Molecular Biology and Evolution, 2025, **42**, 1–5
https://doi.org/10.1093/molbev/msaf059
Advance access publication 18 March 2025
Brief Communications

MBE OXFORD

A Phylogenetic Method Identifies Candidate Drivers of the Evolution of the SARS-CoV-2 Mutation Spectrum

Russ Corbett-Detig [1,2,*]

[1]Department of Biomolecular Engineering, University of California, Santa Cruz, USA
[2]Genomics Institute, University of California, Santa Cruz, USA

***Corresponding author:** E-mail: russcd@gmail.com.
Associate editor: Kelley Harris

Abstract

The molecular processes that generate new mutations evolve, but the causal mechanisms are largely unknown. In particular, the relative rates of mutation types (e.g. C > T), the mutation spectrum, sometimes vary among closely related species and populations. I present an algorithm for subdividing a phylogeny into distinct mutation spectra. By applying this approach to a SARS-CoV-2 phylogeny comprising approximately 8 million genome sequences, I identify ten shifts in the mutation spectrum. I find strong enrichment consistent with candidate causal amino-acid substitutions in the SARS-CoV-2 polymerase, and strikingly three appearances of the same homoplasious substitution are each associated with decreased C > T relative mutation rates. With rapidly growing genomic datasets, this approach and future extensions promise new insights into the mechanisms of the evolution of mutational processes. Keywords: Mutation Spectrum; Phylogenetic Analysis; SARS-CoV-2 Evolution

Keywords: mutation spectrum, phylogenetic analysis, SARS-CoV-2 evolution

Introduction

The molecular machinery that affects the rates and types of new mutations is subject to evolution. In particular, organisms ranging from humans (Harris 2015; Harris and Pritchard 2017) to bacteria (Wei et al. 2022) demonstrate changes in the relative rates of the types of new mutations (i.e. the mutation spectrum). In mammals, the mutation spectrum displays a phylogenetic signal such that more closely related species tend to exhibit similar relative mutation rates (Beichman et al. 2023). Associations between changes in the mutation spectrum and specific genetic elements, such as substitutions that affect the function of genome replication enzymes, have the potential to reveal the factors determining the molecular and evolutionary basis of genome maintenance and repair (e.g. Robinson et al. 2021; Sasani et al. 2022, 2024). Nonetheless, identifying such associations in natural populations is challenging, and there are few cases wherein naturally occurring causal alleles have been identified.

Recent studies leveraged the vast sequencing dataset for SARS-CoV-2 to investigate differences in the mutation spectrum with unprecedented resolution. Ruis et al. (2023) reported a decrease in the rate of G > T mutations in Omicron and descendant lineages. Similarly, (Bloom et al. 2023) corroborated this finding and discovered additional shifts in the mutation spectrum across SARS-CoV-2 lineages since its emergence in human populations. However, these and related studies generally rely on comparisons of mutation spectra using a priori subdivisions of the samples (e.g. clades or populations). This means that if the mutation spectrum changes are not associated with those specific subdivisions, they may be overlooked and the causal mechanisms difficult to discern.

Several phylogenetic methods identify changes in the mutation process across lineages (e.g. Huelsenbeck et al. 2000; Jayaswal et al. 2014). For example, (Blanquart and Lartillot 2006) developed a Bayesian method that can identify branches along the phylogeny where the substitution process changed. In principle, these approaches could facilitate identification of the molecular drivers of mutation spectrum evolution. However, existing maximum likelihood and Bayesian frameworks are limited due to computational demands, and new approaches that can effectively leverage vast genomic datasets are essential for comprehensively exploring the evolution of molecular evolutionary processes.

Results and Discussion

Algorithm Overview

I introduce spectrumSplits, a method for subdividing a phylogeny into nonoverlapping subtrees with distinct mutational spectra. This approach accepts a phylogenetic tree where mutations have been assigned to internal nodes via maximum parsimony using UShER (Turakhia et al. 2021). The algorithm performs a depth-first preorder traversal of the phylogeny and compares the mutation spectra for the subtrees on either side of a given node using a χ^2 test statistic where the categories are the counts of each mononucleotide mutation type (i.e. A > C, A > G, etc.) compared between the two subtrees. If the maximum recorded χ^2 is greater than a specified threshold, the tree is bisected at the corresponding node, resulting in a

Received: November 11, 2024. **Revised:** January 7, 2025. **Accepted:** January 9, 2025

Table 1 Summary of each spectrum split

Spectrum split	Ancestor split	Bootstrap proportion	Distance to nearest split (edges)	Mean Jaccard similarity	Closest lineage root	Δ C > T to ancestor	Δ G > T to ancestor	Associated NSP12 substitutions
node_358749	node_198068	0.573	0.487	0.9999	BA.2.75	−0.057	−0.01	Orf1B:G662S (NSP12:G671S)
node_966581	root	0.476	0.612	0.9358	B.1.617	−0.048	0.007	Orf1B:G662S (NSP12:G671S)
node_250268	node_198068	0.831	0.33	0.9999	XBB	−0.037	0.004	Orf1B:G662S (NSP12:G671S)
node_1145154	node_966581	0.705	0.295	0.9997	AY.43	0.031	0.021	Orf1B:L829I (NSP12:L838)
node_452195	node_421121	0.491	0.799	0.9847	BE.1.1	−0.028	−0.019	Orf1B:Y264H (NSP12:Y271H)
node_80726	root	0.703	0.731	0.8974	B.1.1.7	0.018	−0.014	N/A
node_198055	root	0.489	0.828	0.955	B.1.1.529	−0.013	−0.073	N/A
node_198068	node_198055	0.297	3.033	0.6799	BA.2	0.01	0.0014	N/A
node_1285194	node_966581	1	0	0.9974	AY.4	0.0084	0.017	N/A
node_421121	node_198068	0.613	0.896	0.9181	BA.5	0.003	0.012	N/A

This includes the ancestor spectrum split (or root, as applicable), the proportion of bootstraps that recovered the same node, the distance to the nearest spectrum split identified in bootstraps, the mean maximum Jaccard set similarity between the bootstrapped spectra and split in the full dataset, the closest PANGO lineage root (O'Toole et al. 2021) to each spectrum split, the changes in the C > T and G > T relative mutation rates compared with the ancestor of that spectrum split and the amino-acid substitutions in the polymerase in Orf1B coordinates and NSP12 coordinates in parentheses. The table is sorted by |Δ C > T|.

"spectrum split," and the procedure is repeated on the two resulting subtress until no additional nodes are discovered that exceed the threshold. The first split is therefore performed at the node with the largest χ^2 identified in the entire phylogeny. This is similar to the automated lineage designation method autolin (McBroome et al. 2024). I note that this procedure assumes that the evolution of the mutation spectrum is discontinuous and may not be ideal for gradual changes.

Pseudocode

```
initialize treeVector with tree
  while length(treeVector) > 0:
    for t in treeVector:
      for node in t:
        compute χ²
      if max χ² > threshold
        bisect t at max χ² node into t1, t2
        add t1, t2 to treeVector
      delete t from treeVector
```

Spectrum Partitions

I applied spectrumSplits to the comprehensive public SARS-CoV-2 phylogeny from the UShER project (dated 2024 09 21, McBroome et al. 2021; Turakhia et al. 2021). After obtaining the phylogeny, which receives substantial curation but has different goals than the present analysis (Hinrichs et al. 2024), I applied conservative methods to remove potentially spurious samples and mutations (see Materials and Methods). The tree used in analysis contained 7,593,765 samples with 4,668,537 mutations and is available for interactive visualization via taxonium (Sanderson 2022; Kramer et al. 2023). I identified ten spectrum splits associated with changes in the mutation spectrum. This procedure ran with 1000 bootstraps on 100 threads in 68 min.

Robustness of SpectrumSplits

I used nonparametric bootstraps where I resampled alignment columns with replacement to evaluate the robustness of spectrumSplits using three metrics. First, I computed the proportion of times that a spectrum split in the full dataset is recovered in each bootstrap. Second, because a spectrum split node might be similar but not identical to a bootstrap node, I calculated the mean distance, in number of edges, between each spectrum split identified in the full dataset and the nearest bootstrap spectrum split. Third, I calculated the maximum Jaccard set similarity of the node set descendant from a spectrum split between full and bootstrapped outputs. For each spectrum split, I found that bootstrap support is moderate to high (median 0.61, range 0.3 to 1.0, Table 1), that the distance to the nearest bootstrap split is low (median 0.72 edges, range 0.0 to 3.01), and furthermore that the Jaccard similarity is high (median 0.96, range 0.68 to 1.0). I conclude that the selection of spectrum splits is reasonably robust.

Primary Components of Mutation Spectrum Evolution

A principal component (PC) analysis identified the major axes of variation as the relative rates of C > T and G > T mutation. Just two PCs explain most of the variation among mutation spectra (78% and 21%, Fig. 1a to c). The first is largely related to the relative rate of G > T mutation, and stratification along this PC subdivides omicron and nonomicron lineages (Fig. 1b to d), implying that a large shift in the relative rate of G > T evolution evolved once in SARS-CoV-2. The second is driven by variation in the C > T mutation rate, which is more variable across the phylogeny (Fig. 1b, c, and e). My results are consistent with (Bloom et al. 2023; Ruis et al. 2023), and may also reveal causal mechanisms not previously described.

Candidate Causal Mutations of Spectrum Splits

I identified a set of candidate causal mutations by extracting amino-acid substitutions in the viral polymerase protein, NSP12, that occur near to spectrum splits. In light of the bootstrapping results, I included mutations at nodes wherein the descendant set has a substantial overlap with the spectrum split (see Materials and Methods). Among ten spectrum splits, five

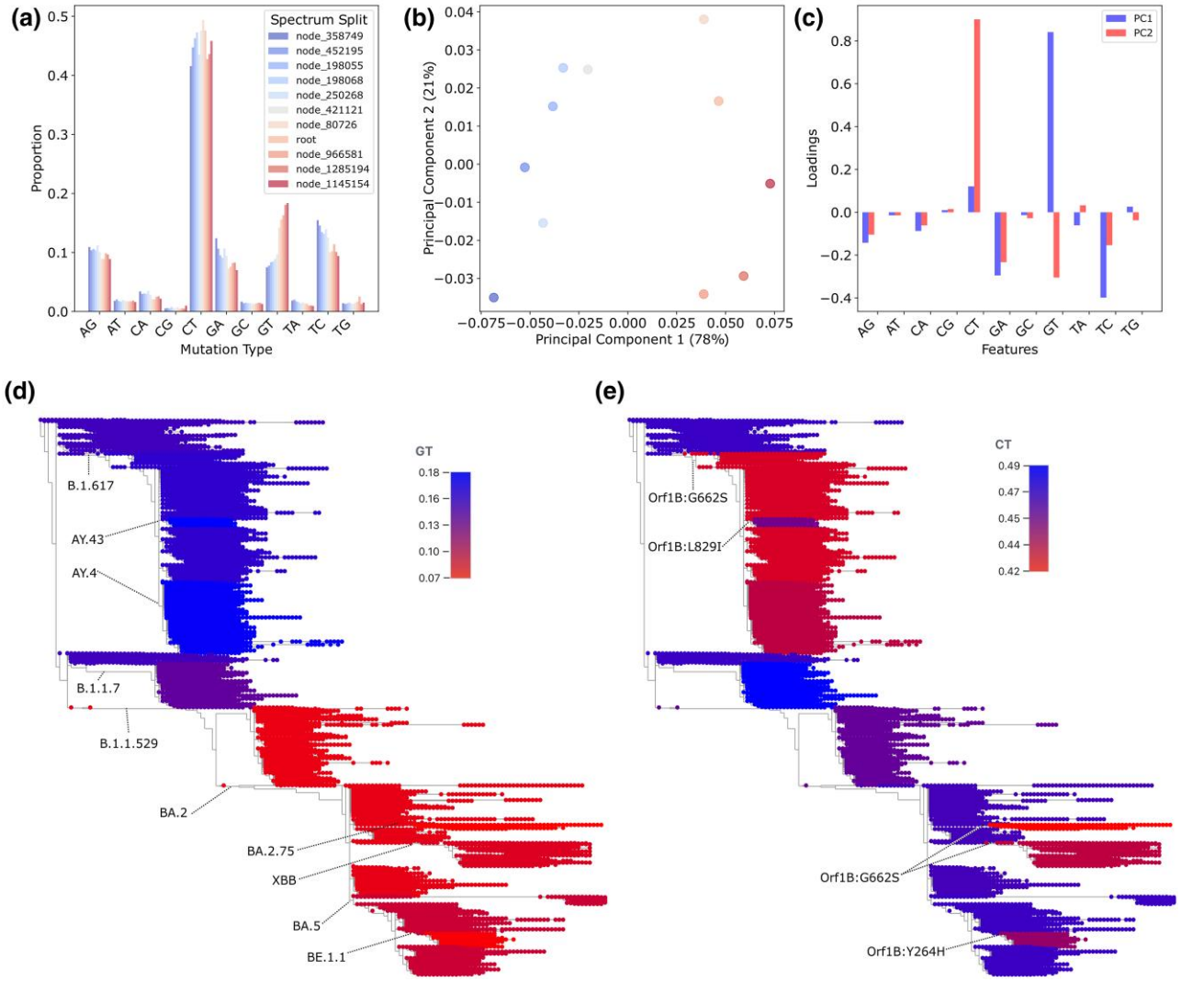

Fig. 1. Results from Spectrum splits. a) Relative proportion of each mutation type in each identified spectrum including the root. b) PC analysis of the 11 mutation spectra obtained from this analysis, colored as in a). c) PCA loadings for the first two PCs across the 12 relative mutation rates. d) Estimated relative rates of G > T substitution in each spectrum split. The nearest PANGO lineage root for each spectrum split is labeled. e) Estimated relative rates of C > T substitution in each spectrum split. Amino-acid substitutions in the viral polymerase are labeled at the nodes where they occurred. An interactive visualization is available here.

are associated with a nonsynonymous substitution in NSP12 (Table 1; $P < 0.0001$, permutation test). Furthermore, the five spectrum splits associated with nonsynonymous substitutions in the polymerase, are those with the largest change in the C > T relative mutation rate (0.028 to 0.057 vs. 0.003 to 0.018; $P = 0.0079$, Mann–Whitney U test). This suggests that substitutions within the polymerase contribute to mutation spectrum evolution in SARS-CoV-2.

The association between decreased C > T mutation rates and recurrent substitutions of Orf1B:G662S (NSP12:G671S) suggests it may be a causal mutation of mutation spectrum evolution. Orf1B:G662S enhances replication possibly by contributing to the stability of the replication complex (Kim et al. 2023) and may enhance viral fitness. I found this substitution in association with the three largest decreases in the relative rate of C > T mutation among all spectrum splits (−0.037, −0.048, and −0.057; $P = 0.0167$, Mann–Whitney U test). In fact, these are the only nodes where Orf1B:G662S evolved with more than 10,000 descendant samples, and it is unlikely that three randomly-selected nodes would be associated with the same amino-acid substitution ($P < 0.0001$, permutation test). Therefore, this mutation is a strong target for future functional investigations of mutation spectrum evolution. Furthermore, I speculate that C > T mutation rate evolution might provide a proxy for polymerase function (either fidelity or speed) when comparing viral lineages and therefore present a valuable metric for functional surveillance of SARS-CoV-2 to identify epidemiologically distinct lineages.

That there are no amino-acid-altering substitutions in NSP12 associated with the origins of Omicron (B.1.1.529), which displays one of the largest changes in mutation spectrum, suggests that distinct factors might drive the evolution of C > T and G > T relative mutation rates. In particular, mutations in other parts of the genome could affect the mutation spectrum. For example, ORF1B:I1566 V (NSP14:I42 V) in the viral proofreading domain is a strong plausible candidate (Bloom et al. 2023). Nonetheless, Omicron also shifted toward a replication niche in the upper lungs (Hui et al. 2022) which might affect exposure to mutagens and impact the mutation spectrum (Ruis et al. 2023). In the absence of replicate

large changes in the relative G>T mutation rates among SARS-CoV-2 lineages, distinguishing among these possibilities is difficult. Future sampling and application of spectrumSplits could help identify which additional factors are responsible for mutation spectrum evolution.

Conclusion

SpectrumSplits rapidly identifies candidate nodes that demonstrate changes in their mutation spectrum, which can then be associated with candidate mutations for downstream experiments and has the potential to empower surveillance of pathogen functional evolution. In particular, the discovery of a strong association between Orf1B:G662S and decreased C>T mutation rates is an appealing future direction for characterizing the functional basis of the evolution of the mutation spectrum. As genomic datasets are rapidly growing in size, spectrumSplits will also enable investigations of the factors that affect evolutionary processes across diverse organisms. Although this method is currently designed for a nonrecombining phylogeny, extensions using ancestral recombination graphs (Ethier and Griffiths 1990; Wong et al. 2024) that consider both the local nonrecombined gene tree surrounding a spectrum-affecting mutation and unlinked gene trees are possible. Similarly, it is possible, and it could be informative, to extend this approach to identify evolution of mutation processes using additional features such as trinucleotide context and multinucleotide mutations. Therefore, spectrumSplits will provide a powerful basis for investigating the causes of the evolution of the mutation processes.

Materials and Methods

Quality Control

Sequencing and assembly errors have been widely documented in SARS-CoV-2 (e.g. Turakhia et al. 2020; De Maio et al. 2020), and if unaddressed, they might lead to spurious inferences of changes in the mutation spectrum. I used two approaches to mitigate their impacts. First, I pruned subtrees where the ratio of mutations:tips exceeded 3, an effect commonly observed with amplicon dropout and resulting assembly errors. Second, I masked individual positions whose relative mutation rates increased dramatically at a particular node in the phylogeny—using an approach that is similar to spectrumSplits but considers mutations at one position vs. all other mutations. I visually inspected the masking results to determine if excluded mutations were likely to be spurious. The set of masked nodes and mutations is available in the GitHub repository associated with this project.

Spectrum Splits

I applied spectrum splits to identify nodes where the descendant mutation spectrum changed relative to its ancestor. I consider all mutation classes because SARS-CoV-2 is a plus-stranded RNA virus, and the strand is an important determinant of the mutation spectra (e.g. Simmonds 2020; Turakhia et al. 2020), rather than collapse mutation types by reverse complementarity. I ran spectrumSplits requiring a minimum χ^2 test statistic of 500. Determining the exact number of statistical tests implied is not straightforward—i.e. the algorithm traverses the tree or resulting subtrees repeatedly, many nodes are evaluated in each traversal, and the number of traversals is not known in advance. Nonetheless, the χ^2 value used is conservative, as the final procedure required fewer than 5 million statistical tests.

Bootstrapping

I used nonparametric bootstrapping to resample alignment columns to quantify the robustness of inferred spectrum splits. I compared bootstrapped datasets with the full dataset in terms of spectrum split identification, phylogenetic proximity to other spectrum splits, and the Jaccard set similarity in the descendant nodes contained within a given spectrum split and its maximally similar bootstrap split.

Permutation Tests

To determine if there is an enrichment for nonsynonymous mutations in the NSP12 associated with spectrum splits, I annotated each node with mutations found at the same node or at adjacent nodes that shared more than 99.99% of the descendant samples. For instance, node_966581 (B.1.617, Delta) includes a few additional samples that apparently contain subsets of the Delta-specific substitutions. I then performed permutation by randomly resampling nodes from the full phylogeny with a minimum of 10,000 descendants and identifying associated nonsynonymous polymorphisms in NSP12, and then determining if this permuted set of nodes is associated with the same or more NSP12 amino-acid altering substitutions than the true spectrum splits. I performed an equivalent test for association of spectrum splits with a single nonsynonymous substitution.

Acknowledgments

I thank Alex Ioannides, Robert Lanfear, Shelbi Russell, Yatish Turkahia, Angie Hinrichs, Landen Gozashti, Pratik Katte, Erik Enbody and members of the Corbett-Detig and Turkahia labs for helpful discussions. I gratefully acknowledge all individuals and groups who sequenced and shared SARS-CoV-2 genome sequences.

Funding

Funding was provided by National Institute of General Medical Sciences R35GM128932.

Conflict of Interest

RBC-D is a paid consultant for International Responder Systems.

Data Availability

SpectrumSplits and ancillary methods are available from https://github.com/russcd/spectrumSplits. The phylogeny used for this analysis is available at http://hgdownload.soe.ucsc.edu/goldenPath/wuhCor1/UShER_SARS-CoV-2/2024/09/28/public-2024-09-28.all.masked.pb.gz.

References

Beichman AC, Robinson J, Lin M, Moreno-Estrada A, Nigenda-Morales S, Harris K. Evolution of the mutation spectrum across a mammalian phylogeny. *Mol Biol Evol.* 2023:40(10):msad213. https://doi.org/10.1093/molbev/msad213.

Blanquart S, Lartillot N. A Bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution. *Mol Biol Evol*. 2006:23(11):2058–2071. https://doi.org/10.1093/molbev/msl091.

Bloom JD, Beichman AC, Neher RA, Harris K. Evolution of the SARS-CoV-2 mutational spectrum. *Mol Biol Evol*. 2023:40(4):msad085. https://doi.org/10.1093/molbev/msad085.

DeMaio N, Walker C, Borges R, Weilguny L, Slodkowicz G, Goldman N. Issues with SARS-CoV-2 sequencing data. In: Virological [Internet]. 2020 [accessed 2020 May 13]. http://virological.org/t/issues-with-sars-cov-2-sequencing-data/473.

Ethier SN, Griffiths RC. On the two-locus sampling distribution. *J Math Biol*. 1990:29(2):131–159. https://doi.org/10.1007/BF00168175.

Harris K. Evidence for recent, population-specific evolution of the human mutation rate. *Proc Natl Acad Sci U S A*. 2015:112(11):3439–3444. https://doi.org/10.1073/pnas.1418652112.

Harris K, Pritchard JK. Rapid evolution of the human mutation spectrum. *Elife*. 2017:6:e24284. https://doi.org/10.7554/eLife.24284.

Hinrichs A, Ye C, Turakhia Y, Corbett-Detig R. The ongoing evolution of UShER during the SARS-CoV-2 pandemic. *Nat Genet*. 2024:56(1):4–7. https://doi.org/10.1038/s41588-023-01622-5.

Huelsenbeck JP, Larget B, Swofford D. A compound poisson process for relaxing the molecular clock. *Genetics*. 2000:154(4):1879–1892. https://doi.org/10.1093/genetics/154.4.1879.

Hui KPY, Ho JCW, Cheung Mc, Ng Kc, Ching RHH, Lai Kl, Kam TT, Gu H, Sit KY, Hsin MKY, *et al*. SARS-CoV-2 Omicron variant replication in human bronchus and lung ex vivo. *Nature*. 2022:603:715–720. https://doi.org/10.1038/s41586-022-04479-6.

Jayaswal V, Wong TKF, Robinson J, Poladian L, Jermiin LS. Mixture models of nucleotide sequence evolution that account for heterogeneity in the substitution process across sites and across lineages. *Syst Biol*. 2014:63(5):726–742. https://doi.org/10.1093/sysbio/syu036.

Kim S-M, Kim E-H, Casel MAB, Kim Y-I, Sun R, Kwak M-J, Yoo J-S, Yu M, Yu K-M, Jang S-G, *et al*. SARS-CoV-2 variants with NSP12 P323L/G671S mutations display enhanced virus replication in ferret upper airways and higher transmissibility. *Cell Rep*. 2023:42(9):113077. https://doi.org/10.1016/j.celrep.2023.113077.

Kramer AM, Sanderson T, Corbett-Detig R. Treenome browser: co-visualization of enormous phylogenies and millions of genomes. *Bioinformatics*. 2023:39(1):btac772. https://doi.org/10.1093/bioinformatics/btac772.

McBroome J, de Bernardi Schneider A, Roemer C, Wolfinger MT, Hinrichs AS, O'Toole AN, Ruis C, Turakhia Y, Rambaut A, Corbett-Detig R. A framework for automated scalable designation of viral pathogen lineages from genomic data. *Nat Microbiol*. 2024:9(2):550–560. https://doi.org/10.1038/s41564-023-01587-5.

McBroome J, Thornlow B, Hinrichs AS, Kramer A, De Maio N, Goldman N, Haussler D, Corbett-Detig R, Turakhia Y. A daily-updated database and tools for comprehensive SARS-CoV-2 mutation-annotated trees. *Mol Biol Evol*. 2021:38(12):5819–5824. https://doi.org/10.1093/molbev/msab264.

O'Toole Á, Scher E, Underwood A, Jackson B, Hill V, McCrone JT, Colquhoun R, Ruis C, Abu-Dahab K, Taylor B, *et al*. Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. *Virus Evol*. 2021:7(2):veab064. https://doi.org/10.1093/ve/veab064.

Robinson PS, Coorens THH, Palles C, Mitchell E, Abascal F, Olafsson S, Lee BCH, Lawson ARJ, Lee-Six H, Moore L, *et al*. Increased somatic mutation burdens in normal human cells due to defective DNA polymerases. *Nat Genet*. 2021:53(10):1434–1442. https://doi.org/10.1038/s41588-021-00930-y.

Ruis C, Peacock TP, Polo LM, Masone D, Alvarez MS, Hinrichs AS, Turakhia Y, Cheng Y, McBroome J, Corbett-Detig R, *et al*. A lung-specific mutational signature enables inference of viral and bacterial respiratory niche. *Microb Genom*. 2023:9(5):001018. https://doi.org/10.1099/mgen.0.001018.

Sanderson T. Taxonium, a web-based tool for exploring large phylogenetic trees. *Elife*. 2022:11:e82392. https://doi.org/10.7554/eLife.82392.

Sasani TA, Ashbrook DG, Beichman AC, Lu L, Palmer AA, Williams RW, Pritchard JK, Harris K. A natural mutator allele shapes mutation spectrum variation in mice. *Nature*. 2022:605(7910):497–502. https://doi.org/10.1038/s41586-022-04701-5.

Sasani TA, Quinlan AR, Harris K. Epistasis between mutator alleles contributes to germline mutation spectrum variability in laboratory mice. *Elife*. 2024:12:RP89096. https://doi.org/10.7554/eLife.89096.

Simmonds P. 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*. 2020:5(3):e00408–e00420. https://doi.org/10.1128/mSphere.00408-20.

Turakhia Y, De Maio N, Thornlow B, Gozashti L, Lanfear R, Walker CR, Hinrichs AS, Fernandes JD, Borges R, Slodkowicz G, *et al*. Stability of SARS-CoV-2 phylogenies. *PLoS Genet*. 2020:16(11):e1009175. https://doi.org/10.1371/journal.pgen.1009175.

Turakhia Y, Thornlow B, Hinrichs AS, De Maio N, Gozashti L, Lanfear R, Haussler D, Corbett-Detig R. Ultrafast sample placement on existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. *Nat Genet*. 2021:53(6):809–816. https://doi.org/10.1038/s41588-021-00862-7.

Wei W, Ho W-C, Behringer MG, Miller SF, Bcharah G, Lynch M. Rapid evolution of mutation rate and spectrum in response to environmental and population-genetic challenges. *Nat Commun*. 2022:13(1):4752. https://doi.org/10.1038/s41467-022-32353-6.

Wong Y, Ignatieva A, Koskela J, Gorjanc G, Wohns AW, Kelleher J. A general and efficient representation of ancestral recombination graphs. *Genetics*. 2024:228(1):iyae100. https://doi.org/10.1093/genetics/iyae100.