactivity data, but I also note that top-ranked G→T 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→T substitutions characterizing major variants now have some of the highest estimated mutational fitnesses for C→T substitutions (**Figure 2**, right).

## Discussion

This analysis shows that it is informative to combine apparent viral fitness, inferred from massive sequencing of SARS-CoV-2 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[4,13]. 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 mechanisms. I suggest that secondary structure, along with other data correlating with substitution frequency, can be used to refine estimates of mutational fitness. More sophisticated analysis can incorporate structural heterogeneity[2] as well as local sequence context[14]. Furthermore, additional measurements of secondary structure for genomes of new variants or modeling may reveal significant changes to secondary structure since 2020. 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[15]. However, it is critical to evaluate uncertainty in any model estimating fitness of a new variant. To this end, initial estimates can be refined by rapid *in vitro* characterization and continued genomic surveillance.

## Acknowledgments

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, Alex Crits-Cristoph, 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).

## Data availability

Data analyzed in this manuscript[2,4] and a Python notebook to reproduce analysis are available at https://github.com/smmlab/SARS2-fitness-secondary-structure

## References

1. Carabelli, A. M. *et al.* SARS-CoV-2 variant biology: immune escape, transmission and fitness. *Nat. Rev. Microbiol.* **21**, 162–177 (2023).
2. Lan, T. C. T. *et al.* Secondary structural ensembles of the SARS-CoV-2 RNA genome in infected cells. *Nat. Commun.* **13**, 1128 (2022).
3. Turakhia, Y. *et al.* Ultrafast Sample placement on Existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. *Nat. Genet.* **53**, 809–816 (2021).
4. Bloom, J. D. & Neher, R. A. Fitness effects of mutations to SARS-CoV-2 proteins. *Virus Evol.* **9**, vead055 (2023).
5. Chan, J. F.-W. *et al.* 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 (2020).

6. De Maio, N. *et al.* Mutation Rates and Selection on Synonymous Mutations in SARS-CoV-2. *Genome Biol. Evol.* **13**, evab087 (2021).
7. VanInsberghe, D., Neish, A. S., Lowen, A. C. & Koelle, K. Recombinant SARS-CoV-2 genomes circulated at low levels over the first year of the pandemic. *Virus Evol.* **7**, veab059 (2021).
8. Sun, L. *et al.* In vivo structural characterization of the SARS-CoV-2 RNA genome identifies host proteins vulnerable to repurposed drugs. *Cell* **184**, 1865-1883.e20 (2021).
9. Li, Y. *et al.* C-to-U RNA deamination is the driving force accelerating SARS-CoV-2 evolution. *Life Sci. Alliance* **6**, (2023).
10. Chen, C. *et al.* CoV-Spectrum: analysis of globally shared SARS-CoV-2 data to identify and characterize new variants. *Bioinformatics* **38**, 1735–1737 (2022).
11. Hodcroft, E. B. *et al.* Spread of a SARS-CoV-2 variant through Europe in the summer of 2020. *Nature* **595**, 707–712 (2021).
12. Ginex, T. *et al.* 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 Pathog.* **18**, e1010631 (2022).
13. Richard, D. *et al.* A phylogeny-based metric for estimating changes in transmissibility from recurrent mutations in SARS-CoV-2. 2021.05.06.442903 Preprint at https://doi.org/10.1101/2021.05.06.442903 (2021).
14. Lamb, K. D. *et al.* Mutational signature dynamics indicate SARS-CoV-2's evolutionary capacity is driven by host antiviral molecules. *PLOS Comput. Biol.* **20**, e1011795 (2024).
15. Dadonaite, B. *et al.* A pseudovirus system enables deep mutational scanning of the full SARS-CoV-2 spike. *Cell* **186**, 1263-1278.e20 (2023).