

## **SARS-CoV-2 saltational events are recurrent and trace to persistent human infections**

Cécile Tran-Kiem<sup>1</sup>, Kathryn Kistler<sup>1,2</sup>, Ryan Hisner<sup>3</sup>, Trevor Bedford<sup>1,2</sup>

1. Vaccine and Infectious Diseases Division, Fred Hutchinson Cancer Center, Seattle, WA, USA
2. Howard Hughes Medical Institute, Seattle, WA, USA
3. Division of Computational Biology, Department of Integrative Biomedical Sciences, Institute of Infectious Diseases and Molecular Medicine, University of Cape Town, Cape Town, South Africa

### **Abstract**

SARS-CoV-2 evolution is characterized by gradual mutation accumulation but has been punctuated by rare yet impactful highly mutated variants. Whether such saltational jumps are a broad feature of SARS-CoV-2 evolution or rare anomalies remains unclear. We systematically investigate SARS-CoV-2 saltational evolution by developing a scalable framework to detect saltational events from 4.4 million high-quality viral genomes. Saltational events occurred at low but detectable rates during the pandemic and post-pandemic periods and across geographies. Their mutational signature closely matches that seen in persistent human infections but is inconsistent with the signatures of mink or deer infections. This points to persistent infection, rather than reverse zoonosis, as their primary source. While most saltational events lack evidence of onward transmission, those that do tend to carry mutations found in successful clades. Our work demonstrates that the emergence of highly mutated SARS-CoV-2 variants reflects a recurrent evolutionary process, with implications for preparedness.

### **Introduction**

The accumulation of mutations in SARS-CoV-2 genomes typically occurs incrementally as the virus spreads throughout the population, with one substitution arising roughly every two weeks. However, SARS-CoV-2 evolution has been punctuated by the emergence of variants carrying unusually high numbers of mutations<sup>1-4</sup>, some of which have rapidly swept worldwide and generated “pandemics within the pandemic”<sup>5</sup>. Because of their major epidemiological consequences, these successful saltational events have attracted considerable attention.

Yet our understanding of SARS-CoV-2 saltational evolution remains largely based on a handful of high-profile variants of concern (VOCs), leaving it unclear whether such events were rare anomalies or a broader underappreciated feature of SARS-CoV-2 evolution. The unprecedented scale of genomic surveillance implemented in response to the pandemic offers an opportunity to address this question by quantifying how frequently saltational events occurred, how they were distributed across space and time and whether there are features associated with their successful spread. Such data may also provide insights into the mechanisms that gave rise to these events. Accelerated evolution during a subset of persistent SARS-CoV-2 infections has emerged as a

very plausible hypothesis<sup>6</sup>, motivated by the overlap between mutations observed in chronically infected individuals and those found in VOCs<sup>7</sup>, with alternative explanations such as accumulation of mutation within animal reservoirs having little evidence to support them<sup>7</sup>.

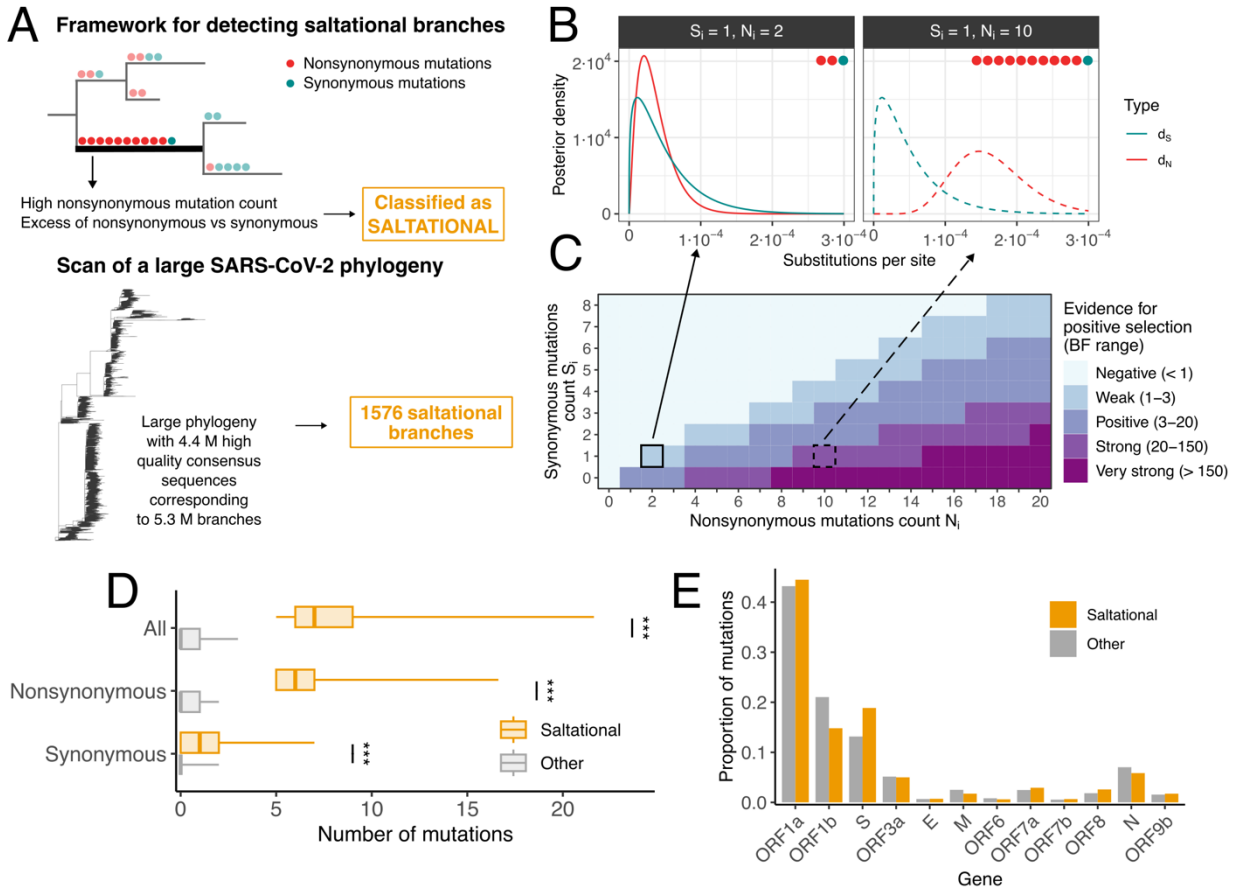
Here, we perform a systematic analysis of SARS-CoV-2 saltational evolution from a large publicly available high-quality SARS-CoV-2 sequence dataset and phylogeny comprising 4.4 million samples<sup>8</sup>. We develop a scalable framework to detect saltational events, defined as branches in the phylogeny with strong evidence for positive selection and a high number of nonsynonymous mutations, and show they are not isolated anomalies but occur repeatedly during SARS-CoV-2 circulation. We find that such events have a distinct mutational signature that strongly resembles that observed in persistent SARS-CoV-2 infections. Finally, we investigate the extent to which saltational events seed onward transmission and identify mutational features associated with their successful spread.

## Results

### *A scalable framework to detect saltational evolution in a mutation-annotated tree*

To systematically study SARS-CoV-2 saltational evolution, we analyze a large SARS-CoV-2 phylogeny built from 4.4 million high-quality consensus sequences<sup>8</sup>, yielding a mutation-annotated tree with 5.3 million branches. We classify saltational branches as branches displaying both (i) an unusually high number of nonsynonymous mutations, (defined as branches with at least 5 nonsynonymous mutations, corresponding to greater than the 99.5th percentile of the distribution across all branches of the phylogeny, see Figure S1), and (ii) evidence of positive selection (Figure 1A). We infer positive selection at the gene or whole-genome level from an excess of nonsynonymous relative to synonymous mutations on a branch, accounting for nonsynonymous and synonymous mutation opportunities ( $d_N/d_S$  ratio above 1). To do so, we develop a Bayesian framework to estimate branch- and gene-specific  $d_N/d_S$  while accounting for sparse mutation counts. This approach allows us to account for uncertainty in branch-level  $d_N/d_S$  estimates and to handle the fact that most branches contain few, if any, synonymous mutations. We then use Bayesian hypothesis testing to identify branches with strong evidence that  $d_N/d_S > 1$  at the gene or whole-genome level (see Methods).

Figure 1B-C illustrates the model's behavior. On a branch carrying 1 synonymous and 2 nonsynonymous mutations in the spike gene (S), we infer a posterior median of  $3.7 \cdot 10^{-5}$  synonymous mutations per site and a posterior median of  $3.1 \cdot 10^{-5}$  nonsynonymous mutations per site, resulting in little evidence for positive selection with  $d_N/d_S$  estimated at 0.84 [95% credible interval: 0.086–14] and a Bayes Factor of 1.1. By contrast, on a branch carrying 10 nonsynonymous mutations in S, the posterior median of  $d_N$  increases to  $1.6 \cdot 10^{-4}$ , resulting in strong support for positive selection with  $d_N/d_S$  estimated at 4.2 [95% credible interval: 0.83–63] and a Bayes Factor of 31. Applying this framework to a tree built from 4.4 million samples collected between January 2020 and June 2024, we identify 1576 saltational branches. Compared with non-saltational branches, these branches carry substantially more mutations, especially nonsynonymous ones (Figure 1D), and are enriched in mutations in Spike (Figure 1E).



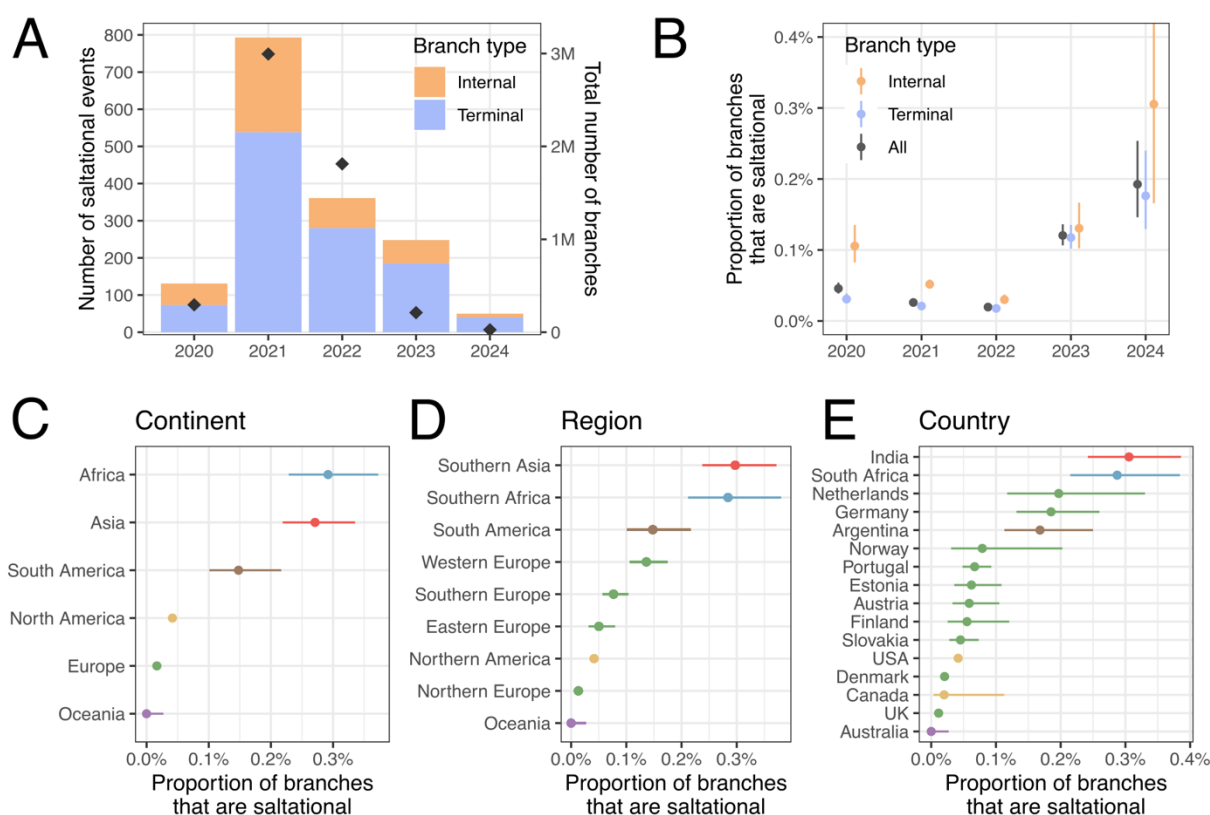
**Figure 1: Framework for detecting saltational branches in a large SARS-CoV-2 phylogeny. A.** Schematic overview of the framework used to classify branches as saltational, which we apply on a large SARS-CoV-2 phylogeny. **B.** Posterior distribution of nonsynonymous ( $d_n$ ) and synonymous ( $d_s$ ) divergence measured as substitutions per site on the spike (S) gene for a branch harboring 1 synonymous mutation and 2 or 10 nonsynonymous mutations. **C.** Strength of evidence for positive selection on the spike gene as a function of the number of nonsynonymous and synonymous mutations in spike occurring on a branch. The range in parenthesis in the legend indicates the range of the Bayes Factor for  $d_n/d_s > 1$  corresponding to each evidence category<sup>9</sup>. **D.** Number of mutations (total, nonsynonymous and synonymous) occurring on saltational compared with non-saltational branches. **E.** Proportion of mutations occurring on each gene on saltational and non-saltational branches. In D, boxplots indicate the 2.5th, 25th, 50th, 75th and 97.5th percentiles.

### *Saltational evolution occurred at a low but detectable rate since SARS-CoV-2 emergence*

We next investigate how the saltational branches we identify are distributed through space and time. We detect saltational branches in every year of the study period, on both terminal and internal branches of the phylogeny (Figure 2A). Their absolute count varies across years, with more events detected during periods where more sequences are collected. Overall, 0.03% of all branches are classified as saltational. Year-specific proportions are higher in 2023 and 2024 than in previous years (Figure 2B): whereas we identify 0.02% of branches as saltational in 2022 [95% CI: 0.02–0.02%], this proportion rises by ten-fold to 0.2% [95% CI: 0.1–0.3%] in 2024. This increase is also observed when subsampling the tree to the same number of samples per year

(Figure S2), indicating that this trend cannot be only attributed to the lower number of sequences available in more recent years.

We identify saltational branches across all continents represented in the dataset we studied, except Oceania (Figure 2C), highlighting their broad geographic distribution. Looking at absolute counts, we detect the largest number of saltational events in the USA (809) and the UK (313), the two countries contributing the largest number of sequences in the studied dataset. The fact that we detect more saltational events in locations with higher sequencing (Figure S3) suggests that saltation is a general feature of SARS-CoV-2 evolution that can be observed when genomic surveillance is sufficiently intense. Focusing on regions with more than 5000 branches in the global phylogeny (see Methods for ancestral reconstruction), we find some variation in the proportion of branches that are saltational (Figure 2D-E), ranging from 0.0% in Australia [95% CI: 0.00–0.03%] to 0.3% in both India and South Africa [95% CI: 0.2–0.4%]. These observed differences could stem from heterogeneity in emergence opportunities and geographic variation in the detectability of saltational events. Taken together, these findings show that saltational evolution is a recurring feature of the COVID-19 pandemic with a broad geographic footprint.



**Figure 2: Spatiotemporal distribution of saltational SARS-CoV-2 branches.** **A.** Number of saltational branches identified by year and branch type (bars with scale on the left). Diamonds depict the total number of branches for each year (right axis). **B.** Proportion of branches classified as saltational by year and branch type. Proportion of branches classified as saltational by **C.** continent, **D.** region and **E.** country. In B-E, segments indicate 95% Wilson confidence intervals.

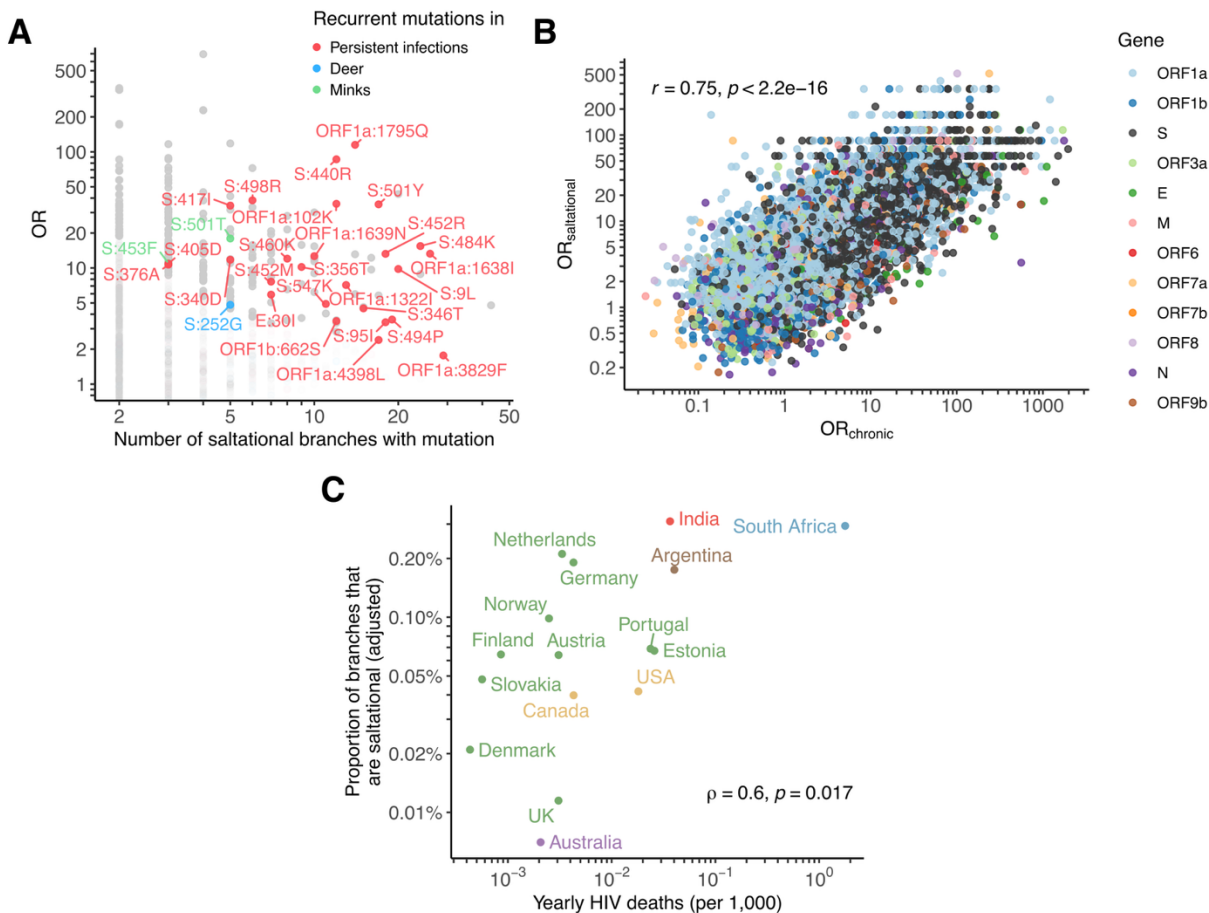


### *SARS-CoV-2 saltation events have a distinct mutational profile*

We identify many residues along the genome that are repeatedly mutated in saltational branches (Table S1). To explore whether saltational branches display a distinct mutational signature, we compare whether a residue is more likely to have a nonsynonymous mutation in saltational branches compared with non-saltational branches. Across the genome, we find many residues that are more likely to be mutated on saltational branches than on non-saltational ones (Figure 3A). These sites concentrate in Spike (Figure 3B), particularly in the receptor-binding domain (RBD) (Figure 3C). These include many functionally relevant residues, known to affect immune escape (S:484 is mutated in 52 saltational branches, OR: 6.2 [95% CI: 4.6–8.1]), receptor binding (S:501 is mutated in 24 saltational branches, OR: 16.2 [95% CI: 10.2–24.7]) or viral entry (S:681 is mutated in 34 saltational branches, OR: 7.6 [95% CI: 5.2–10.7]). We also detect a high concentration of positions that are more likely to be mutated in saltational branches in the Nsp3 region of the ORF1a polyprotein (Figure 3D). This distinct mutational profile includes a higher frequency of mutations at known antigenic sites, including positions associated with antibody escape (Figure 3E) and surface accessible sites (Figure 3F), suggesting that shared selective pressures, such as immune escape, may act across saltational events and leave a detectable signature.

### *The mutational signature of SARS-CoV-2 saltational events overlaps with that observed during persistent infections*

The mutation enrichment profile of saltational branches may thus provide insights into the mechanisms that give rise to saltational events. To investigate this, we identify amino-acid mutations that occur significantly more often on saltational branches than on non-saltational ones (Table S2) and compare them with mutations reported to recur during persistent human infections, as well as after spillover into mink and deer populations since persistent infections and adaptation within other species were two hypotheses potentially explaining VOC-defining saltational branches (Figure 4A). This allows us to assess whether proposed mechanisms for the emergence of saltational events<sup>7,16,17</sup> have recurrent mutations that are characteristic of SARS-CoV-2 saltational evolution. We find a strong overlap between mutations enriched in saltational branches and mutations that occur repeatedly during persistent infections: 26 enriched mutations (found at high frequency in saltational branches) overlap with mutations recurring in persistent infections. This includes mutations outside the Spike protein such as ORF1a:1638I (in 26 saltational branches), ORF1a:1795Q (in 14 saltational branches) or E:30I (in 7 saltational branches), that have been reported in persistent infections but that are observed at very low frequency in global sequences<sup>18–21</sup>. By contrast, only one enriched mutation overlaps with recurrent deer-associated mutations, and 2 overlap with mink-associated mutations. This overlap suggests a role for evolution within chronically infected individuals in the emergence of saltational events, though adaptation to other host species may also contribute to a small proportion of saltational events.



**Figure 4: The mutational signature of SARS-CoV-2 saltational evolution supports the role of persistent human infections in their emergence.** **A.** Odds ratio for a mutation occurring on saltational branches compared with non-saltational ones, as a function of the number of saltational branches carrying that mutation. Mutations that don't reach statistical significance (adjusted p-value greater than 0.05) are shown with greater transparency. Points are colored according to whether the mutation has been reported to recur in persistent infections, deer or minks. **B.** Relationship between the odds ratio of a mutation occurring in saltational branches compared with non-saltational ones as a function of the odds ratio of the same mutation being observed in sequences from chronically infected individuals compared with a high-quality background sequence dataset<sup>22</sup>. **C.** Proportion of branches that are saltational as a function of HIV deaths (per 1,000 inhabitants) estimated for 2022<sup>23,24</sup>. To better visualize the trend in C, variables are displayed on a logarithmic scale and proportions are adjusted using pseudo-counts (Laplace smoothing) to enable displaying the proportion of branches that are saltational in Australia (equal to 0) (see Methods). The Spearman correlation coefficient is computed based on the unadjusted proportion.

Next, we evaluate whether the mutational signature of saltational branches overlaps with that of chronic infections by comparing the odds ratio of mutations occurring on saltational vs non-saltational branches to the odds ratio of the same SARS-CoV-2 mutation occurring in sequences collected from chronic infections relative to other circulating SARS-CoV-2 sequences<sup>22</sup> (Figure 4B). We find a strong positive correlation of log odds ratio (Pearson  $r = 0.75, p < 2.2 \cdot 10^{-16}$ ) between these two metrics, further supporting persistent infections as a likely source of many saltational events.

Persistent SARS-CoV-2 infections have been repeatedly reported in immunocompromised individuals<sup>7</sup>, which suggests that the prevalence of conditions associated with immunosuppression may impact the likelihood of saltational events. Among these conditions, advanced human immunodeficiency virus (HIV) infection is of particular interest as it has been associated with prolonged time to SARS-CoV-2 clearance<sup>6,25</sup>, and heterogeneity in immune suppression levels that may be associated with some residual immunity against SARS-CoV-2 that could provide enough selection pressure to generate immune escape mutations<sup>7,26</sup>. Moreover, several VOCs (including Beta and Omicron) were first detected in regions experiencing a high HIV burden, which has led to the hypothesis that HIV may be linked to the emergence of highly mutated variants<sup>7</sup>. Here, we examine the relationship between spatial variation in the frequency of saltational events and heterogeneity in HIV burden across countries (Figure 4C). We find a moderate positive correlation between the country-level proportion of branches that are saltational and yearly AIDS mortality estimates (Spearman  $\rho = 0.60$ ,  $p = 0.017$ ). This association is reduced but remains consistent when excluding South Africa from the analysis (Spearman  $\rho = 0.52$ ,  $p = 0.051$ ).

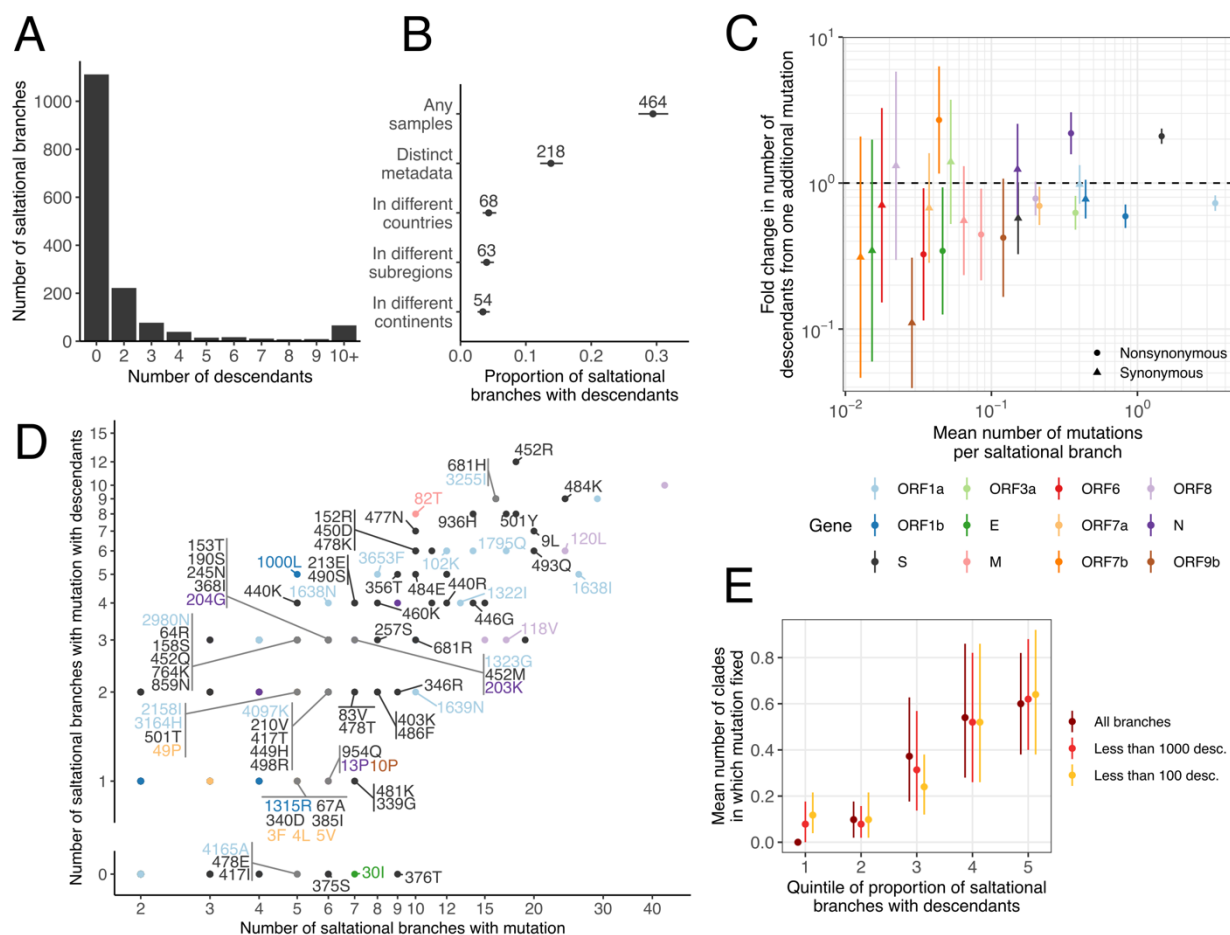
Overall, these results support the role of persistent human infections in the emergence of saltational events.

#### *Evidence of onward spread of a subset of saltational branches*

So far, we have characterized the frequency and likely origin of saltational events. We next investigate their epidemiological consequences and the factors associated with their potential spread. Of the saltational events we identify, 71% [95% CI: 68-73%] occur on terminal branches (Figure 5A) and thus have no detected descendants in the phylogeny. Whether a saltational event occurs on a terminal or internal branch however doesn't directly translate to whether that event seeded a transmission cluster as (i) descendants may not be sampled and (ii) some descendant cluster may reflect repeated sampling of the same individual<sup>27</sup>. To gain insights into saltational branches that seeded transmission clusters, we look for evidence of descendants with distinct characteristics (age, sex and geography). We find evidence for spread to distinct individuals in 218 saltational branches (14%, 95% CI: 12-16%), with a subset spreading across borders: 68 spread to multiple countries (4.3%, 95% CI: 3.4-5.5%) and 54 to multiple continents (3.4%, 95% CI: 2.6-4.5%) (Figure 5B). Among saltational branches with descendants, most remain detectable only briefly (median duration of 9 days), whereas a small fraction has descendants detectable over longer periods (90<sup>th</sup> percentile: 103 days) (Figure S4).

We explore whether the mutational composition of saltational branches is associated with the number of descendants each branch harbors. We model descendant counts as a function of synonymous and nonsynonymous mutation counts in each gene using a negative binomial regression (see Methods). Figure 5C depicts, for each gene and mutation type, the fold-change in expected descendant count associated with one additional mutation, as a function of the frequency of this mutation type in saltational branches. We find that one additional nonsynonymous mutation in spike is associated with a 2.1-fold [95% CI: 1.9-2.4] change in expected descendant counts, while one additional mutation in ORF1a and ORF1b are associated with a 0.73-fold [95% CI: 0.64-0.82] and a 0.59-fold [95% CI: 0.49-0.71] change in expected descendant counts. We identify other significant associations, including a positive effect of

nonsynonymous mutations in N and ORF7b, but these mutations are relatively rare on saltational branches and are hence less likely to explain a large fraction of the overall variation in descendant counts.



**Figure 5: Patterns of onward spread from saltational branches.** **A.** Distribution of descendant counts across saltational branches. **B.** Proportion of saltational branches with descendants across different spread criteria. Segments indicate 95% confidence intervals. **C.** Estimated fold-change in the number of descendants from one additional mutation as a function of the mean number of mutations per saltational branch as by mutation type (synonymous and nonsynonymous) and gene. Vertical segments indicate 95% confidence intervals. **D.** Number of saltational branches that have descendants and a mutation of interest as a function of the number of saltational branches that have that mutation. For clarity, we only write the labels of mutations with an OR of mutation occurring in saltational branches greater than 5 and that occur at least 5 times in saltational branches. **E.** Mean number of clades in which a mutation fixed at the global level by quintile of proportion of saltational branches with that mutation that have descendants. Vertical segments indicate 95% bootstrap confidence intervals. In E, we report the results based on all saltational branches or only saltational branches with less than 1000 or 100 descendants.

Finally, we examine whether specific individual mutations are associated with saltational events leaving descendants (Figure 5D). We identify several mutations that occur repeatedly on saltational branches with descendants. By contrast, there are other mutations frequently observed on saltational branches that never leave detectable descendants. This suggests that the

mutational composition of saltational branches may influence their ability to spread. Interestingly, S:484K, S:452R, S:681H and ORF1a:3255I, which recur frequently on saltational branches producing descendants, all are lineage-defining mutations in at least two emerging VOCs or Omicron sublineages. In comparison, we observe E:30I and the reversion S:376T on 7 and 9 saltational branches respectively, without any detectable descendants, and neither mutation was observed in successful SARS-CoV-2 clades. We find that mutations more frequently observed on saltational branches with descendants also fix in a greater number of clades globally (Figure 5E). This relationship persists when excluding branches with more than 100 descendants, indicating that this result is not driven solely by the identification of clade-defining branches as saltational events.

Overall, we find evidence of onward transmission from a subset of saltational events and identify recurrent mutations within them that may be associated with their ability to spread.

## Discussion

Saltational evolution with the emergence of highly fit SARS-CoV-2 variants has been observed multiple times during SARS-CoV-2 evolution in the human population. Our systematic investigation of SARS-CoV-2 saltational evolution reveals that such events are a recurring feature of SARS-CoV-2 evolution, occurring at a low but detectable rate since the beginning of the pandemic to at least mid-2024 where our study data ends. Their distinct mutational signature is consistent with a role of chronic human infection in their emergence. While most of these saltational events don't display evidence for onward transmission, a subset spread widely.

We find that the mutational signature of SARS-CoV-2 saltational events is informative about the potential mechanism of their emergence. Persistent replication of SARS-CoV-2 in chronically infected individuals with weakened immune system has been proposed to provide an evolutionary setting well-suited for the emergence of highly divergent SARS-CoV-2 variants<sup>6,7</sup>. The strong agreement between the mutational signature of saltational events and that of chronic human SARS-CoV-2 infections supports the role of persistent infections in SARS-CoV-2 saltational evolution (Figure 4B). We also identify a concentration of mutations at antigenically relevant residues in the Spike protein, particularly within the receptor-binding domain (Figure 3, Table S1-S2). This pattern is consistent with evolution during chronic infections of immunocompromised individuals, where selection pressure from partial immunity<sup>28</sup> and treatment administration (including monoclonal antibodies<sup>10-12,14</sup>) may favor the acquisition of immune escape mutations. By contrast, the mutational signature of saltational events is inconsistent with that expected from reverse zoonosis in animal reservoirs such as minks or deer (Figure 4A).

Temporal and geographic variation in the frequency of saltational events (Figure 2) may thus reflect differences in emergence opportunity, which could stem from differences in infection burden, in the prevalence of persistent infections, changes in treatment regimens, or potential lineage-specific differences in mutational tolerance<sup>29</sup>. Variations in the detectability of saltational events may also explain some of these variations, for example with differences in sequencing intensity, in the patient populations captured by surveillance strategies or sequence quality.

While most saltational branches don't display evidence for onward transmission, we identify 218 saltational events with evidence of transmission between individuals, a subset of which also dispersed widely, with for example 68 spreading across country borders (Figure 5). The distinct mutation patterns observed between branches that spread and those that don't suggest that some mutations impact the ability for saltational branches to transmit between hosts. This is further supported by the observation that mutations more frequently found in spreading branches also tend to be successful at the global level. Interestingly, the highly divergent BA.3.2 lineage that was first detected after the end of our study period was characterized by a particularly high number of amino-acid changes in Spike, many at residues that we found to be more frequently mutated in saltational branches with descendants. These include S:9L and S:440R, two changes that we observed repeatedly on saltational branches with descendants but that were not seen in earlier variants of concern. Overall, our findings are consistent with saltational events arising from persistent infections that select for viruses better adapted for persisting and replicating within hosts, but most of these variants are not well suited for transmission between hosts. However, a subset can transmit between hosts<sup>30</sup>, including some with potentially important epidemiological implications<sup>3,4,31</sup>. This may reflect pressure to escape immunity, either from the administration of treatment such as monoclonal antibodies, or from residual immunity, that may confer fitness advantages both at the within and between host level<sup>7</sup>.

Interestingly, although other pathogens can also cause persistent infections in immunocompromised individuals, a recurrent contribution of saltational events to their evolution hasn't been demonstrated<sup>32,33</sup>. One possible explanation is that mutations selected within hosts may not confer an advantage or may even be incompatible with transmission between hosts<sup>7</sup>. However, for some pathogens including influenza, recurrent mutations arising during chronic infections in immunocompromised individuals overlap with mutations successful at the global level<sup>33</sup>, suggesting that similar processes may also occur in those systems. Yet influenza is densely surveilled, and its phylogenies accumulate mutations in a stepwise manner without the long, heavily mutated branches that define SARS-CoV-2 saltational evolution, suggesting this contrast reflects a genuine difference between the two viruses. That said, it remains possible that persistent infections contribute to global influenza evolution, even if the contribution appears on the surface weaker than for SARS-CoV-2.

Given the high potential risk of SARS-CoV-2 evolution during persistent infections, there are both individual and collective benefits associated with detecting and clearing such infections to mitigate the risk of emergence and transmission of new variants that may have major epidemiological consequences<sup>6</sup>. Given persistent infections have been disproportionately reported in immunocompromised patients<sup>21</sup>, efforts to reduce the burden of immunosuppressive conditions, when possible, may indirectly prevent the emergence of highly mutated SARS-CoV-2 variants by decreasing emergence opportunity. Advanced HIV has been associated with chronic SARS-CoV-2 infections during which highly divergent viruses can evolve<sup>25,34,35</sup>, and we find a moderate positive correlation between country-level HIV burden and the proportion of branches that are saltational (Figure 4C). In this context, recent disruptions to global HIV care<sup>36,37</sup>, including reduced access to antiretroviral therapy, could increase the number of individuals at high risk of persistent SARS-CoV-2 infections, thereby potentially amplifying opportunities for high-impact variant emergence. While we don't aim to assess the impact of such disruptions on SARS-CoV-2

evolution, these considerations underline the interconnectedness of public health systems and crises, where the management of one epidemic may have indirect, here evolutionary, consequences for another, well beyond the initially affected populations.

Our ability to identify saltational events depends on the dataset analyzed and the criteria used to define them. Tightening or loosening these criteria would lead to detecting more or less events throughout the pandemic (Figure S5). Here, we thus don't estimate the total number of saltational events that occurred throughout the study period. Instead, we report the subset of events that we can detect under a conservative criterion from a large high-quality sequence dataset. Our estimate of 0.03% of branches that are saltational is thus conservative and would increase when considering a less stringent definition for saltational evolution (Figure S5). Finally, this study is enabled by the considerable scale of SARS-CoV-2 genomic surveillance globally<sup>38</sup>, along with the development of software tools able to handle large sequence datasets<sup>39,40</sup> and efforts to improve sequence quality by reprocessing sequencing reads to correct for systematic errors<sup>8,41,42</sup>. Notably, our first attempts to perform this analysis on uncorrected publicly available consensus sequences<sup>39</sup> mainly led to the identification of outlier branches containing mutations consistent with systematic errors occurring during genome assembly<sup>8</sup>.

To conclude, this work shows that saltational events are a recurrent component of SARS-CoV-2 evolution, linked to within-host evolutionary processes. Because a subset of these events seed onward human-to-human transmission, detecting and treating persistent SARS-CoV-2 infections is a tractable lever to mitigate the emergence of highly mutated, high-burden variants.

**Data and code availability:** Data and code are available at <https://github.com/blab/ncov-saltational>. We directly deposited the files generated when analyzing the sequence data on Figshare to facilitate reproducing figures<sup>43</sup>.

**Acknowledgments:** We thank Jesse Bloom and Sheri Harari for helpful discussions, and Victor Lin for support in downloading metadata from the ENA and SRA. We gratefully acknowledge the researchers and data contributors who collected the specimens, generated and deposited the raw sequence data and metadata into the European Nucleotide Archive (ENA) and Sequence Read Archive (SRA). We also thank the researchers behind the Viridian project for their work in improving the quality of publicly available SARS-CoV-2 sequence data and making this valuable resource available to the community. Analysis of chronic infections used data from GISAID<sup>44,45</sup>. We gratefully acknowledge all data contributors, i.e., the Authors and their Originating laboratories responsible for obtaining the specimens, and their Submitting laboratories for generating the genetic sequence and metadata and sharing via the GISAID Initiative, on which this research is based. The findings about the mutational signature of chronic infections of this study are based on 10,845,376 sequences accessible via <https://doi.org/10.55876/gis8.260618hy>. The supplemental table is available at <https://github.com/blab/ncov-saltational>.

**Funding:** This work is supported by NIH NIGMS R35 GM119774. KK is supported by Howard Hughes Medical Institute. TB was funded as a Howard Hughes Medical Institute Investigator.

**Author contributions:** CTK, KK and TB conceived the study. CTK developed the methods. CTK, KK and RH analyzed the data. CTK wrote the first draft. All authors reviewed and edited the manuscript.

**Declaration of interests:** We declare no competing interests.

## Methods

### Data

#### *Large SARS-CoV-2 mutation-annotated tree of high-quality sequences with associated metadata*

We rely on a publicly available high-quality mutation-annotated tree based on 4,471,579 consensus genomes that were reprocessed in an amplicon-aware manner using the Viridian tool<sup>8</sup>, which was developed to correct for systematic errors in genome assembly. Because the Viridian tree is built using UShER and does not encode indels<sup>39</sup>, our analysis of saltational evolution focuses on substitution patterns and doesn't account for insertions and deletions. We download corresponding associated run metadata from the SRA (geographic location, collection date and isolation source). We exclude samples with isolation sources that were consistent with environmental sampling, non-human hosts, non-respiratory clinical samples or laboratory experimental studies. We also only retain one sequencing run (and hence sequence) per biosample ID. We filter the Viridian tree using the *matUtils* library<sup>46</sup>. We then process the filtered mutation-annotated tree using the *BTE* package<sup>47</sup> to extract child-parent relationships, the number of synonymous and nonsynonymous mutations occurring on each branch and whether these mutations are coded, the number of children and descendants for each node and the number of synonymous and nonsynonymous sites (defined as the number of possible synonymous and nonsynonymous mutations available per gene) at each node.

#### *Recurring mutations in persistent infections, deer and mink infections*

We use mutations that have been reported to recur in studies of persistent human SARS-CoV-2 infections (Table S3)<sup>19–21,27,48</sup> as well as SARS-CoV-2 infections in minks (Table S4)<sup>49–51</sup> and deer (Table S5)<sup>50,52,53</sup>.

We follow the approach described in Hisner et al.<sup>22</sup> to identify sequences from individuals that are likely chronically infected and characterize the mutational signature of SARS-CoV-2 chronic human infections. Briefly, candidate chronic infections are manually identified from metadata associated with sequences deposited on GISAID<sup>44,45</sup>, unusually long branches in the phylogeny or the presence of mutations highly indicative of chronic infections. This approach led to the identification of 3843 sequences from distinct individuals that are likely chronically infected between April 2022 and January 2025. We compare the mutations that occur in these likely persistent infections to other circulating sequences to characterize the mutation profile of chronic infections (see below).

#### *Country-level HIV mortality estimates*

To approximate advanced HIV disease burden, we use AIDS mortality estimates from the Global Burden of Disease for year 2022<sup>23</sup>. We convert these estimates to HIV deaths per 1,000 inhabitants using population size estimates from the World Bank for that year<sup>24</sup>.

### Estimating branch-level $d_N/d_S$ from a mutation annotated tree

#### *Approach*

We aim to identify branches under positive selection by estimating the ratio of the number of nonsynonymous mutations per site to the number of synonymous mutations per site ( $d_N/d_S$ ) for each branch and gene in a mutation-annotated tree. A dedicated approach is required for two reasons. First, most branches are short, and therefore contain few mutations, so reliable inference requires explicit uncertainty quantification and accounting for sparse mutation counts. Second, the size of the dataset requires a method that remains computationally tractable on a mutation-annotated tree containing millions of branches.

#### *Notation*

The index  $i$  refers to branches in the phylogeny and superscript  $g$  to the gene on which we infer the  $d_N/d_S$  ratio. To estimate this quantity, we rely on the number of nonsynonymous and synonymous mutations ( $N_i^g$  and  $S_i^g$ ) on branch  $i$  within gene  $g$ , and the number of nonsynonymous and synonymous sites ( $L_{N,i}^g$  and  $L_{S,i}^g$ ) at the beginning of branch  $i$  within gene  $g$ . For each branch and gene, we introduce three parameters: (i) the number of nonsynonymous mutations per site  $d_{N,i}^g$ , (ii) the number of synonymous mutations per site  $d_{S,i}^g$  and (iii) the ratio between these quantities  $\omega_i^g = d_{N,i}^g/d_{S,i}^g$ .

#### *Likelihood of the data*

We assume that  $N_i^g$  and  $S_i^g$  follow Poisson distributions parametrized as:

$$N_i^g \mid d_{N,i}^g \sim P(d_{N,i}^g \cdot L_{N,i}^g)$$

$$S_i^g \mid d_{S,i}^g \sim P(d_{S,i}^g \cdot L_{S,i}^g)$$

#### *Prior distributions*

We assume that the prior of the number of synonymous and nonsynonymous mutation per site are Gamma distributed:

$$d_{N,i}^g \sim \Gamma(\alpha_N^g, \beta_N^g)$$

$$d_{S,i}^g \sim \Gamma(\alpha_S^g, \beta_S^g)$$

We can derive the prior distribution of  $\omega_i^g$  as the ratio of two independent Gamma distributions:

$$\omega_i^g \sim \frac{\beta_S^g}{\beta_N^g} \cdot \text{BetaPrime}(\alpha_N^g, \alpha_S^g)$$

We set the prior means for the number of nonsynonymous and synonymous mutation per site in segment  $g$  to the corresponding tree-wide empirical ratios:

$$\frac{\alpha_N^g}{\beta_N^g} = \frac{\sum_i N_i^g}{\sum_i L_{N,i}^g}$$

$$\frac{\alpha_S^g}{\beta_S^g} = \frac{\sum_i S_i^g}{\sum_i L_{S,i}^g}$$

We explore a range of values for the coefficient of variation (CV) of the number of nonsynonymous and synonymous mutations per site and used posterior predictive checks (see below) to identify prior parametrizations that best reproduce the observed data.

### *Posterior distribution of parameters*

The Gamma distribution is a conjugate prior of the Poisson distribution with exposure. We thus derive the posterior distribution of model parameters as:

$$d_{N,i}^g \mid N_i^g, L_{N,i}^g \sim \Gamma(\alpha_N^g + N_i^g, \beta_N^g + L_{N,i}^g)$$

$$d_{S,i}^g \mid S_i^g, L_{S,i}^g \sim \Gamma(\alpha_S^g + S_i^g, \beta_S^g + L_{S,i}^g)$$

We deduce the posterior distribution of  $\omega_i^g$  as the ratio of two independent Gamma distributions:

$$\omega_i^g \mid N_i^g, S_i^g, L_{N,i}^g, L_{S,i}^g \sim \frac{\beta_S^g + L_{S,i}^g}{\beta_N^g + L_{N,i}^g} \cdot \text{BetaPrime}(\alpha_N^g + N_i^g, \alpha_S^g + S_i^g)$$

### *Posterior predictive checks*

We perform posterior predictive checks (Figure S6) by drawing new observed counts of nonsynonymous and synonymous mutations ( $N_i^{g,new}, S_i^{g,new}$ ) from:

$$N_i^{g,new} \sim NB\left(r = \alpha_N^g + N_i^g, p = \frac{\beta_N^g + L_{N,i}^g}{2L_{N,i}^g + \beta_N^g}\right)$$

$$S_i^{g,new} \sim NB\left(r = \alpha_S^g + S_i^g, p = \frac{\beta_S^g + L_{S,i}^g}{2L_{S,i}^g + \beta_S^g}\right)$$

We compute the median number of branches  $n_{branches}^{g,new}(N, S)$  with each combination of nonsynonymous and synonymous mutations ( $N, S$ ) across posterior simulations within gene  $g$ . We determine the coefficient of variation for the prior that maximizes the Pearson correlation between  $\log(n_{branches}^{g,obs} + 1)$  and  $\log(n_{branches}^{g,new} + 1)$  pooled across genes and ( $N, S$ ) combinations, where  $n_{branches}^{g,obs}(N, S)$  is the observed number of branches with  $N$  nonsynonymous and  $S$  synonymous mutations in gene  $g$  across the phylogeny (Figure S7-S8) (CV of 1.8).

### *Identifying branches characterized by $d_N/d_S > 1$ at the gene level*

To identify branches characterized by  $d_N/d_S > 1$  on gene  $g$ , we compute the Bayes Factor  $BF_i^g$  between the two competing hypotheses  $H_0: \omega_i^g \leq 1$  and  $H_1: \omega_i^g > 1$  as the ratio of posterior and prior odds for  $\omega_i^g$  being greater than 1:

$$BF_i^g(\omega > 1) = \frac{P[\omega_i^g > 1 \mid N_i^g, S_i^g, L_{N,i}^g, L_{S,i}^g]}{(1 - P[\omega_i^g > 1 \mid N_i^g, S_i^g, L_{N,i}^g, L_{S,i}^g])} \bigg/ \frac{P[\omega_i^g > 1]}{(1 - P[\omega_i^g > 1])}$$

The posterior and prior probabilities  $P[\omega_i^g > 1 \mid N_i^g, S_i^g, L_{N,i}^g, L_{S,i}^g]$  and  $P[\omega_i^g > 1]$  can be computed as:

$$P[\omega_i^g > 1 \mid N_i^g, S_i^g, L_{N,i}^g, L_{S,i}^g] = 1 - I\left(\frac{\beta_N^g + L_{N,i}^g}{\beta_N^g + L_{N,i}^g + \beta_S^g + L_{S,i}^g}; \alpha_N^g + N_i^g, \alpha_S^g + S_i^g\right)$$

$$P[\omega_i^g > 1] = 1 - I\left(\frac{\beta_N^g}{\beta_N^g + \beta_S^g}; \alpha_N^g, \alpha_S^g\right)$$

where  $I(x; a, b)$  is the regularized incomplete beta function evaluated in  $x$  with coefficients  $a$  and  $b$ . We classify branches as characterized by  $\omega_i^g > 1$  if  $BF_i^g$  is above 20, interpreted as strong evidence for model  $H_1$  under the Kass-Raftery scale<sup>9</sup>.

#### *Identifying branches characterized by genome-wide $d_N/d_S > 1$*

To assess evidence for positive selection at the genome level, we define branch-specific genome-wide number of nonsynonymous and synonymous mutations per site ( $d_{N,i}^{genome}$  and  $d_{S,i}^{genome}$ ), and the corresponding genome-wide ratio as:

$$d_{N,i}^{genome} = \frac{\sum_g d_{N,i}^g \cdot L_{N,i}^g}{\sum_g L_{N,i}^g}$$

$$d_{S,i}^{genome} = \frac{\sum_g d_{S,i}^g \cdot L_{S,i}^g}{\sum_g L_{S,i}^g}$$

$$\omega_i^{genome} = d_{N,i}^{genome} / d_{S,i}^{genome}$$

We use a Monte Carlo approach to explore the posterior distribution of  $\omega_i^{genome}$ . To do so, we draw  $M = 1000$  samples from each of the gene-specific posterior distributions of  $d_{N,i}^g$  and  $d_{S,i}^g$ . Using the formula above, we compute the corresponding  $M$  draws  $(\omega_{i,m}^{genome})_{1 \leq m \leq M}$  from the posterior distribution of  $\omega_i^{genome}$ . This enables us to approximate the posterior probability of  $\omega_i^{genome} > 1$  as:

$$P[\omega_i^{genome} > 1 \mid (N_i^g, S_i^g, L_{N,i}^g, L_{S,i}^g)_g] \approx \frac{1}{M} \sum_{m=1}^M \mathbf{1}(\omega_{i,m}^{genome} > 1)$$

Similarly, we approximate the prior probability of  $\omega_i^{genome} > 1$  by drawing from the gene-specific prior distributions of the number of nonsynonymous and synonymous mutations per site. This

enables us to compute  $BF_i^{genome} (\omega > 1)$  by compute the ratio between the posterior and the prior odds of  $\omega_i^{genome} > 1$ .

To validate this genome-wide approach, we perform posterior predictive checks at the genome level by drawing synonymous and nonsynonymous mutation counts from the gene and branch-specific posterior predictive distributions and summing these counts across genes within each branch. Using the CV of 1.8 that maximizes the correlation between true and predicted number of branches at the gene level (see above) also provides a good predictive accuracy at the genome level (Pearson correlation coefficient of 0.9).

## Characterizing the signature of SARS-CoV-2 saltational evolution

### *Definition of saltational events*

We define saltational events as single branches in the mutation-annotated tree with:

1. An unusually high number of nonsynonymous mutations (at least 5 at the genome level, corresponding to greater than the 99.5<sup>th</sup> percentile)
2. Strong evidence for positive selection (measured by a Bayes Factor at the genome level or within a gene greater than 20<sup>9</sup>)

Stricter thresholds rapidly reduce the number of branches classified as saltational (Figure S5). Branches identified as saltational under less stringent criteria still differ from background branches, though with a weaker separation. We choose this conservative threshold combination (Bayes Factor above 20 and at least 5 nonsynonymous mutations) as a compromise between identifying a sufficiently large set of branches for detailed downstream analyses (~1000) while ensuring a stringent definition of saltational evolution. To mitigate the impact of recombination, we exclude from potential saltational events branches annotated as leading to new recombining Pango lineages in the Viridian tree<sup>8</sup>. Here, we classify individual branches in a large phylogeny as consistent with saltational evolution. In practice, some saltational events may be distributed across several consecutive branches, for example because a single infection can span several branches<sup>54</sup> or because of artifactual recombinant placements<sup>55</sup> around long branches, which could decrease our power to detect them by diluting the signal.

### *Temporal signature of saltational events*

We define the year of each branch as the year of the node defining the end of this branch. For internal nodes or terminal nodes with missing sample collection date information, we run Chronumental<sup>40</sup> to infer node times. We compute the proportion of branches that are saltational each year and compute 95% Wilson confidence intervals around these proportions.

### *Geographical signature of saltational events*

We determine internal nodes' geographical locations based on the geographies of their descendants. If all descendants from a node come from the same country, we attribute the internal node to this country. Otherwise, we annotate this node as having "Multiple descendants". We use the same approach to reconstruct subregions (defined using the United Nations geoscheme) and continents of internal nodes.

### *Mutational signature of saltational events*

To evaluate whether saltational branches have a distinct mutational profile, we use Fisher's exact tests to assess whether nonsynonymous mutations are more likely to occur at specific amino acid positions in saltational branches compared with non saltational branches. Using mutations as the statistical unit enables us to account for the higher number of mutations occurring on saltational branches (Figure 1D). By contrast, a branch-level analysis of whether a position is more likely to be mutated in saltational branches would be biased by the excess of mutations in saltational branches, which could lead to identifying positions as more mutated on saltational branches only because mutations occur more frequently on this type of branch. We identify 2,242 amino-acid positions that are mutated (nonsynonymously) at least twice in saltational branches. For each of these amino-acid positions, we run Fisher tests and report p-values adjusted for multiple testing (Benjamini-Hochberg correction). We use a Type I error threshold  $\alpha$  of 0.05.

To complement this position-level analysis, we assess whether specific amino-acid changes are more likely to occur on saltational branches (e.g. S:484K rather than any mutation occurring at position S:484). As not all mutations are possible on all branches (e.g. S:484K is not possible if a K is already present at position 484), we perform Fisher's exact tests using only branches on which that mutation is possible (e.g. branches without a K at position 484 in that specific example). For a given mutation  $m$ , the odds ratio (OR) of occurrence on saltational branches compared with non-saltational ones is equal to:

$$OR_{salt}^m = \frac{N_m^{salt} \cdot N_m^{non\ salt}}{N_{other}^{salt} \cdot N_{other}^{non\ salt}}$$

Here,  $N_m^x$  is the number of occurrences of mutation  $m$  on branches of type  $x$  (saltational or non-saltational) where the mutation is possible, and  $N_{other}^x$  is the number of other mutations on branches of type  $x$  on which mutation  $m$  is possible.

### *Comparison of the mutational signature of saltational events with that of persistent infections*

We next compare the log odds ratio of occurrence of mutations on saltational branches (vs non-saltational branches) with the log odds ratio of occurrence of mutations in chronic infections compared with background sequences<sup>22</sup>. As odds ratio can be null (if the mutation never occurs), we rely on a modified version of odds ratios:

$$\widetilde{OR}_{salt}^m = \frac{(N_m^{salt} + 1) \cdot (N_{other}^{non\ salt} + 1)}{(N_{other}^{salt} + 1) \cdot (N_m^{non\ salt} + 1)}$$

To compute the odds ratio of a mutation occurring in sequences from chronically infected individuals vs background sequences, we rely on a large manually assembled dataset containing 3843 sequences from distinct individuals that are likely chronically infected<sup>22</sup>. We compare the mutations that occur in the sequences from likely chronic infections to other circulating sequences as:

$$\widetilde{OR}_{chronic}^m = \frac{(N_m^{chronic} + 1) \cdot (N_{other}^{circulating} + 1)}{(N_{other}^{chronic} + 1) \cdot (N_m^{circulating} + 1)}$$

where  $N_m^x$  is the number of occurrences of mutation  $m$  occurs on sequences of type  $x$  (chronic or other circulating, and  $N_{other}^x$  is the number of other mutations on sequences of type  $x$ . This dataset doesn't report whether a mutation is possible (if the derived amino-acid differs from the ancestral one). However, we don't expect this to impact the OR computation meaningfully as the OR of observing a mutation in saltational branches is little impacted by accounting for whether a nonsynonymous mutation is possible on a given branch (Figure S9).

To perform the comparison between  $\widetilde{OR}_{chronic}^m$  and  $\widetilde{OR}_{salt}^m$ , we focus on mutations that are observed at least once in either saltational branches or sequences from chronic infection ( $N_m^{salt} > 0$  or  $N_m^{chronic} > 0$ ), and that were observed at least once in both of the OR computations meaning:

$$N_m^{salt} + N_m^{non\ salt} > 0 \text{ and } N_m^{chronic} + N_m^{circulating} > 0.$$

### *Quantifying saltational branches' onward transmission patterns*

We define descendants from saltational branches as terminal tips (corresponding to distinct collected samples) descending from the saltational branch. We determine a saltational branch to have spread to multiple distinct individuals if we can identify descendant samples with distinct metadata (either age, sex or geographic location information). Many samples don't have fine-grained metadata information (age and sex) and it is hence not always possible to determine whether it spread to multiple distinct individuals. We also determine whether there is evidence for a saltational event to have spread across geographies (countries, subregions or continents) if descendant samples are collected in multiple geographies

We investigate whether the mutational composition of saltational branches is associated with the number of descendants from this saltational event. To do so, we model descendant counts as a function of the number of synonymous and nonsynonymous mutations in each gene. We define descendant counts as 0 for branches without detected descendants and as the number of terminal tips descending from that branch minus 1 otherwise. That way, our metric of descendant counts reflects the number of additional descendants observed beyond the initial saltational branch. We estimate the relationship between mutation counts and descendant counts using a negative binomial regression. More specifically, let  $D_i$  denote the number of descendants from saltational branch  $i$ . We model  $D_i$  as:

$$\log(\mu_i) = \beta_0 + \sum_g (\beta_N^g \cdot N_i^g + \beta_S^g \cdot S_i^g)$$

where  $N_i^g$  and  $S_i^g$  are respectively the number of nonsynonymous and synonymous mutations in gene  $g$  on branch  $i$  and  $\beta_N^g$  and  $\beta_S^g$  are parameters to be estimated. We fit the negative binomial regression in R using the *glmmTMB* package<sup>56,57</sup>. We decided to perform a negative binomial regression over a Poisson regression after performing a likelihood ratio test of the two models that yielded a p-value  $< 2.2 \cdot 10^{-16}$  for the negative binomial model over the Poisson one, suggesting overdispersion in the data that needs to be accounted for. As a small number of saltational branches have a very high number of descendants, their inclusion in the analysis can impact parameter estimates (Figure S10). To mitigate the impact of a few branches with many

descendants, we present estimates by restricting the analysis to saltational events that have less than 50,000 descendants in the global phylogeny.

#### *Assessing the dynamics of mutations enriched in saltational branches at the between-host level*

We compare the frequency at which mutations associated with saltational branches give rise to descendants with the frequency at which these mutations fix at the global level. We measure how often a mutation fixed throughout the pandemic by counting the number of non-recombinant Nextstrain clades in the open all-time Nextstrain public tree<sup>58,59</sup> (downloaded on April 20, 2026) in which that mutation fixed. We only focus on mutations that are enriched in saltational branches. We consider that a mutation fixes within a clade if it is present in at least 80% of the tips descending from this clade and if the mutation arose on the branch leading to the clade or within the clade. We then group mutations into quintiles based on the proportion of saltational branches carrying that mutation that have descendants. For each quintile, we compute the mean number of clades in which these mutations fixed. We also compute uncertainty around the mean using 95% bootstrap confidence intervals from 2000 bootstrap draws. As some saltational branches with descendants are clade-defining, we perform a sensitivity analysis by removing saltational branches with more than 100 or 1000 descendants from the computations.

#### **References**

1. Hill, V. *et al.* The origins and molecular evolution of SARS-CoV-2 lineage B.1.1.7 in the UK. *Virus Evol.* **8**, veac080 (2022).
2. Tegally, H. *et al.* Detection of a SARS-CoV-2 variant of concern in South Africa. *Nature* **592**, 438–443 (2021).
3. Faria, N. R. *et al.* Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. *Science* **372**, 815–821 (2021).
4. Viana, R. *et al.* Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa. *Nature* **603**, 679–686 (2022).
5. Volz, E. Fitness, growth and transmissibility of SARS-CoV-2 genetic variants. *Nat. Rev. Genet.* **24**, 724–734 (2023).
6. Machkovech, H. M. *et al.* Persistent SARS-CoV-2 infection: significance and implications. *Lancet Infect. Dis.* **24**, e453–e462 (2024).
7. Sigal, A., Neher, R. A. & Lessells, R. J. The consequences of SARS-CoV-2 within-host persistence. *Nat. Rev. Microbiol.* **23**, 288–302 (2025).

8. Hunt, M. *et al.* Addressing pandemic-wide systematic errors in the SARS-CoV-2 phylogeny. *Nat. Methods* **23**, 653–662 (2026).
9. Kass, R. E. & Raftery, A. E. Bayes Factors. *J. Am. Stat. Assoc.* **90**, 773 (1995).
10. Starr, T. N. *et al.* Prospective mapping of viral mutations that escape antibodies used to treat COVID-19. *Science* **371**, 850–854 (2021).
11. Starr, T. N., Greaney, A. J., Dingens, A. S. & Bloom, J. D. Complete map of SARS-CoV-2 RBD mutations that escape the monoclonal antibody LY-CoV555 and its cocktail with LY-CoV016. *Cell Rep. Med.* **2**, 100255 (2021).
12. Dong, J. *et al.* Genetic and structural basis for SARS-CoV-2 variant neutralization by a two-antibody cocktail. *Nat. Microbiol.* **6**, 1233–1244 (2021).
13. Greaney, A. J. *et al.* Mapping mutations to the SARS-CoV-2 RBD that escape binding by different classes of antibodies. *Nat. Commun.* **12**, 4196 (2021).
14. Dadonaite, B. *et al.* Spike mutations that affect the function and antigenicity of recent KP.3.1.1-like SARS-CoV-2 variants. *J. Virol.* **99**, e0142325 (2025).
15. Zhang, J. *et al.* Structural impact on SARS-CoV-2 spike protein by D614G substitution. *Science* **372**, 525–530 (2021).
16. Markov, P. V. *et al.* The evolution of SARS-CoV-2. *Nat. Rev. Microbiol.* **21**, 361–379 (2023).
17. Carabelli, A. M. *et al.* SARS-CoV-2 variant biology: immune escape, transmission and fitness. *Nat. Rev. Microbiol.* **21**, 162–177 (2023).
18. Velasquez-Reyes, J. M. *et al.* Characterisation of a persistent SARS-CoV-2 infection lasting more than 750 days in a person living with HIV: a genomic analysis. *Lancet Microbe* **6**, 101122 (2025).
19. Ghafari, M. *et al.* Prevalence of persistent SARS-CoV-2 in a large community surveillance study. *Nature* **626**, 1094–1101 (2024).
20. Wilkinson, S. A. J. *et al.* Recurrent SARS-CoV-2 mutations in immunodeficient patients. *Virus Evol.* **8**, veac050 (2022).

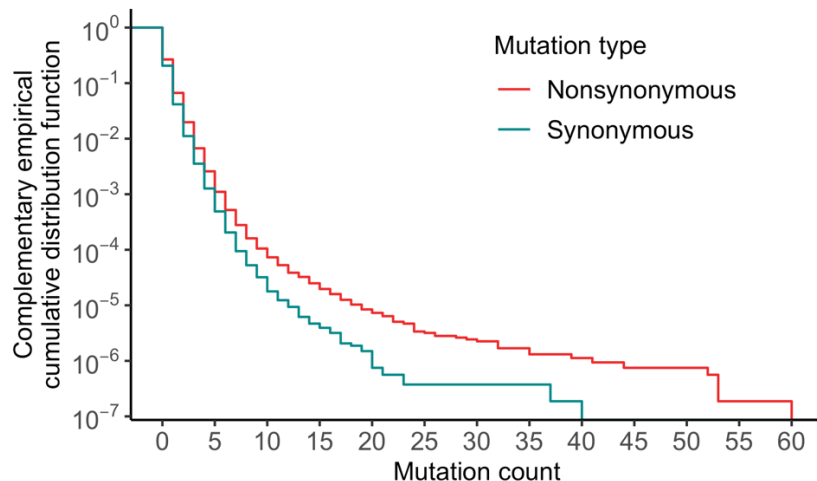
21. Khurana, M. P. *et al.* Large-scale genomic surveillance reveals immunosuppression drives mutation dynamics in persistent SARS-CoV-2 infections. *medRxiv* (2025) doi:10.1101/2025.02.10.25321987.
22. Hisner, R. & Martin, D. P. Genetic evidence indicates the evolutionary importance of the SARS-CoV-2 ORF9b protein. *bioRxiv* (2026) doi:10.64898/2026.02.23.707522.
23. Global Burden of Disease Collaborative Network. Global Burden of Disease Study 2023 (GBD 2023). *Seattle, United States: Institute for Health Metrics and Evaluation (IHME)* (2025).
24. World Bank Open Data. *World Bank Open Data* <https://data.worldbank.org/>.
25. Karim, F. *et al.* Clearance of persistent SARS-CoV-2 associates with increased neutralizing antibodies in advanced HIV disease post-ART initiation. *Nat. Commun.* **15**, 2360 (2024).
26. Raglow, Z. *et al.* SARS-CoV-2 shedding and evolution in patients who were immunocompromised during the omicron period: a multicentre, prospective analysis. *Lancet Microbe* **5**, e235–e246 (2024).
27. Harari, S., Miller, D., Fleishon, S., Burstein, D. & Stern, A. Using big sequencing data to identify chronic SARS-Coronavirus-2 infections. *Nat. Commun.* **15**, 648 (2024).
28. Grenfell, B. T. *et al.* Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* **303**, 327–332 (2004).
29. Sesta, L. & Neher, R. A. Epistasis and the changing fitness landscapes of SARS-CoV-2. *bioRxiv* (2026) doi:10.64898/2026.03.12.711354.
30. Gonzalez-Reiche, A. S. *et al.* Sequential intrahost evolution and onward transmission of SARS-CoV-2 variants. *Nat. Commun.* **14**, 3235 (2023).
31. Volz, E. *et al.* Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. *Nature* **593**, 266–269 (2021).
32. Raglow, Z. & Luring, A. S. Virus evolution in prolonged infections of immunocompromised individuals. *Clin. Chem.* **71**, 109–118 (2025).

33. Xue, K. S. *et al.* Parallel evolution of influenza across multiple spatiotemporal scales. *Elife* **6**, e26875 (2017).
34. García-Martínez de Artola, D. *et al.* A prolonged Delta SARS-CoV-2 infection during the Omicron wave in an HIV immunosuppressed patient. *Int. J. Infect. Dis.* **162**, 108233 (2026).
35. Nabieva, E. *et al.* A highly divergent sample from a nearly extinct SARS-CoV-2 lineage in a patient with long-term COVID-19. *Front. Cell. Infect. Microbiol.* **15**, 1623390 (2025).
36. Cavalcanti, D. M. *et al.* Evaluating the impact of two decades of USAID interventions and projecting the effects of defunding on mortality up to 2030: a retrospective impact evaluation and forecasting analysis. *Lancet* **406**, 283–294 (2025).
37. Brink, D. T. *et al.* Impact of an international HIV funding crisis on HIV infections and mortality in low-income and middle-income countries: a modelling study. *Lancet HIV* **12**, e346–e354 (2025).
38. Brito, A. F. *et al.* Global disparities in SARS-CoV-2 genomic surveillance. *Nat. Commun.* **13**, 7003 (2022).
39. 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).
40. Sanderson, T. Chronumental: time tree estimation from very large phylogenies. *bioRxiv* (2021) doi:10.1101/2021.10.27.465994.
41. Turakhia, Y. *et al.* Stability of SARS-CoV-2 phylogenies. *PLoS Genet.* **16**, e1009175 (2020).
42. De Maio, N. *et al.* Issues with SARS-CoV-2 sequencing data. *virological.org* <https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473> (2020).
43. Tran-Kiem, C. blab/ncov-saltational workflow outputs. *figshare* <https://doi.org/10.6084/M9.FIGSHARE.32736657> (2026).
44. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data – from vision to reality. *Euro Surveill.* **22**, (2017).
45. GISAID - [gisaid.org](https://gisaid.org/). <https://gisaid.org/>.

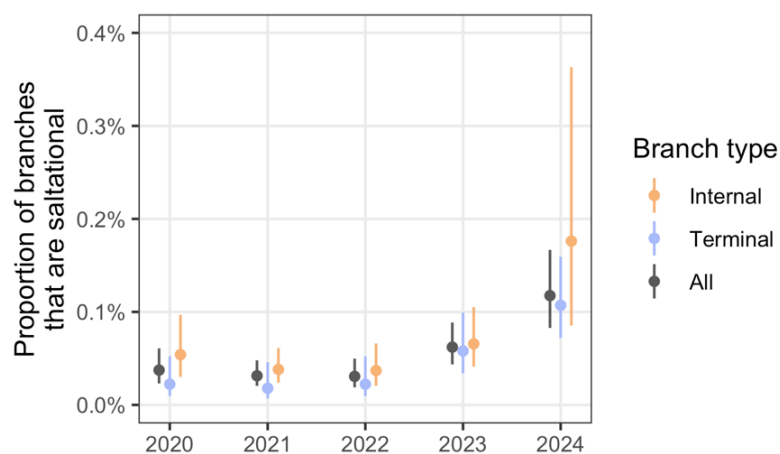
46. McBroome, J. *et al.* A daily-updated database and tools for comprehensive SARS-CoV-2 mutation-annotated trees. *Mol. Biol. Evol.* **38**, 5819–5824 (2021).
47. BTE — usher\_wiki 0.0.2 documentation. <https://usher-wiki.readthedocs.io/en/latest/bte.html>.
48. Harari, S. *et al.* Drivers of adaptive evolution during chronic SARS-CoV-2 infections. *Nat. Med.* **28**, 1501–1508 (2022).
49. Zhou, J. *et al.* Mutations that adapt SARS-CoV-2 to mink or ferret do not increase fitness in the human airway. *Cell Rep.* **38**, 110344 (2022).
50. Tan, C. C. S. *et al.* Transmission of SARS-CoV-2 from humans to animals and potential host adaptation. *Nat. Commun.* **13**, 2988 (2022).
51. Iglesias-Caballero, M. *et al.* Genomic context of SARS-CoV-2 outbreaks in farmed mink in Spain during pandemic: Unveiling host adaptation mechanisms. *Int. J. Mol. Sci.* **25**, 5499 (2024).
52. Marques, A. D. *et al.* Evolution of SARS-CoV-2 in white-tailed deer in Pennsylvania 2021-2024. *PLoS Pathog.* **21**, e1012883 (2025).
53. Feng, A. *et al.* Transmission of SARS-CoV-2 in free-ranging white-tailed deer in the United States. *Nat. Commun.* **14**, 4078 (2023).
54. Didelot, X., Fraser, C., Gardy, J. & Colijn, C. Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. *Mol. Biol. Evol.* msw075 (2017).
55. Turakhia, Y. *et al.* Pandemic-scale phylogenomics reveals the SARS-CoV-2 recombination landscape. *Nature* **609**, 994–997 (2022).
56. Brooks, M. *et al.* GlimmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *R J.* **9**, 378 (2017).
57. McGillicuddy, M., Warton, D. I., Popovic, G. & Bolker, B. M. Parsimoniously fitting large multivariate random effects in glimmTMB. *J. Stat. Softw.* **112**, (2025).
58. nextstrain.org. Nextstrain SARS-CoV-2 all-time global public tree. <https://nextstrain.org/ncov/open/global/all-time>.

59. Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).

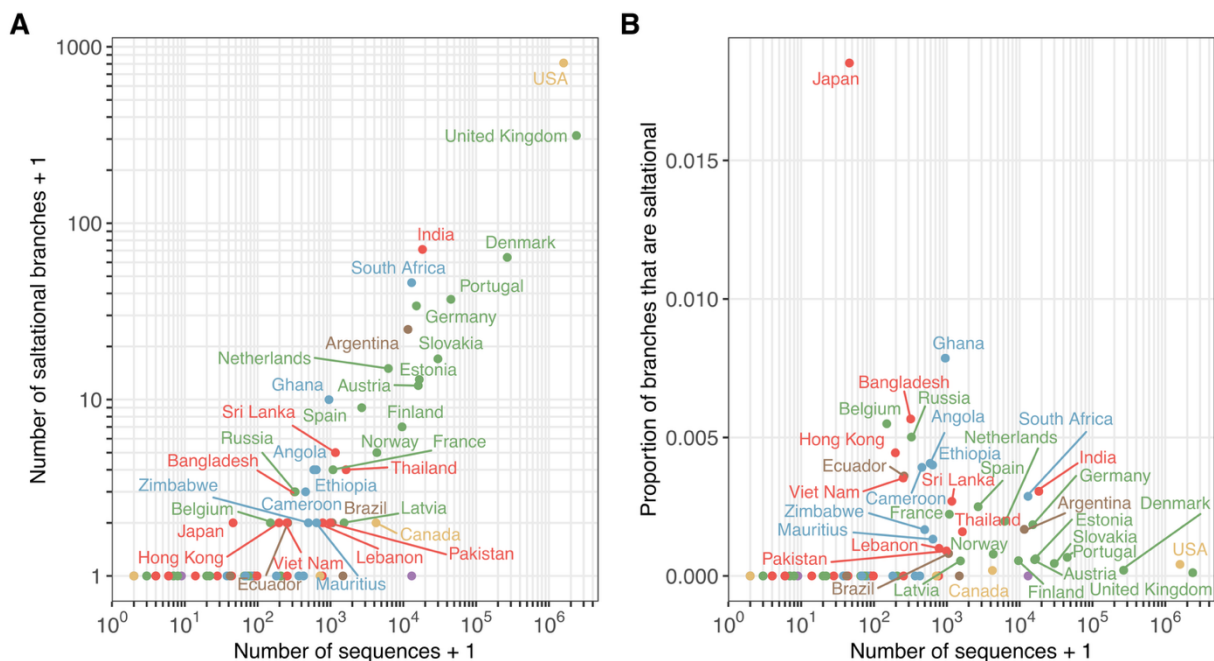
## Supplementary figures



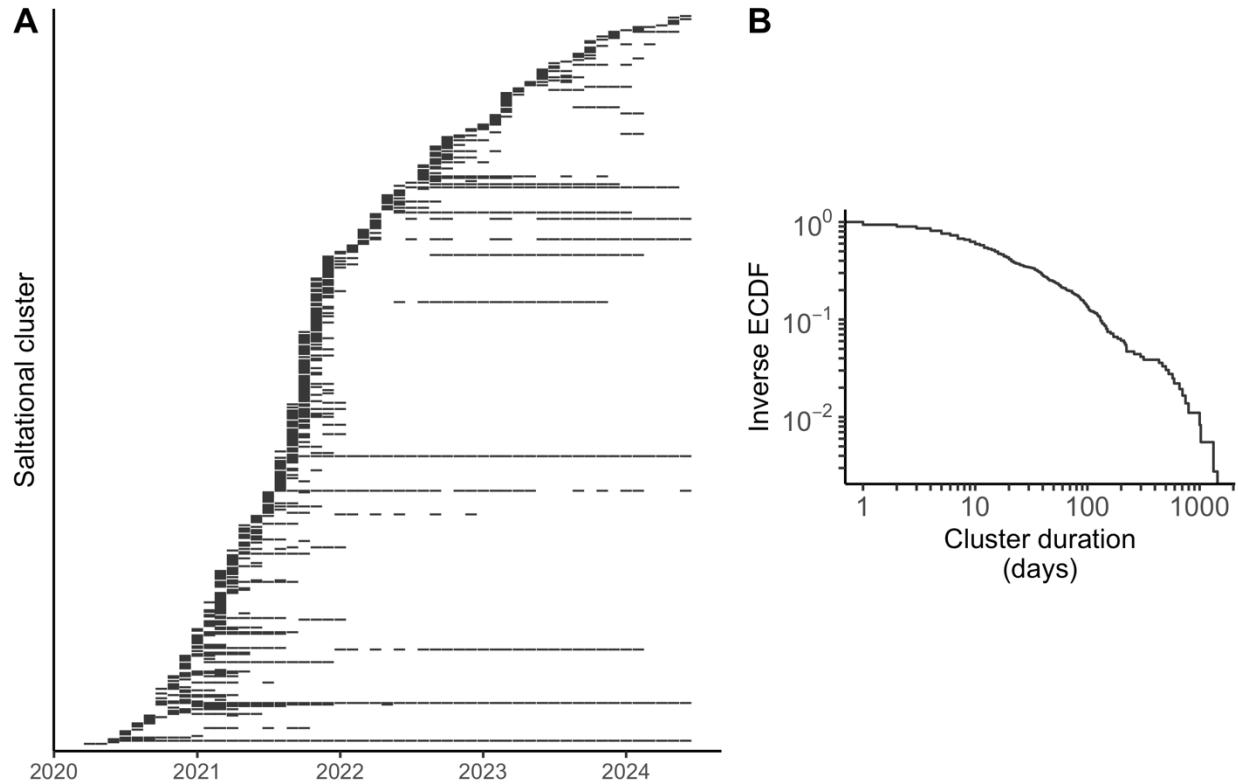
**Figure S1: Complementary empirical cumulative distribution function of mutation count across branches in the phylogeny and by mutation type (nonsynonymous or synonymous).** For each mutation count value on the x-axis, the lines indicate the proportion of branches that have at least that many mutations.



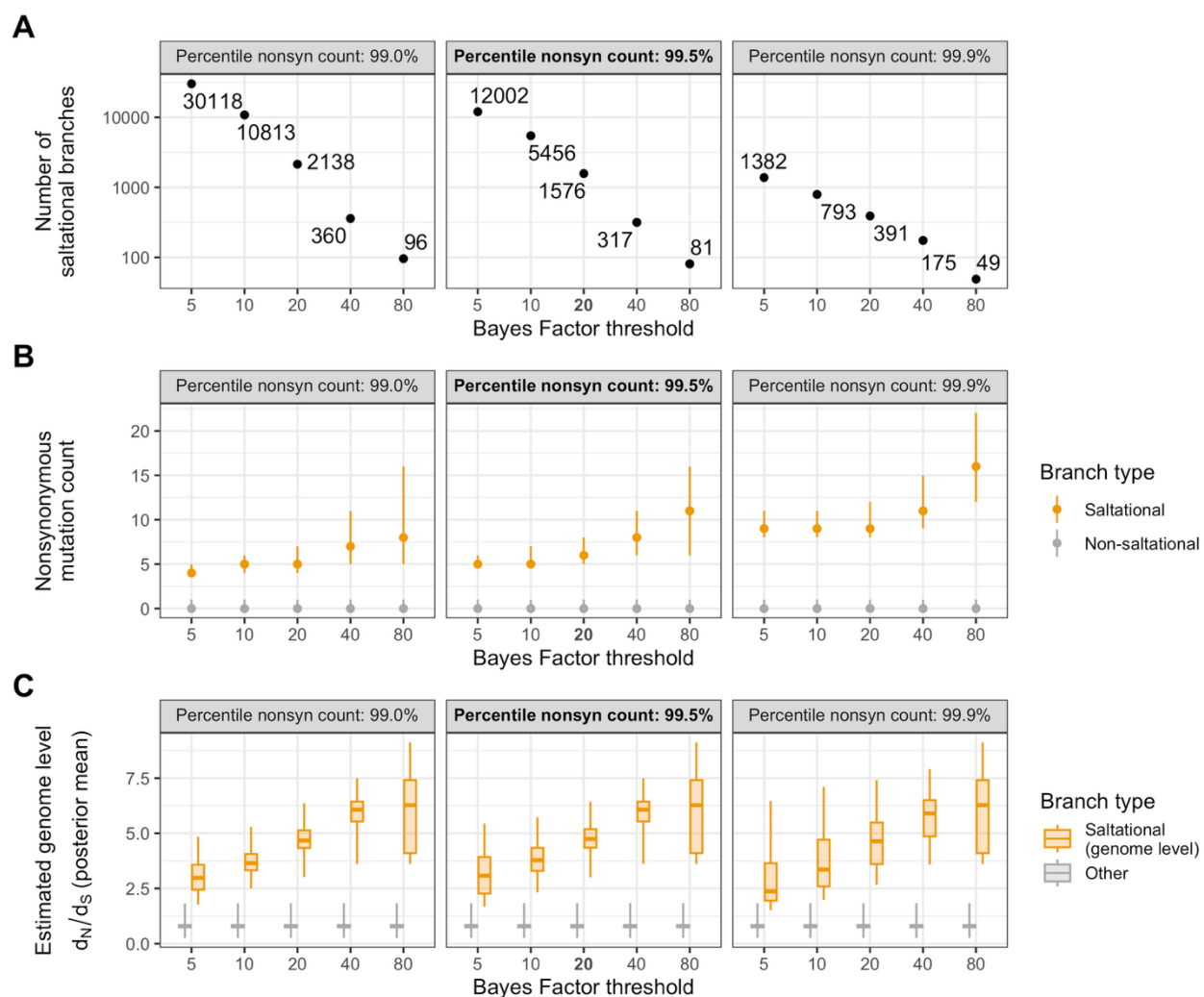
**Figure S2: Sensitivity analysis based on a subsampled tree with the same number of tips every year.** We downsampled the tree used in the analysis by randomly selecting, for each year, the same number of tips as available for 2025. The plot shows the proportion of branches classified as saltational by year.



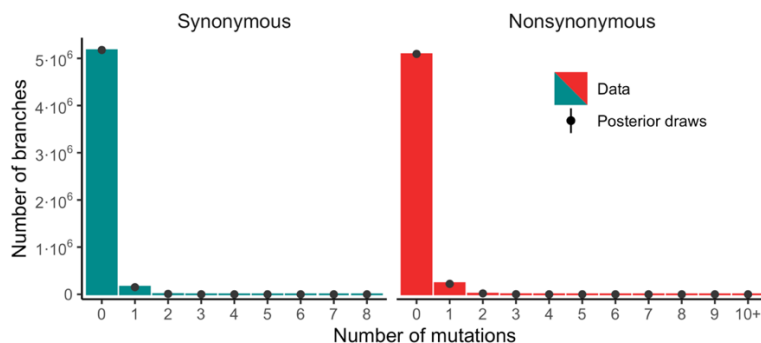
**Figure S3: Relationship between the number of saltational branches identified in a country and the number of sequences from this country included in this analysis. A.** Number of branches identified as saltational + 1 and **B.** proportion of branches that are saltational as a function of the number of sequences from that country.



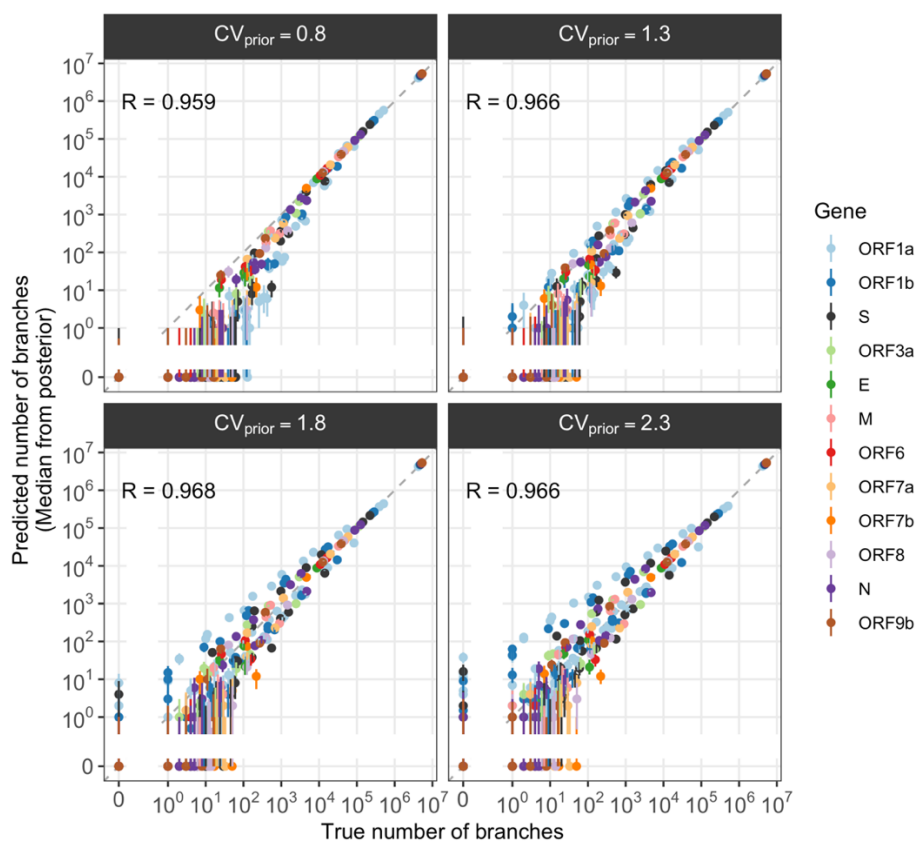
**Figure S4: Persistence of clusters descending from saltational events.** **A.** Temporal window during which clusters descending from saltational events are observed. Each line on the y-axis correspond to a cluster descending from saltational events. Clusters are ordered by inferred saltational branch date. A colored tile indicates that this cluster has sequences collected during the corresponding month. **B.** Inverse empirical cumulative distribution function (ECDF) of cluster durations, measures as the delay between the sequence collection date of the latest and earliest sequence descending from this saltational branch.



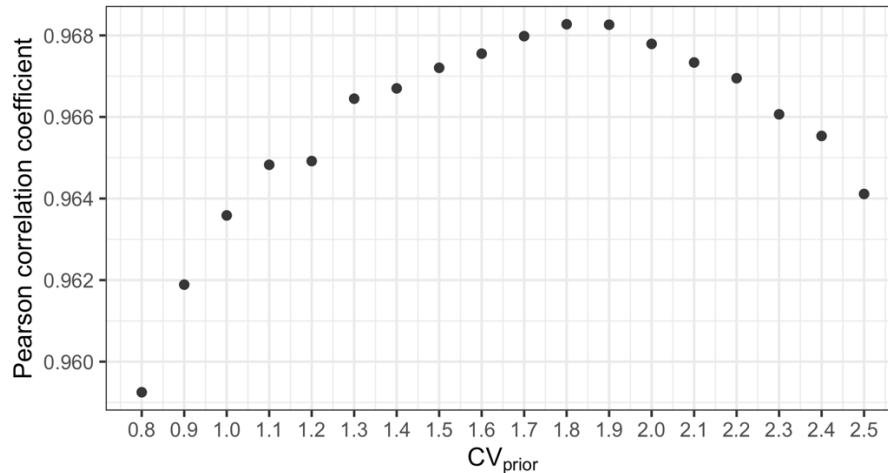
**Figure S5: Impact of saltational branch definition on the characteristics of saltational branches.** We explore the impact of varying the Bayes Factor threshold used to identify branches under positive selection (x-axis) and the minimum genome-wide nonsynonymous mutation count required for classification as saltational (facets). **A.** Number of branches classified as saltational under different criteria. **B.** Median nonsynonymous mutation count on saltational and non-saltational branches under different criteria. Vertical segments indicate interquartile ranges. **C.** Distribution of posterior mean genome-wide  $d_N/d_S$  ratio for branches classified as saltational at the genome level and for other branches. Boxplots indicate 2.5th, 25th, 50th, 75th and 97.5th percentiles.



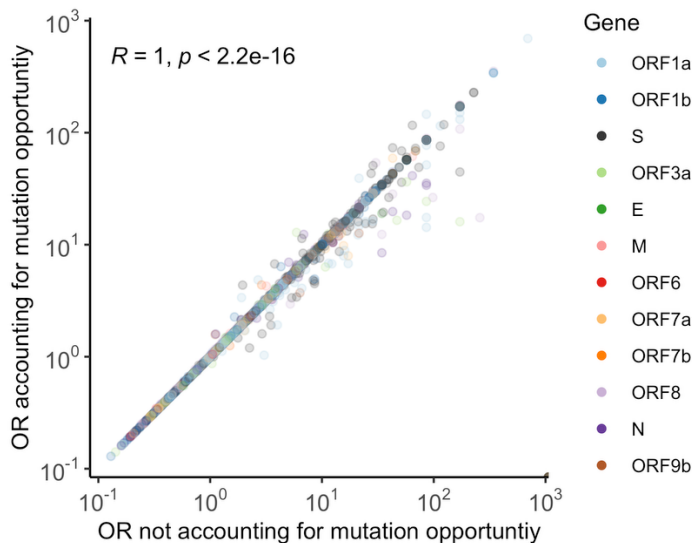
**Figure S6: Distribution of synonymous (left) and nonsynonymous (right) mutation counts on Spike by branch.** Bars depict observed counts in the large SARS-CoV-2 phylogeny. Points and vertical segments depict median and 95% credible intervals from posterior draws.



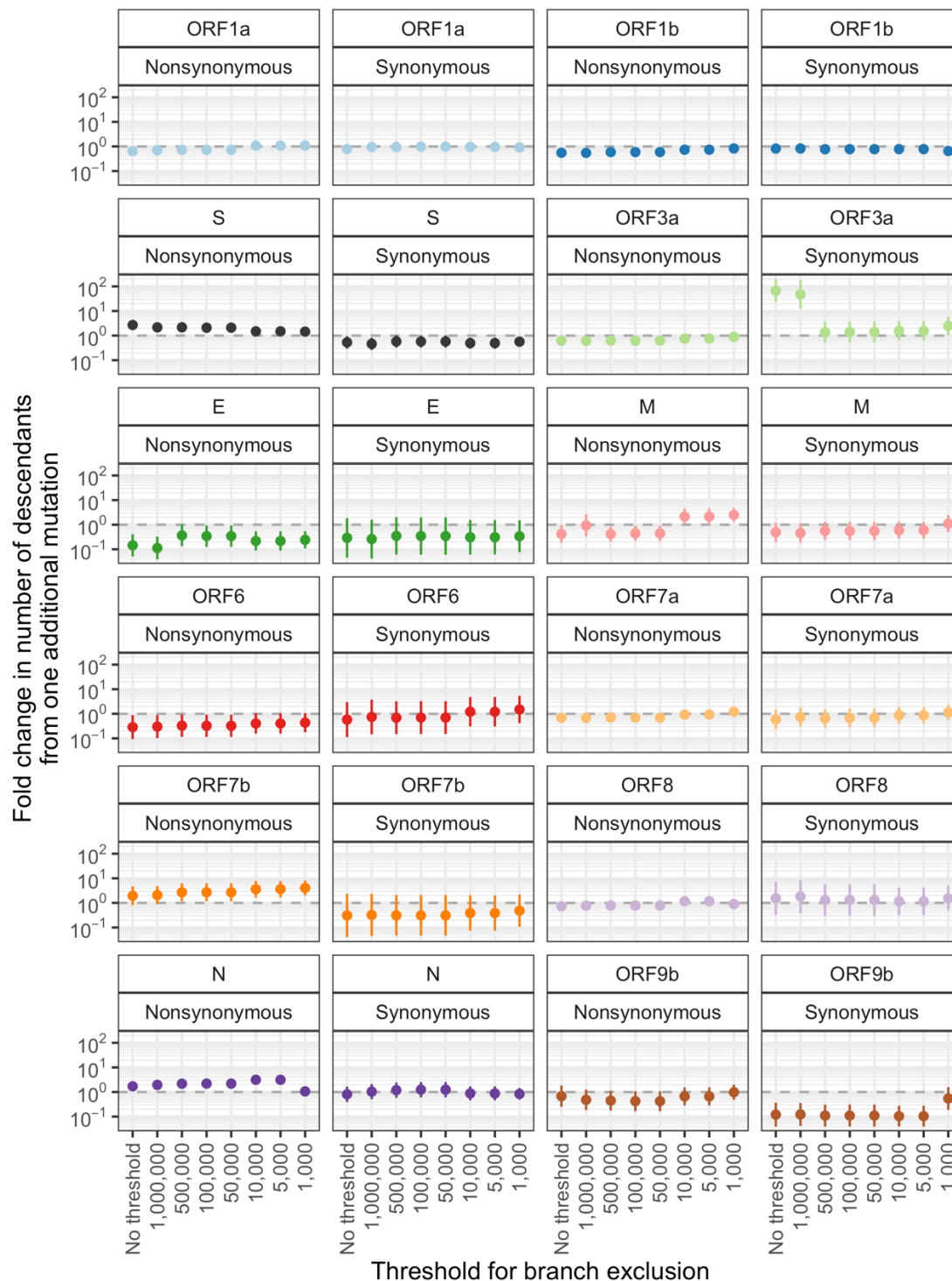
**Figure S7: Posterior predictive checks.** Comparison between the number of branches with a given number of nonsynonymous and synonymous mutations by gene across prior parametrizations. Points and vertical segments depict median and 95% credible intervals from posterior draws. For each prior parametrization (measured by the value of the coefficient of variation  $CV_{prior}$ ), we also report the Pearson correlation coefficient  $R$  between the log-transform predicted and observed number of branches, with a pseudo-count of 1 added prior to transformation (see Methods).



**Figure S8: Impact of prior parametrization on model performance.** Pearson correlation coefficient between the log-transformed predicted number of branches (median from the posterior) and the true number of branches with the same number of synonymous and nonsynonymous mutations across genes as a function of the coefficient of variation  $CV_{\text{prior}}$  used on the prior parametrization. Prior to log-transformation, we add 1 to branch counts (see Methods).



**Figure S9: Impact of accounting for mutation opportunity on the odds ratio for a mutation occurring in saltational vs non-saltational branches.**  $R$  corresponds to the Pearson correlation coefficient between the log ORs.



**Figure S10: Impact of branches with a high number of descendants on parameter estimates from a negative binomial regression.** We depict expected fold-change in the number of descendants from saltational branches from one additional mutation (by gene and mutation type) restricting the saltational branches included in the regression to those lying below the threshold depicted on the x-axis. Vertical segments indicate 95% confidence intervals.

**Table S1: List of positions that are more frequently mutated in saltational than non-saltational branches.** OR were computed using Fisher tests and p-values were adjusted for multiple testing (Benjamini-Hochberg corrections). We use a significance threshold of 0.05 for the adjusted p-value.

Gene	Position	OR of position being mutated in saltational vs non-saltational branches (95% CI)	Saltational branches count with nonsynonymous mutations at position
S	484	6.2 (4.6-8.1)	52
ORF8	121	4 (2.9-5.4)	43
S	346	4.5 (3.2-6.2)	41
ORF8	120	5.9 (4.2-8.2)	39
S	446	2.7 (1.9-3.8)	35
S	478	6.3 (4.4-8.8)	35
S	452	9.1 (6.3-12.9)	34
S	681	7.6 (5.2-10.7)	34
ORF1a	1638	11.3 (7.7-16.1)	33
S	493	14.1 (9.2-20.7)	28
S	936	3.2 (2-4.7)	25
S	501	16.2 (10.2-24.7)	24
ORF8	118	6.7 (4.2-10.1)	23
S	444	3.1 (2-4.7)	23
S	9	5 (3.1-7.8)	21
S	95	5 (3.1-7.8)	21
S	494	2.7 (1.7-4.2)	21
S	157	2.8 (1.7-4.5)	19
S	440	24.6 (14.3-39.9)	19
S	486	9.2 (5.5-14.6)	19
ORF1a	1322	7.7 (4.5-12.3)	18
S	152	3.9 (2.3-6.2)	18
S	339	8.1 (4.7-12.9)	18
ORF1a	3255	13.3 (7.6-21.8)	17
N	203	3.4 (1.9-5.5)	16
ORF1a	102	15.1 (8.5-25.2)	16
S	257	4.6 (2.6-7.6)	16
S	376	7.4 (4.2-12.2)	16
S	417	6.9 (3.9-11.4)	16
ORF1a	1795	29 (15.6-50.5)	15
S	490	3.9 (2.2-6.5)	15
M	3	3 (1.6-5)	14
ORF1a	3209	3.2 (1.8-5.5)	14
S	19	2.6 (1.4-4.3)	14
S	340	7.5 (4.1-12.8)	14
S	477	3.3 (1.8-5.5)	14
S	655	2.9 (1.6-4.9)	14
N	204	2.6 (1.4-4.5)	13
ORF1a	1323	10.2 (5.4-17.9)	13
ORF1b	662	2.7 (1.5-4.7)	13
S	210	9 (4.7-15.7)	13

S	213	4.1 (2.2-7.1)	13
S	214	2.7 (1.4-4.7)	13
S	403	6.7 (3.5-11.6)	13
N	234	3.6 (1.9-6.4)	12
ORF1a	2980	5.2 (2.7-9.2)	12
ORF1a	3201	5.2 (2.7-9.2)	12
S	356	5.5 (2.8-9.7)	12
S	252	3.5 (1.7-6.3)	11
S	445	3.2 (1.6-5.8)	11
S	450	8.2 (4-14.9)	11
ORF1a	1639	9.4 (4.4-17.8)	10
ORF8	117	11.4 (5.3-21.5)	10
S	64	4.7 (2.2-8.8)	10
S	83	3.2 (1.5-6)	10
ORF1a	1500	3.4 (1.5-6.5)	9
ORF1a	3653	10.5 (4.7-20.5)	9
ORF1a	3915	3.4 (1.5-6.5)	9
ORF1a	4097	6.3 (2.8-12.1)	9
ORF1a	4165	8.1 (3.7-15.8)	9
S	375	8.4 (3.8-16.3)	9
S	460	4.5 (2-8.7)	9
S	498	25.4 (11.1-51.3)	9
S	954	8.3 (3.7-16)	9
S	449	18.6 (7.7-38.5)	8
E	30	4.9 (2-10.3)	7
ORF1b	1000	4.9 (2-10.3)	7
ORF8	116	7.2 (2.8-15.2)	7
S	164	6.9 (2.7-14.6)	7
S	373	5.7 (2.3-12)	7
S	385	4.5 (1.8-9.5)	7
S	496	6.9 (2.7-14.6)	7
S	1118	4.2 (1.7-8.7)	7
N	235	5.7 (2.1-12.7)	6
ORF1a	1714	5 (1.8-11.1)	6
ORF1a	2158	5.4 (2-12)	6
ORF1a	2534	5.7 (2.1-12.6)	6
S	368	8.8 (3.2-19.8)	6
S	505	11.1 (4-25.1)	6
ORF1a	47	6.3 (2-15.1)	5
ORF1a	4167	12.3 (3.9-30)	5
S	500	9.3 (3-22.6)	5
S	764	17.9 (5.6-44.7)	5
S	981	11.6 (3.7-28.3)	5
M	19	8.3 (2.2-22)	4
ORF1a	3651	57.2 (13.4-189)	4
ORF1b	799	36.2 (8.9-108.6)	4
ORF1b	800	21.5 (5.5-60.5)	4
S	136	12 (3.2-32.6)	4
S	568	9.8 (2.6-26.3)	4

S	1190	12.5 (3.3-33.9)	4
N	42	102.9 (16-534.7)	3
ORF1a	2390	13.5 (2.7-42.7)	3
ORF1b	804	28.6 (5.4-98)	3
S	332	17.8 (3.5-57.3)	3
S	1002	25.8 (4.9-86.9)	3
ORF1a	2426	42.9 (4.4-216.4)	2
ORF1a	3216	34.3 (3.7-160.8)	2

**Table S2: List of nonsynonymous AA mutations that are more frequent in saltational than non-saltational branches.** OR were computed using Fisher tests and p-values were adjusted for multiple testing (Benjamini-Hochberg corrections). We use a significance threshold of 0.05 for the adjusted p-value. This table only contains mutations occurring at least 5 times on saltational branches.

Gene	Mutation	OR of mutation occurring in saltational vs non-saltational branches (95% CI)	Saltational branches count with mutation
ORF8	121L	4.77 (3.44-6.46)	43
ORF1a	3829F	1.76 (1.18-2.54)	29
ORF1a	1638I	13.27 (8.54-19.8)	26
ORF8	120L	8.85 (5.61-13.34)	24
S	484K	15.44 (9.72-23.47)	24
S	493Q	43.29 (25.48-69.91)	20
S	9L	9.8 (5.91-15.38)	20
S	494P	3.6 (2.15-5.65)	19
S	452R	13.24 (7.67-21.52)	18
S	95I	3.41 (2-5.45)	18
ORF8	118V	5.8 (3.35-9.38)	17
ORF1a	4398L	2.4 (1.39-3.86)	17
S	501Y	35.35 (19.57-60.44)	17
ORF1a	3255I	12.24 (6.79-20.66)	16
S	681H	9.69 (5.45-16.04)	16
ORF8	120V	4.61 (2.56-7.67)	15
S	346T	4.49 (2.5-7.46)	15
ORF1a	1795Q	114.91 (54.05-236.62)	14
S	446G	13.12 (7.01-22.67)	14
S	936H	5.96 (3.23-10.13)	14
ORF1a	1322I	7.13 (3.75-12.38)	13
ORF1a	102K	35.7 (17.45-67.11)	12
S	157L	3.55 (1.82-6.25)	12
ORF1a	3209V	3.38 (1.74-5.95)	12
S	440R	86.22 (39.19-179.41)	12
S	655Y	2.82 (1.45-4.97)	12
ORF1b	662S	3.49 (1.79-6.15)	12
S	478R	4.06 (2.01-7.33)	11
S	547K	4.9 (2.42-8.88)	11
S	152R	6.21 (2.94-11.59)	10
ORF1a	1639N	12.65 (5.93-24)	10

S	450D	9 (4.25-16.93)	10
S	477N	11.4 (5.35-21.58)	10
S	478K	15.38 (7.02-30.28)	10
S	484E	30.11 (13.68-59.64)	10
S	681R	11.93 (5.58-22.72)	10
M	82T	12.04 (5.63-22.91)	10
N	234I	3.71 (1.68-7.11)	9
S	346R	7.25 (3.26-14.08)	9
S	356T	10.2 (4.58-19.89)	9
S	376T	15.8 (6.99-31.44)	9
S	257S	6.55 (2.79-13.14)	8
ORF1a	3653F	28.1 (11.49-59.8)	8
S	403K	21.46 (8.89-44.91)	8
S	446S	3.54 (1.52-7.05)	8
S	460K	12.04 (5.08-24.47)	8
S	486F	16.04 (6.64-33.54)	8
ORF1a	1323G	14.15 (5.52-30.44)	7
ORF1a	1500Y	3.36 (1.34-6.99)	7
N	203K	13.28 (5.19-28.48)	7
S	213E	8.39 (3.32-17.73)	7
E	30I	5.9 (2.34-12.37)	7
S	339G	20.78 (8.02-45.54)	7
S	452M	7.63 (3.02-16.09)	7
S	478T	15.89 (6.21-34.1)	7
S	481K	11.42 (4.48-24.37)	7
S	490S	5.07 (2.02-10.59)	7
S	83V	19.48 (7.27-44.77)	7
ORF9b	10P	9.63 (3.46-21.64)	6
N	13P	9.08 (3.27-20.38)	6
S	153T	6.95 (2.51-15.5)	6
ORF1a	1638N	9.37 (3.36-21.06)	6
S	190S	7.64 (2.75-17.06)	6
N	204G	7.47 (2.69-16.7)	6
S	210V	12.21 (4.36-27.69)	6
S	245N	15.77 (5.59-36.21)	6
S	257D	4.21 (1.53-9.29)	6
S	346K	3.82 (1.39-8.43)	6
ORF1a	3571V	3.9 (1.42-8.6)	6
S	368I	14.55 (5.17-33.2)	6
S	375S	12.2 (4.36-27.67)	6
ORF1a	4097K	8.59 (3.09-19.25)	6
S	417T	8.29 (2.98-18.56)	6
S	449H	33.33 (11.37-81.09)	6
S	498R	38.27 (12.88-95.27)	6
S	70F	3.8 (1.38-8.39)	6
S	859N	13.06 (4.66-29.71)	6
S	954Q	72.5 (23.04-197.03)	6
ORF1b	1000L	7.45 (2.37-18)	5
ORF1a	1074T	4.5 (1.44-10.68)	5

ORF1b	1315R	21.61 (6.62-55.33)	5
S	158S	19.58 (6.06-49.29)	5
ORF1a	165Y	4.48 (1.44-10.63)	5
S	186L	5 (1.6-11.9)	5
ORF9b	18H	4.97 (1.59-11.82)	5
ORF1a	2158I	5.72 (1.83-13.66)	5
S	252G	4.82 (1.53-11.64)	5
ORF1a	2980N	6.4 (2.04-15.32)	5
ORF1a	3164H	6.13 (1.96-14.66)	5
S	340D	11.78 (3.71-28.78)	5
S	385I	5.24 (1.68-12.49)	5
ORF7a	3F	8.18 (2.6-19.7)	5
ORF1a	4165A	31.8 (9.56-83.69)	5
S	417I	34.45 (10.29-91.68)	5
S	417N	4.75 (1.52-11.31)	5
S	439K	4.58 (1.47-10.89)	5
S	440K	41.53 (12.19-113.94)	5
S	452Q	5.71 (1.83-13.64)	5
S	478E	7.75 (2.47-18.64)	5
ORF7a	49P	8.09 (2.57-19.49)	5
ORF7a	4L	11.01 (3.48-26.82)	5
S	501T	17.95 (5.58-44.91)	5
ORF1a	542H	4.64 (1.49-11.03)	5
ORF7a	5V	11.92 (3.76-29.14)	5
S	64R	5.65 (1.81-13.48)	5
S	67A	11.37 (3.62-27.32)	5
S	764K	118.11 (29.6-436.78)	5
S	899S	4.75 (1.52-11.29)	5

**Table S3: Summary of recurrent amino-acid mutations found in persistent SARS-CoV-2 infections.** Summary of recurrent nonsynonymous mutations reported in chronic SARS-CoV-2 infections<sup>19–21,48</sup> or suspected chronic SARS-CoV-2 infections<sup>27</sup>. We report mutations observed in at least 3 infections except in Harari et al.<sup>48</sup> (Nature Medicine) and Ghafari et al.<sup>19</sup> (Nature) where we report mutations observed in at least 2 persistent infections (due to the lower sample size and number of mutations satisfying this criterion).

<b>Mutation</b>	<b>Source</b>
S:E340K	Khurana et al. medRxiv. 2025. <sup>21</sup> Harari et al. Nature Communications. 2024. <sup>27</sup>
S:N405D	Khurana et al. medRxiv. 2025. <sup>21</sup>
ORF1a:L3829F	Khurana et al. medRxiv. 2025. <sup>21</sup>
ORF1b:G662S	Khurana et al. medRxiv. 2025. <sup>21</sup>
ORF1a:T1638I	Khurana et al. medRxiv. 2025. <sup>21</sup> Wilkinson et al. Virus Evolution. 2022. <sup>20</sup> Ghafari et al. Nature. 2024. <sup>19</sup> Harari et al. Nature Medicine. 2022. <sup>48</sup> Harari et al. Nature Communications. 2024. <sup>27</sup>
S:R346T	Khurana et al. medRxiv. 2025. <sup>21</sup> Harari et al. Nature Communications. 2024. <sup>27</sup>
S:K444R	Khurana et al. medRxiv. 2025. <sup>21</sup>
N:R93R	Khurana et al. medRxiv. 2025. <sup>21</sup>
E:T30I	Khurana et al. medRxiv. 2025. <sup>21</sup> Wilkinson et al. Virus Evolution. 2022. <sup>20</sup> Harari et al. Nature Medicine. 2022. <sup>48</sup> Harari et al. Nature Communications. 2024. <sup>27</sup>
S:N417I	Khurana et al. medRxiv. 2025. <sup>21</sup>
ORF1a:D1639N	Khurana et al. medRxiv. 2025. <sup>21</sup>
S:L452M	Khurana et al. medRxiv. 2025. <sup>21</sup>
ORF1b:V783I	Khurana et al. medRxiv. 2025. <sup>21</sup>
S:R346I	Khurana et al. medRxiv. 2025. <sup>21</sup>
S:D405N	Khurana et al. medRxiv. 2025. <sup>21</sup>
S:K764N	Khurana et al. medRxiv. 2025. <sup>21</sup>
ORF1a:T4175I	Khurana et al. medRxiv. 2025. <sup>21</sup> Ghafari et al. Nature. 2024. <sup>19</sup>
S:T19I	Khurana et al. medRxiv. 2025. <sup>21</sup>
ORF3a:S74S	Khurana et al. medRxiv. 2025. <sup>21</sup>
N:P326L	Khurana et al. medRxiv. 2025. <sup>21</sup>
ORF1a:T1322I	Khurana et al. medRxiv. 2025. <sup>21</sup> Harari et al. Nature Medicine. 2022. <sup>48</sup>
S:E340Q	Khurana et al. medRxiv. 2025. <sup>21</sup> Harari et al. Nature Communications. 2024. <sup>27</sup>
S:K444N	Khurana et al. medRxiv. 2025. <sup>21</sup> Harari et al. Nature Communications. 2024. <sup>27</sup>
S:I19T	Khurana et al. medRxiv. 2025. <sup>21</sup>
S:S494P	Khurana et al. medRxiv. 2025. <sup>21</sup> Harari et al. Nature Medicine. 2022. <sup>48</sup>
S:K440R	Khurana et al. medRxiv. 2025. <sup>21</sup>
ORF1a:S4398L	Khurana et al. medRxiv. 2025. <sup>21</sup>
S:P9L	Khurana et al. medRxiv. 2025. <sup>21</sup>

	Harari et al. Nature Medicine. 2022. <sup>48</sup>
S:T95I	Khurana et al. medRxiv. 2025. <sup>21</sup> Wilkinson et al. Virus Evolution. 2022. <sup>20</sup> Harari et al. Nature Communications. 2024. <sup>27</sup>
ORF1a:T1538I	Khurana et al. medRxiv. 2025. <sup>21</sup>
ORF1a:A2529V	Khurana et al. medRxiv. 2025. <sup>21</sup>
ORF1a:E102K	Khurana et al. medRxiv. 2025. <sup>21</sup>
ORF1a:G3072C	Khurana et al. medRxiv. 2025. <sup>21</sup>
S:E484K	Wilkinson et al. Virus Evolution. 2022. <sup>20</sup> Harari et al. Nature Medicine. 2022. <sup>48</sup>
M:H125Y	Wilkinson et al. Virus Evolution. 2022. <sup>20</sup> Harari et al. Nature Medicine. 2022. <sup>48</sup>
ORF1a:T3058I	Wilkinson et al. Virus Evolution. 2022. <sup>20</sup> Harari et al. Nature Medicine. 2022. <sup>48</sup>
S:Q493K	Wilkinson et al. Virus Evolution. 2022. <sup>20</sup>
S:S13I	Wilkinson et al. Virus Evolution. 2022. <sup>20</sup> Harari et al. Nature Medicine. 2022. <sup>48</sup>
ORF1b:V783I	Wilkinson et al. Virus Evolution. 2022. <sup>20</sup> Harari et al. Nature Medicine. 2022. <sup>48</sup>
ORF1a:P1640L	Wilkinson et al. Virus Evolution. 2022. <sup>20</sup> Harari et al. Nature Medicine. 2022. <sup>48</sup>
ORF1a:T1322P	Wilkinson et al. Virus Evolution. 2022. <sup>20</sup>
S:L452R	Wilkinson et al. Virus Evolution. 2022. <sup>20</sup> Ghafari et al. Nature. 2024. <sup>19</sup>
S:Q498R	Wilkinson et al. Virus Evolution. 2022. <sup>20</sup>
S:E484G	Wilkinson et al. Virus Evolution. 2022. <sup>20</sup>
S:F486I	Wilkinson et al. Virus Evolution. 2022. <sup>20</sup>
S:G142V	Wilkinson et al. Virus Evolution. 2022. <sup>20</sup>
ORF1a:K1795Q	Wilkinson et al. Virus Evolution. 2022. <sup>20</sup> Harari et al. Nature Communications. 2024. <sup>27</sup>
S:F490L	Wilkinson et al. Virus Evolution. 2022. <sup>20</sup>
ORF1a:T4175I	Ghafari et al. Nature. 2024. <sup>19</sup>
ORF1b:L1504F	Ghafari et al. Nature. 2024. <sup>19</sup>
S:T376A	Ghafari et al. Nature. 2024. <sup>19</sup>
S:T547K	Ghafari et al. Nature. 2024. <sup>19</sup>
S:R190K	Harari et al. Nature Medicine. 2022. <sup>48</sup>
S:N501Y	Harari et al. Nature Medicine. 2022. <sup>48</sup>
S:E484Q	Harari et al. Nature Medicine. 2022. <sup>48</sup>
ORF7a:A105V	Harari et al. Nature Medicine. 2022. <sup>48</sup>
ORF8:T11I	Harari et al. Nature Medicine. 2022. <sup>48</sup>
ORF1a:V1887I	Harari et al. Nature Communications. 2024. <sup>27</sup>
S:S371F	Harari et al. Nature Communications. 2024. <sup>27</sup>
S:A701V	Harari et al. Nature Communications. 2024. <sup>27</sup>
S:K356T	Harari et al. Nature Communications. 2024. <sup>27</sup>
S:E340D	Harari et al. Nature Communications. 2024. <sup>27</sup>
S:L455S	Harari et al. Nature Communications. 2024. <sup>27</sup>
S:N460K	Harari et al. Nature Communications. 2024. <sup>27</sup>
ORF1a:P1803S	Harari et al. Nature Communications. 2024. <sup>27</sup>
ORF3a:W131C	Harari et al. Nature Communications. 2024. <sup>27</sup>

S:E340A	Harari et al. Nature Communications. 2024. <sup>27</sup>
S:G446V	Harari et al. Nature Communications. 2024. <sup>27</sup>
S:P337S	Harari et al. Nature Communications. 2024. <sup>27</sup>
ORF3a :E239Q	Harari et al. Nature Communications. 2024. <sup>27</sup>

**Table S4: Summary of recurrent amino-acid mutations reported in minks infected with SARS-CoV-2.** We report the 5 mutations found to be strong candidate for mink-specific adaptation by Tan et al.<sup>50</sup>, 3 mutations found to recur in minks in Zhou et al.<sup>49</sup> For Iglesias-Caballero et al.<sup>51</sup>, we extracted recurrent nonsynonymous mutations across clusters from their second figure.

Mutation	Source
S:Y453F	Zhou et al. Cell Reports. 2022. <sup>49</sup> Tan et al. Nature Communications. 2025. <sup>50</sup>
S:F486L	Zhou et al. Cell Reports. 2022. <sup>49</sup> Tan et al. Nature Communications. 2025. <sup>50</sup> Iglesias-Caballero et al. Int. J. Nol. Sci. 2024. <sup>51</sup>
S:N501T	Zhou et al. Cell Reports. 2022. <sup>49</sup> Tan et al. Nature Communications. 2025. <sup>50</sup> Iglesias-Caballero et al. Int. J. Nol. Sci. 2024. <sup>51</sup>
ORF1a:G1477E	Tan et al. Nature Communications. 2025. <sup>50</sup> Iglesias-Caballero et al. Int. J. Nol. Sci. 2024. <sup>51</sup>
ORF3a:L219V	Tan et al. Nature Communications. 2025. <sup>50</sup>
ORF1a:T265I	Iglesias-Caballero et al. Int. J. Nol. Sci. 2024. <sup>51</sup>

**Table S5: Summary of recurrent amino-acid mutations reported in deers infected with SARS-CoV-2.** We report a mutation found to be a strong candidate for deer-specific adaptation by Tan et al.<sup>4</sup>, 2 mutations found to occur at least 3 times independently in Marques et al.<sup>52</sup> and mutations found in 3 independent samples in Feng et al.<sup>53</sup>

Mutation	Source
ORF1a:L1853F	Tan et al. Nature Communications. 2025. <sup>50</sup>
S:V252G	Marques et al. PLoS Pathogens. 2025. <sup>52</sup>
ORF1a:S3149F	Marques et al. PLoS Pathogens. 2025. <sup>52</sup>
ORF1a:H110Y	Feng et al. Nature Communications. 2023. <sup>53</sup>
ORF1a:S443P	Feng et al. Nature Communications. 2023. <sup>53</sup>
ORF1a:S944L	Feng et al. Nature Communications. 2023. <sup>53</sup>
ORF1a:T1000I	Feng et al. Nature Communications. 2023. <sup>53</sup>
ORF1a:P1640S	Feng et al. Nature Communications. 2023. <sup>53</sup>
ORF1a:T1854I	Feng et al. Nature Communications. 2023. <sup>53</sup>
ORF1a:T1881I	Feng et al. Nature Communications. 2023. <sup>53</sup>
ORF1a:T2016I	Feng et al. Nature Communications. 2023. <sup>53</sup>
ORF1a:S2103F	Feng et al. Nature Communications. 2023. <sup>53</sup>
ORF1a:S2224F	Feng et al. Nature Communications. 2023. <sup>53</sup>
ORF1a:S2242F	Feng et al. Nature Communications. 2023. <sup>53</sup>
ORF1a:S2255F	Feng et al. Nature Communications. 2023. <sup>53</sup>
ORF1a:S2273F	Feng et al. Nature Communications. 2023. <sup>53</sup>
ORF1a:T2283I	Feng et al. Nature Communications. 2023. <sup>53</sup>
ORF1a:H2357Y	Feng et al. Nature Communications. 2023. <sup>53</sup>
ORF1a:K2511N	Feng et al. Nature Communications. 2023. <sup>53</sup>
ORF1a:V3017I	Feng et al. Nature Communications. 2023. <sup>53</sup>
ORF1a:H3076Y	Feng et al. Nature Communications. 2023. <sup>53</sup>

ORF1a:T3082I	Feng et al. Nature Communications. 2023. <sup>53</sup>
ORF1a:S3099L	Feng et al. Nature Communications. 2023. <sup>53</sup>
ORF1a:L3116F	Feng et al. Nature Communications. 2023. <sup>53</sup>
ORF1a:S3149F	Feng et al. Nature Communications. 2023. <sup>53</sup>
ORF1a:T3150I	Feng et al. Nature Communications. 2023. <sup>53</sup>
ORF1a:A3548V	Feng et al. Nature Communications. 2023. <sup>53</sup>
ORF1a:T3646I	Feng et al. Nature Communications. 2023. <sup>53</sup>
ORF1a:L3808F	Feng et al. Nature Communications. 2023. <sup>53</sup>
ORF1a:T3904I	Feng et al. Nature Communications. 2023. <sup>53</sup>
ORF1a:L4111F	Feng et al. Nature Communications. 2023. <sup>53</sup>
ORF1b:A88V	Feng et al. Nature Communications. 2023. <sup>53</sup>
ORF1b:S2338L	Feng et al. Nature Communications. 2023. <sup>53</sup>
ORF1b:S2339F	Feng et al. Nature Communications. 2023. <sup>53</sup>
ORF1b:T2376I	Feng et al. Nature Communications. 2023. <sup>53</sup>
S:T22I	Feng et al. Nature Communications. 2023. <sup>53</sup>
S:T29I	Feng et al. Nature Communications. 2023. <sup>53</sup>
S:H146Y	Feng et al. Nature Communications. 2023. <sup>53</sup>
S:H245Y	Feng et al. Nature Communications. 2023. <sup>53</sup>
S:S640F	Feng et al. Nature Communications. 2023. <sup>53</sup>
S:S680F	Feng et al. Nature Communications. 2023. <sup>53</sup>
S:L1203F	Feng et al. Nature Communications. 2023. <sup>53</sup>
ORF3a:A54V	Feng et al. Nature Communications. 2023. <sup>53</sup>
ORF3a: L108F	Feng et al. Nature Communications. 2023. <sup>53</sup>
ORF7a:T39I	Feng et al. Nature Communications. 2023. <sup>53</sup>
ORF7b:S5L	Feng et al. Nature Communications. 2023. <sup>53</sup>
ORF8:S43F	Feng et al. Nature Communications. 2023. <sup>53</sup>
ORF8:K68E	Feng et al. Nature Communications. 2023. <sup>53</sup>
N:T271I	Feng et al. Nature Communications. 2023. <sup>53</sup>
N:T334I	Feng et al. Nature Communications. 2023. <sup>53</sup>