medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Topological data analysis identifies emerging adaptive mutations in SARS-CoV-2

Michael Bleher[2†*], Lukas Hahn[2†*], Maximilian Neumann[1], Juan Ángel Patiño-Galindo [4], Mathieu Carrière[5], Ulrich Bauer[6], Raúl Rabadán[3*], Andreas Ott[1,2†*]

[1]Mathematics Department, Karlsruhe Institute of Technology, Karlsruhe, Germany
[2]Institute for Mathematics, Heidelberg University, Heidelberg, Germany
[3]Program for Mathematical Genomics, Department of Systems Biology, Columbia University, New York, NY, USA
[4]Department of Microbiology, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[5]DataShape, Centre Inria d'Université Côte d'Azur, Biot, France
[6]TUM Department of Mathematics and Munich Data Science Institute, Munich, Germany

[†]These authors contributed equally to this work.
[*]Corresponding authors:
mbleher@mathi.uni-heidelberg.de (M.B.)
lhahn@mathi.uni-heidelberg.de (L.H.)
rr2579@cumc.columbia.edu (R.R.)
andreas.ott@kit.edu (A.O.)

Abstract

The COVID-19 pandemic has initiated an unprecedented worldwide effort to characterize its evolution through the mapping of mutations of the coronavirus SARS-CoV-2. The early identification of mutations that could confer adaptive advantages to the virus, such as higher infectivity or immune evasion, is of paramount importance. However, the large number of currently available genomes precludes the efficient use of phylogeny-based methods. Here we present CoVtRec, a fast and scalable Topological Data Analysis approach for the surveillance of emerging adaptive mutations in large genomic datasets. Our method overcomes limitations of state-of-the-art phylogeny-based approaches by quantifying the potential adaptiveness of mutations merely by their topological footprint in the genome alignment, without resorting to the reconstruction of a single optimal phylogenetic tree. Analyzing millions of SARS-CoV-2 genomes from GISAID, we find a correlation between topological signals and adaptation to the human host. By leveraging the stratification by time in sequence data, our method enables the high-resolution longitudinal analysis of topological signals of adaptation. We characterize the convergent evolution of the coronavirus throughout the whole pandemic to date, report on emerging potentially adaptive mutations, and pinpoint mutations in Variants of Concern that are likely associated with positive selection. Our approach can improve the surveillance of mutations of concern and guide experimental studies.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Introduction

The COVID-19 pandemic, caused by the coronavirus SARS-CoV-2, has led to millions of lost human lives and devastating economic impact worldwide. As the virus continues to spread through the world, it is acquiring new mutations in its genome, and although most mutations will be deleterious or neutral, a few of them could be advantageous for the virus, for instance, by increasing its infectivity or by helping it to avoid the immune system. As more people develop immune protection by previous viral infections or through vaccination, it is important to rapidly and effectively identify mutations that could confer the virus some fitness advantage [1, 2].

One approach to identify adaptive mutations consists of experimentally mutating many positions and testing the effect on certain phenotypes, like binding to the human receptor or immune evasion [3–6]. However, experimental approaches are limited by the vast number of possible variations, especially when different genetic backgrounds are considered [7].

A more data-driven approach, which solely relies on the genomic information of the virus, is to look for mutations in a particular genomic locus that occur multiple times. In this approach, one usually reconstructs an estimated phylogenetic tree and identifies mutations that occur in independent branches [8–16]. If a mutation gives some sort of advantage to the virus, we expect it to occur in several places in the phylogeny independently and its prevalence to increase with time. For instance, the `D614G` mutation in the Spike gene was identified early in the pandemic and is now found in virtually all virus isolates [17].

In the COVID-19 pandemic, an unprecedented worldwide effort to sequence viruses resulted in a growing number of millions of genomes available to the scientific community [18]. Ideally, one would like to leverage all this genomic information at real-time to rapidly report the emergence of potential mutations of concern [1, 2]. Several phylogeny-based tools and methods are available [7, 19–31]. Phylogenetic approaches, however, become daunting as the number of sequences increases, and are computationally challenging when the number of genomes exceeds the tens of thousands [13, 19, 32]. In addition, homoplasies caused by the independent reemergence of mutations, as well as recombination events, confuse the generation of phylogenetic trees [33, 34]. It has moreover been observed that certain methods for constructing a single optimal phylogeny for SARS-CoV-2 can be problematic as the number of isolates is large while the genetic diversity is low [32].

Here we establish a novel method, based on Topological Data Analysis (TDA) and implemented in the CoVtRec pipeline (see `https://tdalife.github.io/covtrec`), that can efficiently identify emerging potentially adaptive mutations without the need to choose a single optimal phylogenetic tree in a vast range of equally plausible tree reconstructions. It systematically detects convergent events in viral evolution merely by their topological footprint in the genome alignment, thus overcoming limitations of current phylogenetic inference techniques. Thanks to our use of highly optimized algorithms, it easily scales to hundreds of thousands of distinct genomes. Our approach has the ability to map the dynamics of topological signals of adaptation at daily resolution over any period of time, which enables not only the real-time monitoring but also the retrospective assessment of adaptive processes in the evolution of the coronavirus. We evaluate how our new method compares, both in performance and in results, with state-of-the-art phylogeny-based methods for the construction and analysis of phylogenetic trees, such as UShER [29], IQTree [35] and HomoplasyFinder [9], and for the assessment of fitness effects of mutations as developed by Bloom & Neher [7].

We perform CoVtRec analyses of large alignments with millions of genomic sequences shared via GISAID, the global data science initiative [18, 36], in three different phases of the pandemic: (i) in February 2021, retrospectively monitoring the convergent evolution of the virus from the

medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

beginning of the pandemic in a largely immunologically naive population; (ii) in January 2022, in the presence of increasing host immunity and dominant variants, with the Omicron variant on the rise; and (iii) in March 2023, in order to obtain a comprehensive picture of the dynamics of adaptation over the whole course of the pandemic. Invoking published experimental and epidemiological data into our analysis, we investigate the correlation between topological signals and host adaptation. As we show, our method can identify adaptive mutations at an early stage, often several months before they become recognizable by their prevalence in the population. We characterize potentially adaptive mutations in different phases of the pandemic, and identify mutations in Variants of Concern that are likely due to convergent evolution.

Results

Quantification of topological recurrence

Chan *et al.* [37] initiated the use of *persistent homology*, a method from Topological Data Analysis, to extract global evolutionary features from large genomic datasets. This method detects topological cycles in the dataset, which correspond to reticulate events in the phylogeny and may cause phylogenetic inference methods to produce ambiguous tree topologies. All information is compiled into a stable and unbiased descriptor known as a *persistence barcode* (see Figure 1, Figure 2 and Methods).

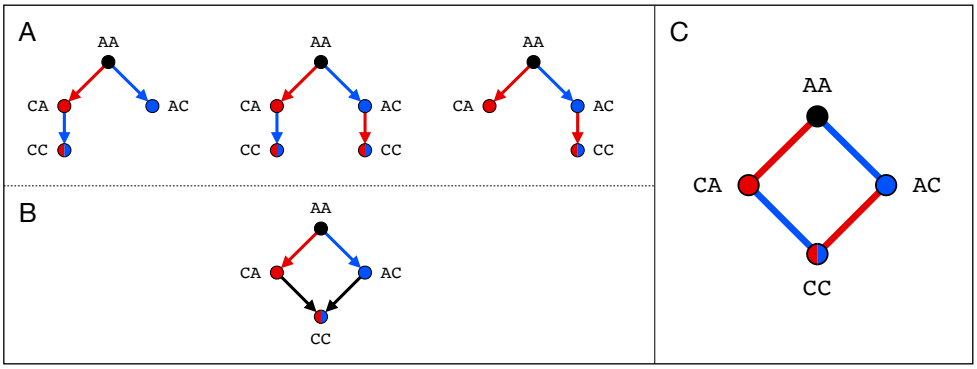


Figure 1. Reticulate events in molecular evolution create topological cycles. (A) and (B) show possible evolutionary histories on the level of individuals in the example of a genome with only two nucleotides. The coloring of the edges corresponds to the acquisition of a specific mutation, while the coloring of the nodes represents individuals carrying this mutation. Convergent evolution (A) or recombination (B) leads to the presence of four alleles for which there is no single consistent phylogeny (four-gamete test). On the genomic level, genetically identical individuals cannot be distinguished and incompatible phylogenies are represented by a topological cycle in the corresponding phylogenetic network (C).

There are several scenarios that can lead to reticulate events. For instance, if a genome of an organism imports genetic material from a different genome, in lateral gene transfer for instance, we will observe that parts of the newly generated genome resemble the parent, while others resemble the genome of the organism that exported that material. Recombination and reassortments are common phenomena observed in viruses where two parental strains co-infect the same host cell generating a new virus containing genetic material from both parental strains. But similarity between genomes can also be generated at smaller scales, when the same mutation occurs independently twice, making the two strains more similar than expected. Persistent homology captures all these events, and also the scale of the events. Although in general it requires care to infer the biological origin of a given topological cycle, in viral evolution

medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

one expects bars at small scales to correspond mostly to homoplasies, while well-supported recombination events typically produce topological features at larger scales, as entire blocks of genetic material are exchanged in the process [37–39].

Here we define a novel index of recurrence that is based on persistent homology and does not rely on a possibly ambiguous tree reconstruction. We use a specifically designed algorithm, implemented in `Ripser` [40], that associates to relevant bars in the persistence barcode explicit *SNV cycles*, given by a series of isolates that approximates all evolutionary steps as faithfully as possible in terms of single nucleotide variations (SNV). We define the *topological recurrence index (tRI)* of a specific mutation as the total number of SNV cycles in the reticulate phylogeny that contain an edge corresponding to this mutation. This index provides a lower bound for the number of independent occurrences of the mutation in the phylogeny and is therefore a measure for convergent evolution (see Figure 2 and Methods).

Our method enables the temporal mapping of topological signals of recurrence at daily or monthly resolution by leveraging the stratification by time in genomic datasets given by the collection dates of genomic isolates (see Supplementary Information Figure 13, Figure 5, Figure 8 and Methods).

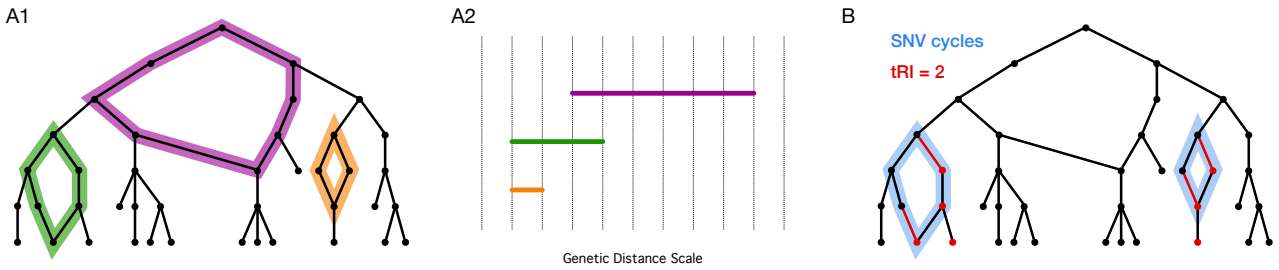


Figure 2. Topological data analysis quantifies convergent evolution. (A) Persistent homology detects reticulate events in viral evolution by means of a persistence barcode. Each bar in the barcode (A2) corresponds to a topological cycle in the reticulate phylogeny (A1). Bars at small genetic distance scales are expected to correspond mainly to homoplasies, while recombination events typically produce topological features at larger scales. (B) SNV cycles are topological cycles in the reticulate phylogeny for which adjacent sequences differ by single nucleotide variations (SNV) only. Under the assumption of single substitutions per site, any SNV in each such cycle appears twice, independently of each other, and distributed across both possible lineages. The topological recurrence index (tRI) of a specific mutation is the total number of SNV cycles in the reticulate phylogeny in which this mutation is acquired. In the displayed example phylogeny, the red edges indicate the acquisition of a specific mutation, while the red nodes represent viruses carrying this mutation. The mutation is acquired in two SNV cycles (shaded in blue) and therefore has a tRI of 2.

Comparison with phylogeny-based methods

Standard phylogenetic methods for the detection of convergent evolution are based on the reconstruction of a phylogenetic tree and therefore have an unfavorable scaling, due to the rapid growth of the number of trees representing evolutionary histories that are compatible with the observations [32]. Persistent homology provides a new approach by measuring convergent evolution purely in terms of topological cycles, without the need to construct any phylogenetic trees. It enables a rapid and scalable analysis of large alignments with hundreds of thousands of distinct sequences. A performance analysis showed that when applied to sample alignments of up to 5,000 SARS-CoV-2 sequences, the topological method was faster than state-of-the-art phylogeny-based methods such as UShER and HomoplasyFinder [9, 29, 35] (see Figure 3).

medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Moreover, the topological recurrence analysis of current alignments with up to 13 million SARS-CoV-2 sequences was accomplished in less than a day (see Methods).

The UShER phylogenetic tree [42] has recently been used to estimate fitness effects of mutations in the evolution of SARS-CoV-2 by means of a "fitness index" [7]. We compared results from this fitness analysis with results from our topological recurrence analysis based on GISAID data as of March 2023. For Spike gene amino acid changes, we found a strong correlation between positive topological recurrence index (tRI) and positive "fitness index" (Fisher's exact test, $p < 10\text{e}{-10}$) with a Spearman correlation coefficient of 0.63 ($p < 10\text{e}{-99}$) (see Supplementary Information Figure 12).

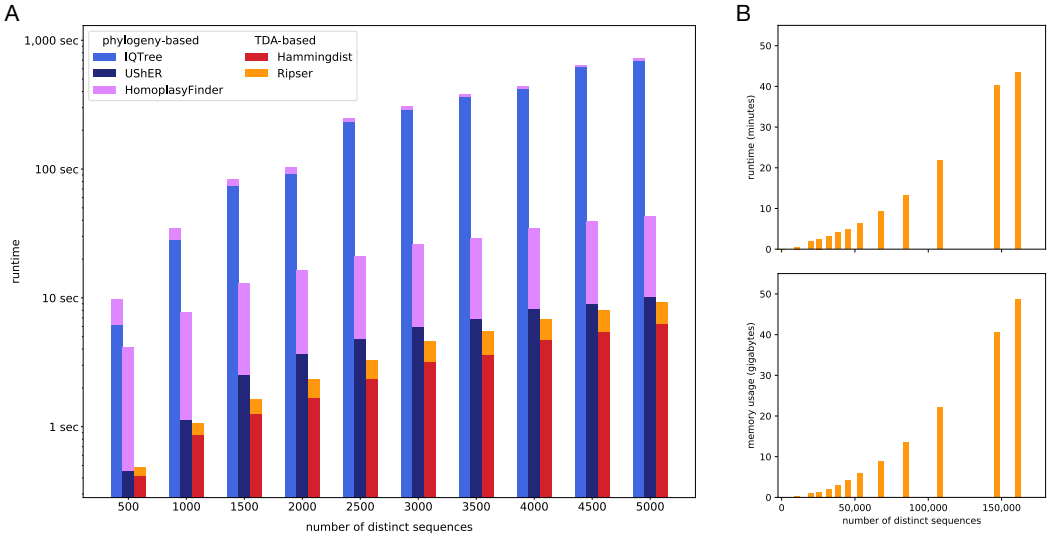


Figure 3. Performance analysis and comparison with phylogeny-based methods. (A) Basic runtime comparison between Topological Data Analysis (TDA)-based methods and phylogeny-based methods for random samples of up to 5,000 SARS-CoV-2 genomes. We used `IQTree` [35] and `UShER` [29] to reconstruct phylogenetic trees. The subsequent homoplasy analysis was performed with `HomoplasyFinder` [9]. For the TDA-approach we used `Hammingdist` [41] to generate genetic distance matrices and `Ripser` [40] for the subsequent computation of persistence barcodes (see Methods). (B) Runtime and memory usage for the computation of persistence barcodes with `Ripser` [40] for monthly sub-alignments of the GISAID alignment with 161,024 genetically distinct whole genome sequences covering the first year of the pandemic (see Methods).

Topological analysis of the first year of the pandemic

We analyzed topological signals for convergent evolution of the coronavirus SARS-CoV-2, both across the whole genome and on the Spike gene, during the first year of the pandemic from its beginning in December 2019 until February 2021. To that end, we performed a topological recurrence analysis for a curated alignment of 303,651 high-quality SARS-CoV-2 whole genome sequences from GISAID [18, 36] (see Methods). The resulting persistence barcode features 2,899 bars, 58% of which (corresponding to 1,684 SNV cycles) concentrate at small genetic distance scales ≤ 2 and are therefore expected to be associated mainly with homoplasic events (see Supplementary Information Figure 1). We did not observe bars at genetic distances that are associated with well-supported recombination events, which is compatible with previous reports that recombination did not play a substantial role in the evolution during the first year of the pandemic [30, 43].

The topological recurrence index (tRI) is designed in such a way that it naturally excludes recombination events. It is defined in terms of SNV cycles built exclusively from single nu-

medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

cleotide variations (see Figure 2 and Methods), while recombination events in SARS-CoV-2 typically involve the exchange of larger parts of the genome. This has also been documented for later phases of the pandemic, and in particular for circulating recombinant forms during the Omicron waves [44, 45]. It is therefore unlikely that recombination events significantly contribute to tRI signals in SARS-CoV-2 evolution.

In large genomic datasets, topological cycles in persistent homology may arise randomly, causing a certain amount of noise in the topological recurrence analysis. However, from simulations we inferred that at least 60% of the topological cycles found in the GISAID dataset must be real features due to increased mutation probabilities, selection, recombination, or sequencing errors (see Supplementary Information Figure 9 and Methods). We implemented several filtration rules to make sure that sequencing errors have no significant effect on results of the topological recurrence analysis (see Methods).

We detected 401 non-synonymous and 299 synonymous mutations across the whole genome showing significant tRI signal (see Figure 4 and Supplementary Information Table 1). Here signals with tRI ≥ 2 are statistically significant ($p < 0.05$; see Methods).

A1 Whole genome topological recurrence analysis (February 2021)

SNV	SAAV	tRI	prevalence	filter†
G11083T	ORF1a:L3606F	141	5%	1, 2
C21575T	S:L5F	97	2%	1, 2
C15324T	ORF1b:N619syn	38	3%	1
C3037T	ORF1a:F924syn	34	96%	1
A17615G	ORF1b:K1383R	33	6%	
C2453T	ORF1a:L730F	31	2%	
A10323G	ORF1a:K3353R	29	2%	1
A23403G	S:D614G	29	96%	
G25563T	ORF3a:Q57H	28	26%	1
C27964T	ORF8:S24L	27	10%	
A20268G	ORF1b:L2267syn	26	8%	
C16887T	ORF1b:Y1140syn	24	1%	2
G29553A	none	21	< 1%	2
C3177T	ORF1a:P971L	20	1%	
C18877T	ORF1b:L1804syn	20	6%	

B1 Spike gene topological recurrence analysis (February 2021)

SNV	SAAV	tRI	prevalence	filter†
C22227T	S:A222V	776	19%	
C21575T	S:L5F	466	2%	1, 2
G22992A	S:S477N	433	4%	
A23403G	S:D614G	400	96%	
C21614T	S:L18F	362	8%	
C21855T	S:S98F	342	2%	
C22444T	S:D294syn	276	2%	
G23593T	S:Q677H	251	1%	
C23731T	S:T723syn	203	1%	
C24334T	S:A924syn	193	4%	
C22388T	S:L276syn	161	< 1%	
A22255T	S:I231syn	145	< 1%	
G25088T	S:V1176F	133	< 1%	
C21637T	S:P25syn	127	< 1%	
T24076C	S:G838syn	123	1%	

† **Filtration rule:** 1, 2 indicates that the corresponding site is identified as highly homoplasic in 1 *Turakhia et al.* [33, 46] and/or in 2 *De Maio et al.* [47] and is normally removed or masked.

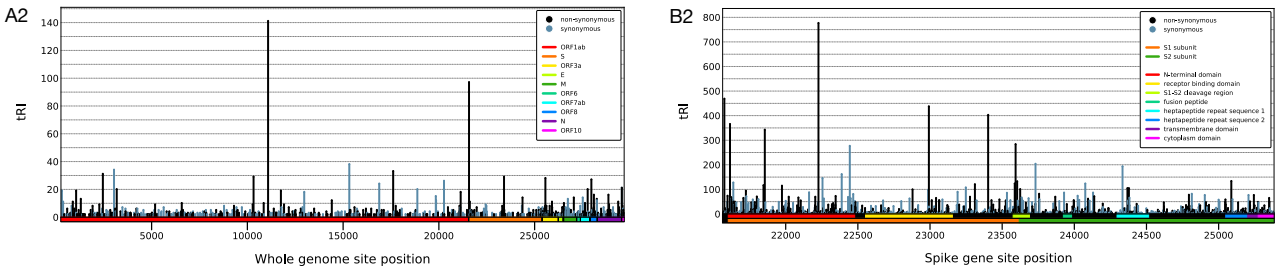


Figure 4. Topological signals during the first year of the pandemic. Topological recurrence analysis for the whole genome and the Spike gene from December 2019 until February 2021. The tables in (A1) and (B1) list the topological recurrence index (tRI) and the prevalence for mutations with strongest topological signal. In each table the last column reports filtration rules recommended in Turakhia *et al.* [33, 46] and De Maio *et al.* [47] for highly homoplasic sites, problematic sites and potential artifacts that are normally removed or masked in SARS-CoV-2 phylogenetic analyses. For a complete list see Supplementary Information Table 1 and Supplementary Information Table 2. (A2) Histogram showing the distribution of topological signals across the whole genome. In every region of the genome, reticulate events play a crucial role in the evolution of the virus. (B2) Histogram showing the distribution of topological signals across the Spike gene. There is an accumulation of topologically recurrent mutations in the S1 subunit and the S1-S2 cleavage region, as well as in the signal peptide at the beginning of the Spike gene.

medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Several of the mutations with strong tRI signal, such as `G11083T`, `C21575T`, `C15324T` and `C3037T`, are known to be highly homoplasic for various reasons and cause stability issues in the construction of phylogenetic trees [11, 32, 33, 47]. While the corresponding sites are typically masked in phylogenetic analyses, this is not necessary in our Topological Data Analysis approach as the computation of the topological recurrence index (tRI) does not rely on the construction of phylogenetic trees. We prefer to take an unbiased approach and include additional information on highly homoplasic and potentially problematic sites only in the downstream analysis of mutations with significant topological signal. Furthermore, we augment our tables of tRI scores by common masking schemes suggested in Turakhia *et al.* [33, 46] and De Maio *et al.* [47] (see Figure 4 and Supplementary Information Table 1).

Host RNA editing processes have been reported to be a factor in SARS-CoV-2 mutagenesis [48–50]. We investigated the role of innate immune editing in the topological analysis of the first year of the pandemic (see Supplementary Information Figure 7). Among mutations on the whole genome flagged by a significant tRI signal, we observed a preponderance of `C>T` transitions with a strong bias towards the positive-sense strand: 58% of the non-synonymous mutations, and 73% of the synonymous mutations were C-to-T changes. A cosine similarity analysis between the trinucleotide signature of mutations flagged with significant tRI and COSMIC signatures [51] revealed a high resemblance with signature SBS 2 that is attributed to the action of the APOBEC family of cytidine deaminases (cosine similarity of 0.39 for non-synonymous mutations and of 0.44 for synonymous mutations; see Methods). These findings suggest that topological signals are also driven by `C>T` hypermutation caused by APOBEC-mediated human host responses [52–54].

Notably, we observed that 43% (299 out of 700) of the topologically recurrent single nucleotide variations on the whole genome were synonymous. The excess and imbalance in frequencies of C-to-T transitions (73% synonymous vs. 58% non-synonymous) among topologically recurrent mutations (see Supplementary Information Figure 7) suggests that the action of APOBEC enzymes is one explanation for this observation. Moreover, it is known that synonymous mutations are not always neutral or almost neutral [55], and it is likely that the topological recurrence index also captures signals of the adaptation of the SARS-CoV-2 codon usage to the human host via codon optimization and deoptimization, processes which can affect SARS-CoV-2 protein translation efficiency [56–59]. A substantial number of fitness increasing synonymous mutations was also reported in the phylogeny-based assessment of fitness effects by Bloom & Neher [7, 60]. These observations suggest that the assessment of the adaptiveness of topologically recurrent synonymous mutations requires care, for example by invoking other indicators of adaptation. Our downstream analysis of the potential adaptiveness of topologically recurrent mutations in the present work focuses on protein changing substitutions as they are more interpretable in terms of protein selection.

The topological recurrence analysis can also be carried out for single genes, based on alignments of appropriately truncated genome sequences (see Methods). In this case, evolutionary processes outside the specific gene are ignored, which typically leads to the creation of more topological cycles and hence to a more detailed picture of the ongoing convergent evolution in the respective gene. For the Spike gene alignment, we found a total of 322 non-synonymous and 196 synonymous Spike mutations with significant tRI signal (see Figure 4, Supplementary Information Table 2 and Methods). Here signals with tRI ≥ 8 are statistically significant ($p < 0.05$; see Methods). We observed a distinct accumulation of topological signals in the S1 subunit, which is associated with host receptor recognition and contains epitopes for antibody binding [61–65], and in the Spike protein signal peptide as well as in the S1-S2 cleavage region [66] (see Figure 4).

medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

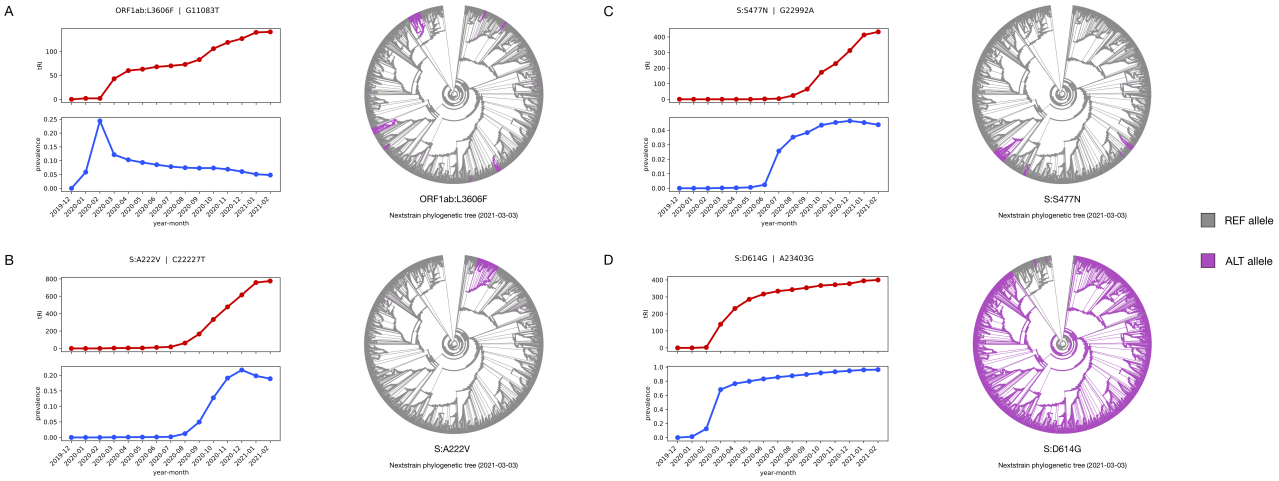


Figure 5. Typical patterns of topological signals during the first year of the pandemic. Each panel shows a comparative time series analysis (topological recurrence index (tRI) vs. prevalence) with monthly timesteps ranging from December 2019 until February 2021, in comparison with an ancestral state reconstruction obtained with standard phylogenetic methods. The tRI time series maps the dynamics of topological signals for specific mutations and therefore allows for the retrospective assessment of adaptive processes over a longer period of time. (A) Topological footprint of a highly recurrent mutation in the example of the mutation `ORF1ab:G11083T`. A time series analysis shows a monthly increase of the tRI, while the prevalence stays low. This indicates that the mutation has been reemerging frequently and steadily since the beginning of the pandemic. The consistently low prevalence suggests that the mutation is neutral or deleterious, as a beneficial substitution would be expected to establish itself in larger subpopulations. (B, C) Recurrence of the mutations `S:A222V` and `S:S477N` persists after its initial surge in prevalence in mid-2020. (D) The mutation `S:D614G` shows a pattern typical for an adaptive mutation—after a rapid increase in tRI and prevalence the tRI reaches a plateau once the mutation has become dominant, superseding the wild type in the early phase of the pandemic. All ancestral state reconstructions are based on the Nextstrain tree [19] of a curated subsample of 3,507 sequences from the GISAID dataset as of 3 March 2021 [18, 36] (see Methods). The respective amino acid substitutions are highlighted in purple.

We found a strong topological signal for the mutation `S:D614G`. A comparative time series analysis (tRI vs. prevalence) covering the first year of the pandemic revealed a steady increase in both tRI and prevalence until the new variant eventually superseded the wild type in the population (see Figure 5 and Methods). This mutation is known to increase transmissibility [17, 67] and in vitro infectiousness [68–70]. The mutations `S:A222V` and `S:S477N` are among those with strongest topological signal for recurrence (see Figure 5). Both are associated with lineage B.1.177 / 20E (EU1) which emerged in Europe in mid-2020, and `S:S477N` is now also seen in the Omicron variant B.1.1.529 [71]. While `S:S477N` is known to affect the binding affinity to the ACE2 receptor [3, 72] leading to a slight increase in fitness, there is no conclusive evidence yet whether or not `S:A222V` also results in higher transmissibility [73]. Our time series analysis for the latter shows that the particularly strong tRI signal is still rising after the initial surge in prevalence in the European summer of 2020 (see Figure 5). This suggests that `S:A222V` is notably recurrent, independently of its impact on viral fitness.

We noticed that in particular on the receptor-binding domain (RBD), several of the topologically significant amino acid changes (tRI ≥ 8) are found in Variants of Interest (VOI) or Variants of Concern (VOC) [71], with a distinct accumulation in the receptor-binding motif (see Table 1, Supplementary Information Figure 2 and Supplementary Information Figure 3). Specifically, the substitutions `S:N501Y`, `S:E484K`, `S:L452R`, `S:K417N`, `S:F490S` and `S:S494P` all result in reduced binding of polyclonal convalescent plasma [4] and exhibit a distinct increase in tRI and prevalence starting in late 2020 (see Supplementary Information Figure 3). This pattern is likely due to selective pressures induced by immune evasion in a host population

medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

with rising immunity. While both `S:N501Y` and `S:N501T` produce comparable tRI signals and induce a slight antibody escape, the fact that `S:N501Y` has a comparatively high prevalence of 19%, and is seen in several VOCs, is probably due to the additional increase in ACE2-binding affinity caused by the asparagine-to-tyrosine substitution. Topological signals for the mutations `S:Y453F` and `S:F486L` exclusively originated from a small subpopulation in minks [74–76]. The fact that tRI signals remain low (tRI = 8) despite becoming significant already in June/May 2020 suggests that both mutations have an adaptive effect in minks but do not confer a significant fitness advantage in humans.

SAAV	tRI	tendency	significant since	notable variants
S477N	433	↗	Jul 2020	Iota*, Omicron†
N439K	88	↗	Apr 2020	
S494P	64	↗	Sep 2020	Alpha*
N501Y	55	↗	Sep 2020	Alpha, Beta, Gamma, Mu†, Omicron†
N501T	50	↗	Oct 2020	
E484K	49	↗	Sep 2020	Alpha*, Beta, Gamma, Eta, Iota*, Mu†, Zeta
A520S	35	↗	May 2020	
L452R	28	↗	Dec 2020	Delta†, Epsilon, Iota*, Kappa†
V367F	27	→	March 2020	
A522S	27	↗	April 2020	
F490S	19	↗	Dec 2020	Lambda†
K417N	13	↗	Feb 2021	Beta, Delta†*, Omicron†*
Y453F	8	→	Jun 2020	Mink (Cluster 5)
F486L	8	→	May 2020	Mink
T478K	7	↗	–	Delta†, Omicron†
E484Q	6	→	–	Kappa†

* mutation found in some sequences but not all
† designated as VOI/VOC after analysis was completed in March 2021

Table 1. Amino acid changes on the receptor-binding domain with strong topological signal of convergence as of February 2021. The table displays the top ten amino acid substitutions on the receptor-binding domain with significant topological recurrence index (tRI ≥ 8), plus a few more selected mutations, together with the tendency of the tRI and the date of initial acquisition of a significant tRI signal. Several of the mutations with strong tRI signal occur in lineages designated as VOI/VOC [71]. For the full table see Supplementary Information Figure 2.

Correlation with host adaptation

Our analysis of the first year of the pandemic revealed that on the receptor-binding domain, there is a strong correlation between significant topological signals (tRI ≥ 8) and an increase in ACE2-binding affinity [3] compared to the wild type (Fisher's exact test, $p < 0.01$; Figure 6). We did not find a similar correlation between significant tRI and an increase in plasma antibody escape [4], which is plausible as immune evasion has become a relevant factor for the evolution of the virus only towards the end of 2020. However, among those mutations with strong topological signal of convergence (tRI > 12) we found a correlation between increased mean plasma antibody escape > 0.025 and the initial acquisition of a significant tRI signal after October 2020 (Fisher's exact test, $p < 0.05$; Figure 6). This indicates that beginning in late 2020, immune escape became an increasingly important evolutionary factor.

The latter observation led us to investigate to what extent topological signals of recurrence for amino acid changes on the Spike gene are modulated by the dynamic SARS-CoV-2 fitness landscape [77] shaped by increasing host immunity and changes in viral genetic background through the emergence of new dominant virus variants. To that end, we performed a topological recurrence analysis of GISAID data exclusively from the United Kingdom covering the whole pandemic (December 2019 until March 2023; see Figure 7 and Methods). We recorded a significant increase in amino acid variations newly flagged by significant tRI during the second

medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

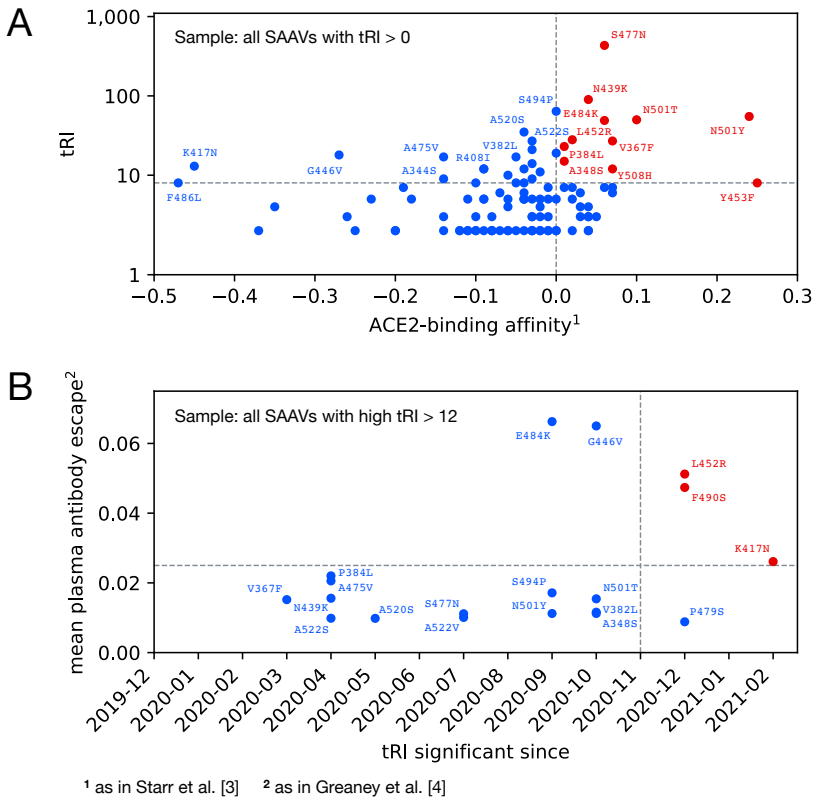


Figure 6. Topological recurrence and SARS-CoV-2 phenotypic variation during the first year of the pandemic. Both diagrams visualize contingency tables underlying Fisher's exact test. (A) There is a strong correlation between a significant topological recurrence index (tRI ≥ 8) and an increase in ACE2-binding affinity in the year from December 2019 until February 2021. (B) Among highly topologically reccurent amino acid changes (tRI > 12), more recent ones show increased mean plasma antibody escape. There is a significant correlation between increased mean plasma antibody escape > 0.025 and the initial acquisition of a significant tRI signal after October 2020 (Fisher's exact test, $p < 0.05$). Experimental data taken from Starr *et al.* [3, Table S2] and Greaney *et al.* [4, Table S3].

half of the year 2020, with a distinct accumulation in the S1-S2 cleavage region, which is a determinant of SARS-CoV-2 infectiousness and transmissibility [61, 66, 78], and to a lesser extent in the N-terminal domain and the receptor binding domain, which contain epitopes for antibody binding [61–65]. This is compatible with the observation that host adaptation during the first year of the pandemic was mainly enhancing infectiousness and transmissibility in a largely immunologically naive host population, as manifested in the Alpha variant, while immune escape became an evolutionary factor only towards the end of the year 2020 when infection-induced population immunity was on the rise [77]. After a longer period with only few tRI signals we observed another significant increase in amino acid variations newly flagged by significant tRI beginning in the second half of the year 2021, with a distinct accumulation in the N-terminal domain and the receptor binding domain in the S1 subunit as well as in the fusion peptide and the heptapeptide repeat sequence 2 in the S2 subunit, which are targets for the action of neutralizing antibodies [62–65]. This confirms that during this second phase of the pandemic, immune evasion became the dominant factor of adaptation in the presence of widespread population immunity, as manifested in the Delta and Omicron variants, after the great majority of the UK population had received a second dose of vaccination by the third quarter of 2021 [77].

medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

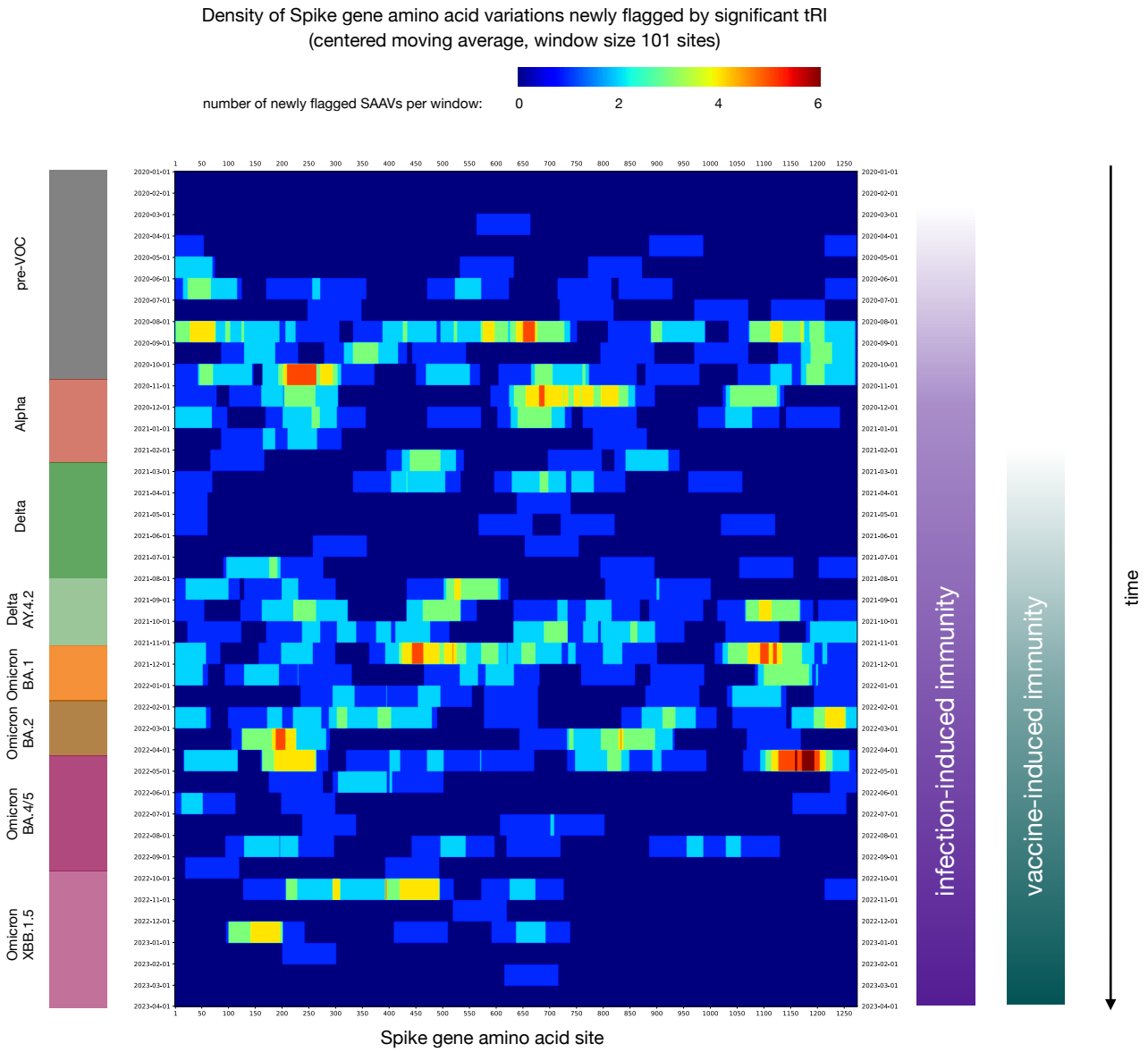


Figure 7. Topological recurrence and the SARS-CoV-2 fitness landscape. Dynamics of the interaction between topological signals of recurrence and host immunity/waves of dominant variants in the United Kingdom. For each month from January 2020 until March 2023, the heatmap shows the density of Spike gene amino acid variations for which the topological signal (tRI) turned significant during that month (density is centered moving average with window size of 101 sites). Topological recurrence analysis based on GISAID data from the UK covering the whole pandemic from December 2019 until March 2023 (see Methods); additional epidemiological information taken from Carabelli *et al.* [77, Figure 4].

We observed that many, though not all, of the mutations that are of high prevalence also tend to show a strong topological signal (see Figure 4, Supplementary Information Table 1 and Supplementary Information Table 2). While from a conceptual point of view tRI and prevalence are independent quantifiers of viral evolution, this phenomenon can largely be explained by the observation that adaptiveness is a confounding factor for both of them: evolutionary advantages lead to an increase in both tRI and prevalence. One of the advantages of the topological recurrence index, however, lies in its great sensitivity and specificity towards signals associated with convergent evolution. While the prevalence depends heavily on the dynamics of the pandemic, including founder effects, the topological signal does not.

medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

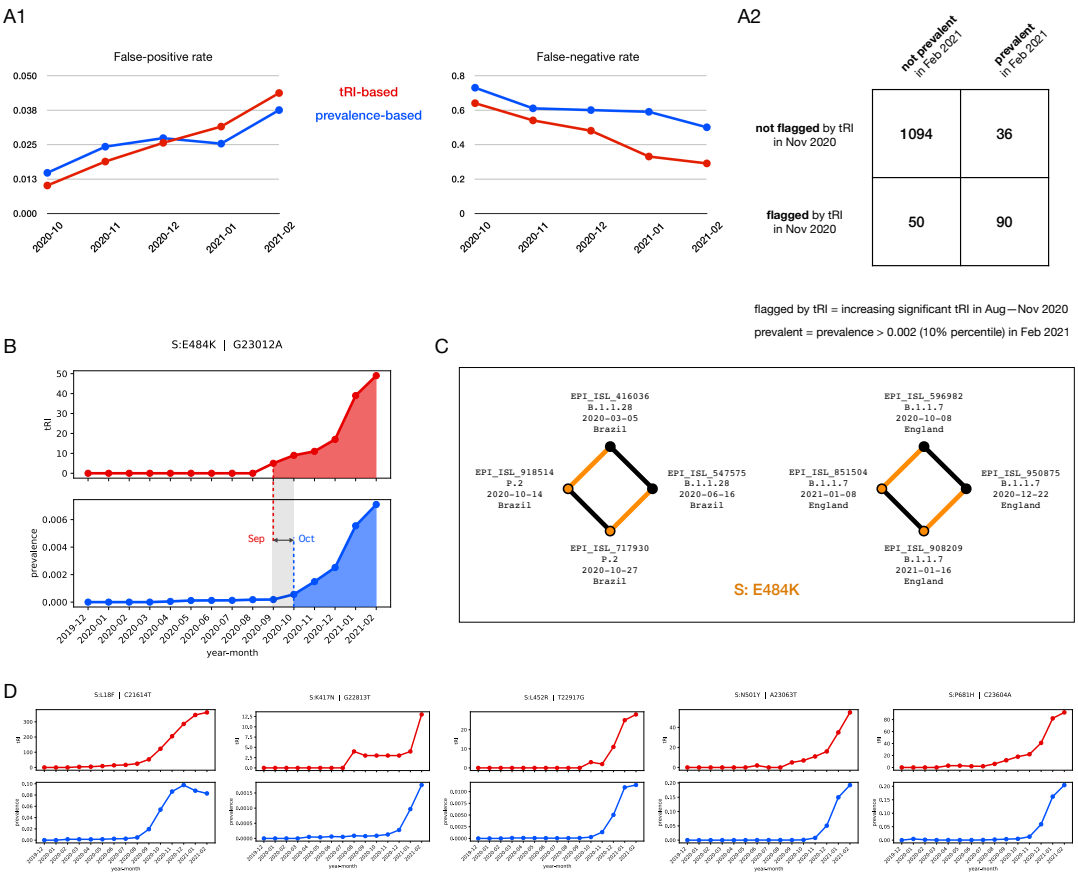


Figure 8. Surveillance of emerging potentially adaptive Spike gene mutations. (A) Specificity and sensitivity of tRI-based vs. prevalence-based identification of variable Spike amino acid sites likely to become highly prevalent (prevalence contained in 10% percentile of all mutations) three months thereafter. (A1) Reliability of identifications for each of the months October 2020 through February 2021. While false-positive rates of tRI-based and prevalence-based analyses are similar and remain small, the false-negative rate of tRI-based analyses is better than that of prevalence-based analyses (29% vs. 50%). (A2) Contingency table for tRI-based analyses for February 2021. Variable Spike amino acid sites were flagged if they exhibited a significant and increasing tRI signal from August until November 2020, and designated prevalent if their prevalence was contained in the 10% percentile of all mutations in February 2021. (B) Typical comparative time series pattern (tRI vs. prevalence) of an emerging adaptive mutation as observed in the escape mutation `S:E484K`. The tRI rises to a significant level ($\text{tRI} = 5$) in September 2020, while the prevalence stays very low $< 0.02\%$ and shows a visible increase to 0.06% by a factor of 3 for the first time several weeks later in October 2020. Both tRI and prevalence show a rising tendency in February 2021. (C) Cycle localization in the genome alignment yields geographic and temporal information about the independent acquisition of `S:E484K`. The diagram shows SNV cycles created by reticulate events involving the Zeta variant (P.2) in Brazil in October 2020, and within the Alpha variant (B.1.1.7) in England in January 2021. (D) Typical comparative time series pattern observed in emerging adaptive Spike mutations, with the onset of significant tRI signals preceding the rise in prevalence by several weeks.

Surveillance of emerging potentially adaptive mutations

The topological recurrence analysis has the capability to identify specific mutations at an early stage that are likely to confer a fitness advantage to the virus and may become prevalent in the future. We observed that the sensitivity and specificity of the topological recurrence index (tRI) is sufficient to detect signals of convergent evolution already at very low mutation frequencies, often several months before the respective mutation becomes recognizable by its prevalence in the population. As an initial assessment of the reliability of this method, we used topological signals to identify mutations at a given point in time that are likely to become

medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

highly prevalent three months later. A variable amino acid site was flagged if it exhibited a significant and increasing tRI signal that persisted from August until November 2020. We found that flagged mutations were strongly correlated with a high prevalence contained in the 10% quantile of all mutations in February 2021 (Fisher's exact test, $p < 10e{-}10$; Figure 8). Results of similar quality were obtained for mutations flagged in any of the months October 2020 through January 2021. In contrast, a similar analysis based on prevalence (replacing tRI) had about the same false-positive rate, but a significantly higher false-negative rate, especially at low mutation frequencies (see Figure 8). Further validation comes from the observation that the onset of a significant tRI signal precedes the actual rise in prevalence by several weeks for the fitness increasing mutations `S:L18F` [79], `K417N` [4], `L452R` [4, 80], `S:E484K` [4, 80], `S:N501Y` [3, 4] and `S:P681H` [81, 82] seen in the VOCs Alpha, Beta, Gamma, Delta and Omicron [71] (see Figure 8). Moreover, looking at the four VOCs Alpha, Beta, Gamma and Delta which occurred during the first year of the pandemic, among defining Spike variable amino acid sites we found a substantial number of sites expressing a significant topological signal of recurrence already months before the variant was first reported (see Supplementary Information Figure 5).

In practice, one may wish to improve the tRI-based identification of emerging potentially adaptive mutations by invoking other indicators of positive selection, such as the prevalence and its tendency, or experimental data. Focusing on the RBD, we found that the amino acid changes `S:A348S`, `S:N354D`, `S:P384L`, `S:N439K`, `S:G446V`, `S:A475V`, `S:E484G`, `S:F490S`, `S:N501T` and `S:Y508H` developed a significant tRI signal over the course of the first year of the pandemic, showed a rising prevalence in many cases, and were all associated with an increased mean plasma antibody escape > 0.01 [4, 83], but had low prevalence $< 5\%$ and had not been seen in any VOI/VOC as of February 2021 (see Table 1, Supplementary Information Figure 2 and Supplementary Information Figure 4). Mutations at these RBD residues are likely to confer a fitness advantage to the virus and might therefore appear in future variants. In fact, several months after our analysis was completed, the immune escape mutations `S:F490S` [4, 80] and `S:G446S` [4, 84], which had not been observed in notable variants before, occurred in the Lambda [85, 86] and Omicron [87] variants, which were designated as VOI/VOC in June/November 2021 [71]. By investigating explicit representatives of topological cycles in the genomic dataset, we were able to extract geographic and temporal information about the independent acquisition of topologically recurrent mutations during the pandemic (see Figure 8).

Ongoing convergent evolution and the Omicron variant

We analyzed topological signals of adaptation also at later phases of the pandemic, by applying our CoVtRec surveillance pipeline to two curated GISAID Spike gene alignments [18, 36] covering the period from December 2019 until January 2022 and March 2023, respectively (see Methods). As of January 2022, a total of 312 non-synonymous and 192 synonymous Spike mutations with significant tRI signal were found (see Figure 9, Supplementary Information Table 4 and Methods). Here signals with tRI ≥ 76 are statistically significant ($p < 0.05$; see Methods). We observed that on the receptor-binding domain, amino acid changes with significant topological signal accumulate in the receptor-binding motif (see Figure 9).

Of particular interest are the defining amino acid substitutions in the Omicron variant B.1.1.529 [87, 88], which was designated as VOC in November 2021 [71] and was rapidly spreading all over the world in January 2022 [89]. Our analysis revealed that 38% (18 out of 48) of the Spike amino acid substitutions seen in B.1.1.529 (BA.1, BA.2 and BA.3 [90, 91]) —with 50% (15 out of 30) in the sublineage BA.1 and 50% (14 out of 28) in the sublineage BA.2— show significant topological signal of recurrence (see Figure 9). This includes, among

medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

others, the following Spike amino acid substitutions: `S:G339D`, `S:N440K`, `S:S477N`, `S:T478K` and `S:N501Y`, which are known to enhance binding to the human ACE2 receptor; `S:K417N`, `S:G446S` and `S:E484A`, which may reduce polyclonal antibody binding; `S:P681H` at the S1-S2 cleavage region [66, 78, 92]. Our findings provide further insight into the possible origins of

A Selected Spike amino acid changes (January 2022)

SAAV	tRI	notable variants
G142D†	18226	Delta, Omicron
T95I	6600	Omicron*, Iota, Mu
L5F†	4217	Iota
A222V	3220	Delta*
D950N†	2949	Delta, Mu
V1264L	1542	
S112L	1370	
L18F	1252	Beta, Gamma
V1104L	1111	Delta*
S98F	1101	
Q677H	1052	Eta
S477N	850	Omicron
A701V	818	Beta, Iota
L452R	707	Delta, Kappa, Omicron*
P681H	632	Alpha, Omicron, Mu
E484K	598	Alpha*, Beta, Gamma, Eta, Iota, Mu
N501Y	548	Alpha, Beta, Gamma, Omicron, Mu
H655Y	543	Gamma, Omicron
A67V	380	Omicron*, Eta

C Omicron BA.1-3 Spike mutations (January 2022)

SAAV	tRI	SAAV	tRI
T19I	64	N460K	0
A27S	**395**	**S477N**	**850**
A67V	**380**	**T478K**	**299**
T95I	**6600**	T478R	6
G142D†	**18226**	**E484A**	**181**
Y145D	0	F486P	0
L212I	0	F486S	7
V213G	0	**F490S**	**213**
G339D	**92**	Q493R	41
G339H	0	G496S	0
R346T	12	Q498R	22
L368I	14	**N501Y**	**548**
S371F	5	Y505H	47
S371L	0	T547K	75
S373P	37	**D614G**	**787**
S373F	0	**H655Y**	**543**
S375F	39	N679K	64
T376A	4	**P681H**	**632**
D405N	8	**N764K**	**76**
R408S	4	**D796Y**	**218**
K417N	**429**	N856K	16
N440K	**201**	Q954H	5
V445P	0	N969K	5
G446S	**155**	L981F	30

Annotation: † highly homoplasic site identified in *Turakhia et al.* [33, 46] and/or in *De Maio et al.* [47]
in (C) boldface highlights significant tRI

B RBD amino acid substitutions (January 2022)

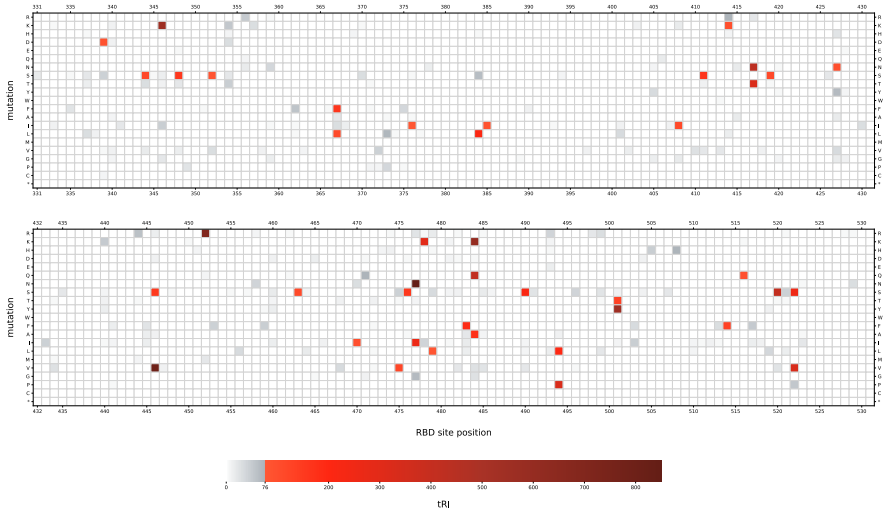


Figure 9. Topological signals for Spike gene mutations as of January 2022 and the Omicron variant. (A) The table displays selected amino acid substitutions on the Spike gene with significant topological recurrence index (tRI $\geq$ 76). For the full table see Supplementary Information Table 4. (B) Heatmap of all amino acid variations across the RBD showing topological signals of convergence. Significant signals (tRI $\geq$ 76; shown in red) accumulate in the receptor-binding motif. (C) Topological signals of convergence for the defining Spike mutations in the Omicron variant (B.1.1.529; BA.1, BA.2 and BA.3 [88, 90, 91]). 38% of these mutations are significantly topologically recurrent with tRI $\geq$ 76 (highlighted in boldface) and are likely due to convergent evolution. Corresponding tRI time series analysis charts are provided in Supplementary Information Figure 13.

medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

the Spike mutations observed in the Omicron variant [93]. In fact, all 18 substitutions with significant topological recurrence index (tRI ≥ 76) likely arose from convergent evolution, while the emergence of substitutions showing weak topological signal may be due to other reasons (see Figure 9).

Finally, a topological recurrence analysis covering the whole pandemic from December 2019 through March 2023 revealed a total of 328 non-synonymous and 212 synonymous Spike mutations with significant tRI signal (see Supplementary Information Table 5 and Methods). Here signals with tRI ≥ 75 are statistically significant ($p < 0.05$; see Methods). These mutations are presently topologically recurrent. They potentially confer a fitness advantage to the virus and could appear in future variants.

Discussion

In the current COVID-19 pandemic, the early identification of emerging adaptive mutations in large SARS-CoV-2 genomic datasets is of paramount importance. Such mutations could be associated with vaccine resistance or higher transmissibility, among other concerning attributes [94]. We present here an effective and unbiased method, based on a technique from Topological Data Analysis known as persistent homology, that can rapidly identify the presence of these mutations and is able to efficiently deal with the ever-increasing wealth of genomic sequencing data created by global public health surveillance. While phylogeny-based approaches rely on the construction of a phylogenetic tree, often in combination with phylogenetic model assumptions, our method quantifies the potential adaptiveness of mutations merely by their topological footprint in the genome alignment and is agnostic about biological reasons. It enables the longitudinal analysis of the dynamics of topological signals of adaptation at daily resolution over any given period of time. This makes it a useful tool both for the real-time monitoring and for the retrospective characterization of the convergent evolution of the coronavirus starting from the very beginning of the pandemic.

In benchmark tests, our method performed faster than state-of-the-art phylogeny-based tools such as IQTree, UShER and HomoplasyFinder [9, 29, 35]. In particular, it accomplishes the analysis of current alignments with millions of SARS-CoV-2 genomes within just one day. Orthogonal biological validation for our method comes from a strong correlation between mutations flagged as adaptive by our Topological Data Analysis approach and mutations flagged as fitness increasing in a recent phylogeny-based assessment of fitness effects of mutations by Bloom & Neher [7].

Drawing from data from GISAID [18, 36], we applied our CoVtRec analysis pipeline to characterize the evolution of the coronavirus in three different phases of the pandemic. The analysis of the first year ranging from December 2019 until February 2021 revealed a total of 700 (518) topologically recurrent mutations distributed across the whole genome (Spike gene), which shows that convergence plays a significant role in the evolution of SARS-CoV-2. This is compatible with results from phylogeny-based analyses of the adaptive evolution during the first year of the pandemic [13, 95]. We found that our method reliably identifies highly homoplasic sites across the whole genome, making it a useful tool in the design of unbiased masking schemes in phylogenetic inference.

We studied the correlation between topological signals of recurrence and host adaptation, including experimental [3, 4] and epidemiological data [77] into our topological recurrence analysis. For amino acid changes on the receptor-binding domain we found a correlation between significant topological signals and an increase in ACE2-binding affinity compared to the wild type of the virus. Moreover, our analysis suggests that topological signals of recurrence are modulated by the SARS-CoV-2 fitness landscape. We inferred that during the first year of

medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

the pandemic, adaptation was mainly enhancing infectiousness and transmissibility in a largely immunologically naive host population, while in the second phase of the pandemic beginning in the year 2021, immune evasion became an important factor of adaptation in the presence of increasing population immunity.

As we observed, our method likely also captures topological footprints of APOBEC-mediated host RNA editing processes which can give rise to recurrent hypermutable `C>T` transitions. These mutations may show significant topological recurrence index (tRI) in our analysis but could nevertheless not be fitness increasing [52, 53]. We found that this in particular applies to synonymous mutations, which make up a substantial proportion of single nucleotide variations with significant topological recurrence index. However, it is likely that other processes, such as adaptation to the human codon usage, could also drive topological signals of synonymous mutations. Caution is therefore required in the biological interpretation of significant topological signals of recurrence.

We demonstrated that our method is sufficiently sensitive to detect emerging adaptive mutations already at an early stage when mutation frequencies are still very low, and provides geographic and temporal information about virus isolates involved in concrete reticulate events. We identified mutations on the Spike gene that showed significant topological signals of recurrence with a rising tendency in February 2021. These mutations had the potential to confer a fitness advantage to the virus in a largely immunologically naive host population. In fact, several of these candidate mutations have later received special attention as features of the Lambda and Omicron Variants of Interest/Concern [71], not until several months after our analysis was completed.

We explained how our results can aid the understanding of possible reasons for the occurrence of certain mutations in Variants of Concern such as the Omicron variant. We found that as of January 2022, 38% of the Spike gene mutations observed in the Omicron variants are likely due to convergent evolution.

The topological analysis of the current phase of the pandemic in March 2023 revealed a total of 540 topologically recurrent mutations on the Spike gene. Updated CoVtRec analyses are available at `https://tdalife.github.io/covtrec`.

Epistasis has been reported to play a key role in SARS-CoV-2 evolution [5, 96–101], but the phylogeny-based computational characterization of this phenomenon across all sites along the whole genome remains challenging due to the rapidly growing combinatorial complexity in possible interactions [102]. While in the present study we confined ourselves to the assessment of the adaptive potential of single mutations in the presence of specific genetic backgrounds, our method opens new ways for the computational mapping of the epistatic interaction in the evolution of the coronavirus. For example, high-resolution time series analyses can help to identify networks of interacting mutations by comparing the onsets of significant topological signals of adaptation, possibly after restricting to fixed genetic backgrounds. In addition, topological footprints of networks of interacting mutations should be accessible by means of evaluating topological cycles in which adjacent genomes are allowed to differ by more than just one nucleotide substitution.

Our method has several limitations. First, in contrast to phylogeny-based approaches, our approach based on topological footprints will resolve homoplasies only on small genetic distance scales and will not distinguish between forward and backward mutations. Also, in its present form the topological analysis does not include insertions and deletions. Second, sequencing errors [103, 104] may tamper topological signals to some extent. We implemented several filtration rules in order to control this effect. In this regard it is also important to note that persistent homology is robust with respect to noise in the genomic dataset [105]. Third, the

medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

topological recurrence analysis is agnostic about biological reasons, and mutations flagged by significant topological recurrence index (tRI) need not always be adaptive. The downstream analysis of mutations with significant tRI therefore requires care and should invoke additional biological information, such as masking schemes for problematic sites, epidemiological data or experimental evidence of adaptation available from the literature.

Based on these insights, we propose persistent homology as a surveillance system for emerging potentially adaptive mutations in the ongoing evolution of the coronavirus SARS-COV-2 as well as for the retrospective characterization of the dynamics of adaptation for specific mutations. We foresee this capability also in future pandemics of other pathogens. Our method allows for a targeted mapping of topologically recurrent mutations in any region of the genome, compiling a shortlist of flagged mutations which can guide and motivate the experimental study of adaptive effects of specific mutations, also in genes other than the Spike gene. We envision a combined effort between public health organizations with extensive sequencing of viral genomes, the computational characterization of potentially adaptive variants, and the experimental phenotypic characterization of these variants.

Methods

Data acquisition and data preparation

We used three separate datasets of SARS-CoV-2 genome sequences provided by the GISAID EpiCoV Database [18, 36]. The first dataset was obtained by downloading all available SARS-CoV-2 whole genome sequences as of 28 February 2021, isolated from human and animal hosts, that carried the following attributes: "complete", "high coverage", "low coverage excluded", "collection date complete". This dataset comprised 450,473 sequences. We aligned all sequences with `MUSCLE (v3.8.31)` [106], using as reference genome the sequence Wuhan/Hu-1 with accession number `EPI_ISL_402125`, truncated at the start codon of `ORF1ab` (reference site position 266) and the stop codon of `ORF10` (reference site position 29,674). Subsequently all sequences containing at least one ambiguous site (letter "N") were removed. This resulted in an alignment comprising 303,651 complete SARS-CoV-2 genomes of length 30,130nt. The second dataset is the GISAID alignment `msa_0126.fasta` from January 2022. It comprises 6,703,606 SARS-CoV-2 whole genome sequences that have been aligned to the reference sequence Wuhan/WIV04 with GISAID accession number `EPI_ISL_402124` using `MAFFT (v7.497)` [107]. Sequences were truncated to the Spike gene (reference site positions 21,563 to 25,384), and finally sequences containing any letters other than A, C, T or G were removed. This resulted in an alignment comprising 318,851 distinct complete SARS-CoV-2 Spike genes of length 4,796nt. The third dataset is the GISAID alignment `msa_0324.fasta` from March 2023. It comprises 13,766,674 SARS-CoV-2 whole genome sequences that have been aligned to the reference sequence Wuhan/WIV04 with GISAID accession number `EPI_ISL_402124` using `MAFFT (v7.497)` [107]. Sequences were truncated to the Spike gene (reference site positions 21,563 to 25,384), and finally sequences containing any letters other than A, C, T or G were removed. Subsequently two sub-alignments were selected: one with sequences exclusively from the United Kingdom (England, Wales, Scotland and Northern Ireland), resulting in an alignment comprising 91,063 distinct complete SARS-CoV-2 Spike genes of length 7,942nt; and a second one with sequences from all over the world that occur in at least two identical copies, resulting in an alignment comprising 220,978 distinct complete SARS-CoV-2 Spike genes of length 7,942nt. A list of accession numbers of all GISAID sequences used in this work, along with an acknowledgement of the contributions of both the submitting and the originating laboratories, can be retrieved from https://doi.org/10.55876/gis8.230731by.

medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

The experimental data on viral phenotypes by Starr *et al.* and Greaney *et al.* was downloaded from [3, Table S2] and [4, Table S3]. The values for the mean plasma antibody escape used in Supplementary Information Figure 2 were computed by averaging mutation escape values for every substitution in [4, Table S3] over all subjects.

The data on fitness effects of Spike gene amino acid changes as of March 2023 by Bloom & Neher was downloaded from [108, Spreadsheet `aa_fitness.csv`]. Subsequently the columns "gene", "aa_site", "aa" and "fitness" were extracted.

The data on filtration rules for highly homoplasic sites, problematic sites and potential artifacts in SARS-CoV-2 sequence alignments by Turakhia *et al.* and De Maio *et al.* was taken from [47] and downloaded from [46, Spreadsheet `problematic_sites_sarsCov2.vcf`]. Subsequently the columns "POS", "FILTER" and "INFO" were extracted.

The data on COSMIC mutational signatures for Genome `GRCh37` by COSMIC was downloaded from [109]. Subsequently the column "SBS2" was extracted.

Distance matrices

We used `Hammingdist (v0.19.0)` [41] to compute the *genetic distance matrix* of a given alignment of genome sequences. For any pair of sequences in the alignment, this matrix gives the Hamming distance between the two sequences, which is the number of site positions at which the nucleotides in the two aligned sequences differ. Noteworthy, our convention in this work is that insertions and deletions (dashes "-" in aligned sequences) do not contribute to the genetic distance.

From the whole genome alignment covering the first year of the pandemic we created 15 time buckets, each ranging from December 2019 to one of the months between December 2019 and February 2021. For each time bucket, a *time bucket sub-alignment* of all genetically distinct sequences whose collection dates belong to the given time bucket was created by selecting isolates by their date stamp and removing genetically identical sequences (Hamming distance = 0). The largest time bucket sub-alignment ranging from December 2019 to February 2021 contained 161,024 genetically distinct sequences. Then for each such sub-alignment the corresponding genetic distance matrix, which is a sub-matrix of the distance matrix of the whole alignment, was derived. We obtained 15 distance matrices of whole genome time bucket sub-alignments. This process was repeated for all sub-alignments after truncating sequences to the Spike gene (reference site positions 21,563 to 25,384). We obtained 15 distance matrices of Spike gene time bucket sub-alignments.

The genetic distance matrices for the other two Spike gene alignments (as of January 2022 and March 2023) were computed in a similar fashion. Subsequently a Vietoris-Rips transformation as implemented in the MuRiT algorithm was applied [110].

Topological Data Analysis and viral evolution

Topological Data Analysis (TDA) is a field of data science that aims to study the shape of large datasets, by extracting topological structures and patterns. Such topological structures have associated dimensions: structures of dimension zero can be thought of as the connected components, and structures of dimension one are essentially the loops, or topological cycles, of the dataset. Structures of higher dimensions can also be defined, but are also more difficult to interpret. Here we are interested in reticulate evolutionary processes, thus we choose to focus on topological structures in dimension one, since topological cycles can be interpreted as signals of divergence from phylogenetic trees (see Figure 2).

medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Datasets often come as point clouds: in our setting, each point corresponds to a virus genome sample, and lies in a high-dimensional space where each nucleotide of the genome is a dimension. A common way to extract the phylogenetic network from this point cloud simply amounts to connecting samples as soon as their genetic distance is less than a given threshold $r > 0$. This results in a (neighborhood) graph, whose set of cycles provides candidates for the topological structures in dimension one of the true underlying network. However, a main limitation of this approach comes from the fact that relevant topological structure typically appears at multiple scales (see Supplementary Information Figure 8).

The most common way to handle this issue in Topological Data Analysis is to actually compute and track the cycles for all possible values of r ranging from 0 to $+\infty$. As r increases, some cycles can appear, and some already existing cycles can disappear, or get filled in. The whole point of Topological Data Analysis is to record, for each cycle, its radius of appearance, or birth time, and radius of disappearance, or death time. This is called the *persistent homology* of the point cloud. The construction, based on a scale parameter r, can be summarized as follows. The input is a distance matrix describing the dataset, considered as a finite metric space. First, consider the *geometric graph* at scale r, whose vertices are the data points, with any two points connected by an edge whenever their distance is at most r. Generalizing this construction, the *Vietoris–Rips complex* at scale r connects any subset of the data points with a simplex (an edge, a triangle, a tetrahedron, or a higher-dimensional generalization thereof) whenever all pairwise distances in the subset are at most r. A Vietoris–Rips complex is a particular type of *simplicial complex*, a higher-dimensional generalization of graphs which is of crucial interest in algebraic topology, in particular in homology theory. The family of Vietoris–Rips complexes for all parameters r is called the *Vietoris–Rips filtration*. It provides a multiscale method to extract cycles of various sizes, and to encode them in a so-called *persistence barcode*: each bar, or interval, in this barcode corresponds to a cycle representing a topological feature (a reticulate evolutionary process in our case), and the bar endpoints correspond to its radii of birth and death (the maximum genetic distance between consecutive samples forming the cycle, and, roughly, the maximum pairwise genetic distance between samples forming the cycle).

Each bar indicates the presence of a reticulate event, implying that the evolutionary history cannot be fully explained in terms of a single phylogenetic tree. The mathematical background of this phenomenon is a classical theorem due to Rips, which asserts that trees have trivial persistent Vietoris–Rips homology [111]. The corresponding cycle in the associated reticulate phylogeny can then be localized in the sequence alignment by tracing it back to the isolates that constitute the reticulate event. Moreover, the length of the bar represents the cycle size. In our case, this corresponds to the length of the reticulate evolutionary process, which allows to distinguish, for instance, between homoplasies and recombinations (see Figure 2).

Computation of persistent homology

`Ripser` is a state-of-the-art software for the computation of persistent homology based on the topological construction of Vietoris–Rips complexes, developed by one of the authors [40]. For the computation of the persistence barcode, `Ripser` resorts to various optimizations, which are crucial when handling datasets of the size considered here. Notably, `Ripser` computes persistent cohomology, which is not based on cycles but instead on cocycles, often described intuitively as *cuts* that tear open a hole. In order to obtain the requisite cycles representing the features in persistent homology, we used a custom version of `Ripser` that subsequently carries out a second computation, this time based on cycles instead of cocycles. While a naive computation based on homology would be prohibitively expensive, the previous computation of the persistence barcode based on cocycles makes the subsequent computation of representative

medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

cycles feasible.

For our computations, we used a customized version of `Ripser` to compute the representative cycles for the persistent homology of the Vietoris–Rips filtration associated to the genetic distance matrix for each time bucket sub-alignment (whole genome and Spike gene). As we are only interested in SNV cycles, the computation of persistence barcodes for the time bucket sub-alignments was restricted to small genetic distance scales (`Ripser` scale parameter threshold set to 2), which greatly increases the speed of the computation.

The homological features identified by persistent homology admit different representative cycles. In order to obtain cycles that fit tightly to the data points, our customized version of Ripser uses a method called *exhaustive reduction* [112, 113], which can be roughly summarized as follows. Whenever a representative cycle contains an edge that also appears in another cycle as the longest edge, a tighter representative can be obtained by replacing the edge with the remaining edges from the other cycles, which all have shorter length.

Topological features are statistically significant

We estimated the expected number of topological cycles in persistent homology that are created by random homoplasic events in the GISAID dataset covering the first year of the pandemic with 161,024 genetically distinct whole genome sequences collected from December 2019 until February 2021. To this end, we simulated several evolutionary scenarios under the following assumptions: uniform probability distribution for substitutions across the genome, no variations in fitness, and zero recombination rate.

We generated forward simulations of viral evolution based on a Wright-Fisher model using `SANTA-SIM (v1.0)` [114] with fixed parameters: number of generations ($N = 10,000$), number of sequences sampled from the population per time step ($n = 15$), recombination rate ($\rho = 0$), and variable parameters: mutation rate per site per generation, effective initial population, carrying capacity, population growth rate per generation. We considered five scenarios: In scenarios I-III we varied the mutation rate under the assumption of fixed population size, while in scenarios IV and V we investigated the effects of logistic growth of the viral population (see Supplementary Information Figure 9). The range of mutation rates in scenarios I-III were chosen such that the diversity in the simulated phylogenies are in close correspondence to the observed diversity in the GISAID dataset. While a mutation rate of $\mu = 0.75E-7$ substitutions per generation per site systematically underestimates the maximal distances to the root, the highest value of $\mu = 1.25E-7$ produces slightly larger maximal values. In fact, scenario II with $\mu = 1.00E-7$ reproduced the observed maximal distance accurately and provides a good approximation of the GISAID dataset (see Supplementary Information Figure 9). Major differences between simulations and the GISAID dataset are likely due to epidemiological phenomena in the ongoing pandemic such as variable population size and sequencing rate, and the spread of certain variants.

For each of the simulated datasets, we computed its persistent homology in dimension one using `Hammingdist` [41] and `Ripser` [40]. In order to keep overall computational expenses at a reasonable level, we resorted to extrapolations from smaller simulated datasets to the size of the GISAID dataset with 161,024 sequences. For all scenarios we produced 100 simulations for each of the following values of the effective population p (resp. carrying capacity c): 100, 500, 1000, 2500, 5000, 7500, 10^4, 10^5. Additionally, we included five simulations for $p = 10^6$ to achieve a better support of the extrapolation fit. For each value of p we randomly chose 60% of the simulations as training data, used to determine the parameters of different models in a non-linear least squares fit, while the remaining 40% were reserved for later validation and comparison of the models.

medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

For each scenario we considered a quadratic, cubic, powerlaw and exponential model for the observed points (x_i, y_i), and linear and powerlaw fits for the squared residuals $(y_i - y_{\text{fit}}(x_i))^2$ in the training data (see Supplementary Information Figure 11). In each model, we then used the resulting fits mean(x) and var(x) as estimators for the mean and variance of an underlying Panjer $(a, b, 0)$-class distribution [115, 116]. The quantiles of the observed number of cycles in the training data fit the quantiles of the Panjer distribution with corresponding mean and variance remarkably well (see Supplementary Information Figure 10). We then determined the likelihood $L = \prod_i P_{\text{Panjer}}(y = y_i|\text{mean}(x_i), \text{var}(x_i))$ to observe the validation data $\{(x_i, y_i)\}$. For each model, the corresponding log-likelihoods are listed alongside the corresponding fits in Supplementary Information Figure 11. According to the log-likelihoods, the variance of the Panjer distribution is generally best described by a powerlaw behaviour. An exception is scenario II, for which the small sample of 5 simulations at $p = 10^6$ has an uncharacteristically small variance that skews the fits and corresponding likelihoods. Among the models that assume a powerlaw dependence of the variance, again with exception of scenario II, the cubic-powerlaw model yields maximum likelihoods. Finally, we determined the 95% prediction intervals for the expected numbers of random cycles by use of the cubic-powerlaw extrapolation of mean and variance of a Panjer distribution (see Supplementary Information Figure 9). The validation data of scenarios I, IV and V, which were all based on the same mutation rate, are well described by the prediction intervals of scenario I. The prediction intervals of scenario V differ significantly from the other two scenarios only at high numbers of distinct sequences. This difference arises because simulations in scenario V generally produce fewer distinct sequences than scenario I and IV, such that a steeper extrapolation is not sufficiently penalized. Hence, the prediction intervals of scenario V illustrate the error margins of the extrapolations, but are not likely to faithfully represent the expected number of one-dimensional cycles. We also observe that higher mutation rates in scenarios II and III lead to smaller numbers of one-dimensional cycles in the dataset.

In conclusion, the 95% prediction interval of scenario V yields an upper bound between 1023 and 1171 expected random cycles in a dataset comparable to the GISAID dataset with 161,024 distinct sequences. Moreover, since the diversity of the GISAID dataset is better approximated by scenario II than by scenarios I, IV or V, it is reasonable to rely on the prediction interval of scenario II, which predicts that in 95% of the cases we expect between 362 and 408 random cycles (see Supplementary Information Figure 9).

Topological recurrence analysis

We performed a topological recurrence analysis both for the whole genome and for the Spike gene. In principle, this analysis can be done for any region of the genome, by working with intervals in the sequence alignment consisting of appropriately truncated genomes [38, 39]. As it turns out, the number of topological cycles, and hence the resulting tRI scores, will heavily depend on the choice of these intervals. Generally speaking, as evolutionary processes outside the specific genome region are ignored, genetic distances between isolates will decrease. This typically leads to the formation of more topological cycles at small genetic distance scales, and hence to stronger tRI signals. There is a trade-off between the strength of tRI signals and the genomic scope of the analysis. This explains why in our analysis of the first year of the pandemic, tRI signals in the Spike gene analysis are normally stronger than corresponding signals in the whole genome analysis (see Figure 4). We remark that there is in general no relation between tRI scores for different genomic regions.

Regarding the alignments covering the first year of the pandemic, we proceeded as follows: For each time bucket sub-alignment (whole genome and Spike gene) a complete list of SNV

medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

cycles (topological cycles all of whose edges correspond to single nucleotide variations) in this alignment was generated from the corresponding `Ripser` output. For each edge in an SNV cycle the endpoints of the edge correspond to a pair of uniquely determined sequences in the alignment that differ in exactly one nucleotide site position and hence determine an SNV. Then for each such SNV, its topological recurrence index (tRI) is by definition the total number of all SNV cycles containing an edge that gives rise to the given SNV. We restricted our analysis to SNVs with the following two properties: (i) one of the two nucleotides involved in the SNV agrees with the nucleotide in the reference sequence `EPI_ISL_402125` at that site position, and (ii) the SNV is isolated in the sense that at the two preceding and following site positions the nucleotides are the same as in the reference sequence. These two conditions ensure that the corresponding SAAV is uniquely determined by the SNV. We used custom code implemented in Python to compute the tRI of each such SNV for every time bucket sub-alignment (whole genome and Spike gene). Moreover, for every whole genome time bucket sub-alignment the prevalence of every SNV was computed as the quotient of the number of all sequences carrying that SNV by the number of all sequences in that sub-alignment. Note that the sub-alignments entering into this computation consisted of genetically distinct sequences. Finally, for every SNV the measurements of both tRI (whole genome and Spike gene) and prevalence for all time buckets were combined into a time series analysis chart.

Even if all SNV cycles arose through random processes, it is expected that the resulting tRIs are distributed uniformly among all observed mutations. So the probability for a given mutation to have tRI $\geq k$ is given by a binomial distribution where the number of trials corresponds to the number of mutations in SNV cycles, and the probability for success is the inverse of the number of mutations that are realized in the dataset. From this we deduce that in the whole genome analysis a tRI ≥ 2 is highly significant ($p < 0.01$), while for the Spike gene analysis any signal with tRI ≥ 8 is significant ($p < 0.05$).

For the topological recurrence analysis of the Spike gene alignments as of January 2022 and March 2023, we used Vietoris-Rips transformations in multipersistent homology as implemented in the MuRiT algorithm [110] to compute tRI time series analysis charts at daily resolution from the natural stratification by time induced by collection dates of viral isolates. For each mutation, we computed the tRI growth rate as the 14 days moving average of tRI data. We remark that the tRI growth rate is tampered by computational artifacts at the time when the tRI first attains a positive value. Moreover, towards the end of the period covered by the sequence data the sampling rate is normally still low, which will result in reduced tRI growth rates.

The effect of sequencing errors on topological results

Persistent homology is robust with respect to noise in datasets [105]. Nevertheless, sequencing errors in the SARS-CoV-2 alignments [103, 104] might tamper results of the topological recurrence analysis. We implemented the following filtration rules in order to control this effect. During data preparation, sequences not labeled "high coverage" and "complete" by GISAID, or containing any letters other than A, C, T or G, were removed from the alignment. Moreover, in subsequent analyses we only used significant tRI signals, while non-significant signals were discarded. Here the tRI significance level was computed from a test statistics with the null hypothesis assuming that tRI scores are distributed randomly across all sites on the genome. The combination of these filtration rules keeps the effect of sequencing errors on tRI signals to a minimum.

Our tRI significance levels are corroborated by the following estimate of the expected amount of erroneous tRI signal per mutation. Here we assume that the probability of a sequencing

medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

error at a given nucleotide site is of the order of 10^{-3} for typical GISAID consensus genomes. The creation of a tRI signal for a given SNV requires the formation of a topological SNV cycle, a process that involves at least *two* edges associated with some SNV (see Figure 1). The probability for this to be due to sequencing errors is therefore of the order of $10^{-3} \times 10^{-3} = 10^{-6}$. The GISAID alignments we used contain about 300,000 resp. 4,000,000 distinct genomes. Hence for a given mutation the expected errors in tRI are $10^{-6} \times 300,000 = 0.3$ resp. $10^{-6} \times 4,000,000 = 4$, which is below the tRI significance levels we observed in our analyses.

In view of the large number of available sequences, a common method to contain sparse sequencing errors is to confine the analysis to subalignments of genomes that appear in multiple identical copies in the alignment. We observed that the qualitative structure of the tRI landscape along the Spike gene did not change under this additional filter and characteristic and significant signals persisted (see Supplementary Information Figure 6). On the other hand, the topological analysis is sensitive enough to detect cycles that arise from unique genomes in the alignment which may nevertheless be correctly sequenced and carry important information about the evolution of the virus. As these valuable signals might otherwise have been lost, we preferred not to use this additional filter in our analyses.

Performance analysis

We performed a basic runtime comparison between Topological Data Analysis (TDA)-based methods and standard phylogeny-based methods for a random sample of 5,000 SARS-CoV-2 genomes, and ten nested sub-alignments thereof, drawn from the GISAID alignment covering the first year of the pandemic. We used `IQTree` [35] (v2.1.3, with default settings and fast search option) and `UShER` [29] (v0.5.6, with default settings) to reconstruct phylogenetic trees. Here for each sub-alignment, UShER placed the 500 newly added samples using a mutation-annotated tree obtained from the previous sub-alignment. The subsequent homoplasy analysis was performed with `HomoplasyFinder` [9] (with default settings). For the TDA-analysis we used `Hammingdist` [41] (v0.15.0) to generate genetic distance matrices, and a custom version of `Ripser` [40] (with scale parameter threshold set to 2) for the subsequent computation of persistence barcodes and representative cycles. All computations were carried out on a machine with Intel Xeon E7-4850 v4 processors and 128 kernels. The resulting runtimes for each sample are shown in Figure 3.

The computation of the genetic distance matrix for the whole genome alignment covering the first year of the pandemic with 303,651 sequences was carried out with `Hammingdist` [41] (v0.13.0) on a machine with Intel Xeon Gold 6230R processors and 52 kernels. The runtime was 57 minutes and the memory usage was 36 gigabytes for the whole genome analysis (49 seconds and 2 gigabytes for the Spike gene analysis).

The computation of the persistence barcodes for all monthly sub-alignments of the corresponding alignment with 161,024 genetically distinct sequences was carried out with `Ripser` [40] (with scale parameter threshold set to 2) on an Intel Xeon Gold 6230R processor. The runtime and memory usage for each sub-alignment are shown in Figure 3. For the largest time bucket sub-alignment ranging from December 2019 to February 2021 with 161,024 genetically distinct sequences, the runtime was 45 minutes and the memory usage was 49 gigabytes for the whole genome analysis (59 seconds and 2.1 gigabytes for the Spike gene analysis).

For the GISAID alignment as of March 2023 with 13,766,674 SARS-CoV-2 Spike gene sequences the total computing time for the topological recurrence analysis was 12.7 hours.

medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Ancestral state reconstruction analysis

For the study of the evolutionary histories of topologically highly recurrent mutations (see Figure 5) we performed ancestral state reconstruction analyses using `Mesquite (v3.61)` [117]. As input we used a curated alignment of 3,507 genome sequences and its corresponding Maximum-Likelihood tree, downloaded from Nextstrain [19] on 3 March 2021. The tree was rooted using the oldest sequence available (`EPI_ISL_406798`, collected on 26 December 2019). We inferred the evolution of each amino acid of interest along this SARS-CoV-2 tree using a parsimony approach.

Mutational patterns and signatures in topological signals

We performed a frequency analysis for nucleotide transitions with significant tRI signal across the whole genome during the first year of the pandemic from December 2019 until February 2021 (see Supplementary Information Figure 7). The analysis was carried out separately for non-synonymous mutations and synonymous mutations, based on tRI scores taken from Supplementary Information Table 1. We analyzed the local sequence context of these transitions by determining the signature of relative allele frequencies of trinucleotide motifs centered at a given single nucleotide variation. Here the relative allele frequency was computed as the quotient of the total number of trinucleotide motifs `N[X>Y]N` centered at a mutation `X>Y` with significant tRI by the total number of triplets `NXN` in the reference genome Wuhan/Hu-1 `EPI_ISL_402125`. In this way, we obtained mutational signatures both for non-synonymous and synonymous mutations. We compared the resulting signatures with the COSMIC signatures [51] for Genome `GRCh37` taken from [109] by restricting to trinucleotide motifs centered at any of the six substitutions `C>A`, `C>G`, `C>T`, `T>A`, `T>C` and `T>G`. Subsequently we computed the cosine similarity between our signatures and COSMIC signature SBS 2 for non-synonymous (cosine similarity of 0.39) and synonymous mutations (cosine similarity of 0.44).

Data availability

The SARS-CoV-2 genome data used in this work are available from the GISAID EpiCov Database [18, 36] at `https://www.gisaid.org`. To view the contributors of each individual sequence with details such as accession number, Virus name, Collection date, Originating Lab and Submitting Lab and the list of Authors, visit `https://doi.org/10.55876/gis8.230731by`. Experimental data on viral phenotypes by Starr *et al.* and Greaney *et al.* is available from [3, Table S2] and [4, Table S3]. Data on fitness effects of Spike gene amino acid changes by Bloom & Neher is available from [108]. Data on filtration rules for highly homoplasic sites, problematic sites and potential artifacts in SARS-CoV-2 sequence alignments by Turakhia *et al.* and De Maio *et al.* is available from [46, 47]. Data on COSMIC mutational signatures for Genome `GRCh37` by COSMIC is available from [109].

Code availability

Code used for the analyses is available at `https://github.com/ssciwr/hammingdist` and `https://github.com/Ripser/ripser/tree/tight-representative-cycles`. All other code is available from the corresponding authors upon request.

medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

References

1. Hodcroft, E. B., De Maio, N., Lanfear, R., *et al.* Want to Track Pandemic Variants Faster? Fix the Bioinformatics Bottleneck. *Nature* **591,** 30–33 (2021). doi:10.1038/d41586-021-00525-x.
2. Schrörs, B., Riesgo-Ferreiro, P., Sorn, P., *et al.* Large-scale analysis of SARS-CoV-2 spike-glycoprotein mutants demonstrates the need for continuous screening of virus isolates. *PLOS ONE* **16** (ed Khudyakov, Y. E.) e0249254 (2021). doi:10.1371/journal.pone.0249254.
3. Starr, T. N., Greaney, A. J., Hilton, S. K., *et al.* Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell* **182,** 1295–1310.e20 (2020). doi:10.1016/j.cell.2020.08.012.
4. Greaney, A. J., Loes, A. N., Crawford, K. H., *et al.* Comprehensive Mapping of Mutations in the SARS-CoV-2 Receptor-Binding Domain That Affect Recognition by Polyclonal Human Plasma Antibodies. *Cell Host & Microbe* **29,** 463–476.e6 (2021). doi:10.1016/j.chom.2021.02.003.
5. Zahradník, J., Marciano, S., Shemesh, M., *et al.* SARS-CoV-2 variant prediction and antiviral drug design are enabled by RBD in vitro evolution. *Nature Microbiology* **6,** 1188–1198 (2021). doi:10.1038/s41564-021-00954-4.
6. Starr, T. N., Greaney, A. J., Addetia, A., *et al.* Prospective Mapping of Viral Mutations That Escape Antibodies Used to Treat COVID-19. *Science* **371,** 850–854 (2021). doi:10.1126/science.abf9302.
7. Bloom, J. D. & Neher, R. A. Fitness effects of mutations to SARS-CoV-2 proteins. *bioRxiv* (2023). doi:10.1101/2023.01.30.526314.
8. Sagulenko, P., Puller, V. & Neher, R. A. TreeTime: Maximum-Likelihood Phylodynamic Analysis. *Virus Evolution* **4** (2018). doi:10.1093/ve/vex042.
9. Crispell, J., Balaz, D. & Gordon, S. V. HomoplasyFinder: A Simple Tool to Identify Homoplasies on a Phylogeny. *Microbial Genomics* **5** (2019). doi:10.1099/mgen.0.000245.
10. Van Dorp, L., Richard, D., Tan, C. C. S., *et al.* No Evidence for Increased Transmissibility from Recurrent Mutations in SARS-CoV-2. *Nature Communications* **11,** 5986 (2020). doi:10.1038/s41467-020-19818-2.
11. Van Dorp, L., Acman, M., Richard, D., *et al.* Emergence of Genomic Diversity and Recurrent Mutations in SARS-CoV-2. *Infection, Genetics and Evolution* **83,** 104351 (2020). doi:10.1016/j.meegid.2020.104351.
12. Zahradník, J., Nunvar, J. & Schreiber, G. SARS-CoV-2 Convergent Evolution as a Guide to Explore Adaptive Advantage. *bioRxiv* (2021). doi:10.1101/2021.05.24.445534.
13. Rochman, N. D., Wolf, Y. I., Faure, G., *et al.* Ongoing Global and Regional Adaptive Evolution of SARS-CoV-2. *Proceedings of the National Academy of Sciences* **118,** e2104241118 (2021). doi:10.1073/pnas.2104241118.
14. Obermeyer, F., Jankowiak, M., Barkas, N., *et al.* Analysis of 6.4 million SARS-CoV-2 genomes identifies mutations associated with fitness. *Science* **376,** 1327–1332 (2022). doi:10.1126/science.abm1208.
15. Lee, B., Sohail, M. S., Finney, E., *et al.* Inferring effects of mutations on SARS-CoV-2 transmission from genomic surveillance data (2022). doi:10.1101/2021.12.31.21268591.
16. Maher, M. C., Bartha, I., Weaver, S., *et al.* Predicting the mutational drivers of future SARS-CoV-2 variants of concern. *Science Translational Medicine* **14** (2022). doi:10.1126/scitranslmed.abk3445.
17. Korber, B., Fischer, W. M., Gnanakaran, S., *et al.* Tracking Changes in SARS-CoV-2 Spike: Evidence That D614G Increases Infectivity of the COVID-19 Virus. *Cell* **182,** 812–827.e19 (2020). doi:10.1016/j.cell.2020.06.043.

medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

18. Khare, S., Gurry, C., Freitas, L., *et al.* GISAID's Role in Pandemic Response. *China CDC Weekly* **3,** 1049–1051 (2021). doi:10.46234/ccdcw2021.255.

19. Hadfield, J., Megill, C., Bell, S. M., *et al.* Nextstrain: Real-Time Tracking of Pathogen Evolution. *Bioinformatics* **34,** 4121–4123 (2018). doi:10.1093/bioinformatics/bty407.

20. *Nextstrain* https://nextstrain.org/ visited on 2023-01-10.

21. O'Toole, Á., Scher, E., Underwood, A., *et al.* Assignment of Epidemiological Lineages in an Emerging Pandemic Using the Pangolin Tool. *Virus Evolution* (2021). doi:10.1093/ve/veab064.

22. O'Toole, Á., Pybus, O. G., Abram, M. E., *et al.* Pango lineage designation and assignment using SARS-CoV-2 spike gene nucleotide sequences. *BMC Genomics* **23** (2022). doi:10.1186/s12864-022-08358-2.

23. *Cov-Lineages* https://cov-lineages.org/ visited on 2023-01-10.

24. Ferreira, R.-C., Wong, E., Gugan, G., *et al.* CoVizu: Rapid analysis and visualization of the global diversity of SARS-CoV-2 genomes (2021). doi:10.1101/2021.07.20.453079.

25. *CoVizu* https://filogeneti.ca/covizu/ visited on 2023-01-10.

26. Weaver, S., Shank, S. D., Spielman, S. J., *et al.* Datamonkey 2.0: A Modern Web Application for Characterizing Selective and Other Evolutionary Processes. *Molecular Biology and Evolution* **35,** 773–777 (2018). doi:10.1093/molbev/msx335.

27. Cherian, S., Potdar, V., Vipat, V., *et al.* Phylogenetic classification of the whole-genome sequences of SARS-CoV-2 from India & evolutionary trends. *Indian Journal of Medical Research* **153,** 166 (2021). doi:10.4103/ijmr.ijmr_3418_20.

28. *Datamonkey* https://http://covid19.datamonkey.org/ visited on 2023-01-10.

29. Turakhia, Y., Thornlow, B., Hinrichs, A. S., *et al.* Ultrafast Sample placement on Existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. *Nature Genetics* **53,** 809–816 (2021). doi:10.1038/s41588-021-00862-7.

30. Turakhia, Y., Thornlow, B., Hinrichs, A., *et al.* Pandemic-scale phylogenomics reveals the SARS-CoV-2 recombination landscape. *Nature* **609,** 994–997 (2022). doi:10.1038/s41586-022-05189-9.

31. *UCSC Genome Browser* http://genome.ucsc.edu visited on 2023-01-10.

32. Morel, B., Barbera, P., Czech, L., *et al.* Phylogenetic Analysis of SARS-CoV-2 Data Is Difficult. *Molecular Biology and Evolution* (ed Malik, H.) msaa314 (2020). doi:10.1093/molbev/msaa314.

33. Turakhia, Y., De Maio, N., Thornlow, B., *et al.* Stability of SARS-CoV-2 Phylogenies. *PLOS Genetics* **16,** e1009175 (2020). doi:10.1371/journal.pgen.1009175.

34. Ignatieva, A., Hein, J. & Jenkins, P. A. Ongoing Recombination in SARS-CoV-2 Revealed through Genealogical Reconstruction. *Molecular Biology and Evolution* **39** (ed Thorne, J.) (2022). doi:10.1093/molbev/msac028.

35. Minh, B. Q., Schmidt, H. A., Chernomor, O., *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution* **37,** 1530–1534 (2020). doi:10.1093/molbev/msaa015.

36. Shu, Y. & McCauley, J. GISAID: Global Initiative on Sharing All Influenza Data – from Vision to Reality. *Eurosurveillance* **22** (2017). doi:10.2807/1560-7917.ES.2017.22.13.30494.

37. Chan, J. M., Carlsson, G. & Rabadan, R. Topology of Viral Evolution. *Proceedings of the National Academy of Sciences* **110,** 18566–18571 (2013). doi:10.1073/pnas.1313480110.

medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

38. Cámara, P. G., Levine, A. J. & Rabadán, R. Inference of Ancestral Recombination Graphs through Topological Data Analysis. *PLOS Computational Biology* **12** (ed Pond, S. L. K.) e1005071 (2016). doi:10.1371/journal.pcbi.1005071.

39. Rabadán, R. & Blumberg, A. J. *Topological Data Analysis for Genomics and Evolution* doi:10.1017/9781316671665 (Cambridge University Press, 2019).

40. Bauer, U. Ripser: efficient computation of Vietoris-Rips persistence barcodes. *Journal of Applied and Computational Topology* (2021). doi:10.1007/s41468-021-00071-5.

41. Keegan, L. & Kempf, D. *Hammingdist: A Fast Tool to Calculate Hamming Distances* version 0.19.0. 2022. https://github.com/ssciwr/hammingdist visited on 2022-07-01.

42. McBroome, J., Thornlow, B., Hinrichs, A. S., *et al.* A Daily-Updated Database and Tools for Comprehensive SARS-CoV-2 Mutation-Annotated Trees. *Molecular Biology and Evolution* **38** (ed Lu, J.) 5819–5824 (2021). doi:10.1093/molbev/msab264.

43. VanInsberghe, D., Neish, A. S., Lowen, A. C. & Koelle, K. Recombinant SARS-CoV-2 Genomes Circulated at Low Levels Over The First Year of The Pandemic. *Virus Evolution* (2021). doi:10.1093/ve/veab059.

44. Jackson, B., Boni, M. F., Bull, M. J., *et al.* Generation and transmission of interlineage recombinants in the SARS-CoV-2 pandemic. *Cell* **184,** 5179–5188.e8 (2021). doi:10.1016/j.cell.2021.08.014.

45. Shiraz, R. & Tripathi, S. Enhanced recombination among Omicron subvariants of SARS-CoV-2 contributes to viral immune escape. *Journal of Medical Virology* **95** (2023). doi:10.1002/jmv.28519.

46. Turakhia, Y., De Maio, N., Thornlow, B., *et al. Problematic Sites SARS-CoV-2* Github. https://github.com/W-L/ProblematicSites_SARS-CoV2 visited on 2021-07-25.

47. De Maio, N., Walker, C., Borges, R., *et al. Issues with SARS-CoV-2 Sequencing Data - SARS-CoV-2 Coronavirus / nCoV-2019 Genomic Epidemiology* Virological. https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473 visited on 2020-05-05.

48. Giorgio, S. D., Martignano, F., Torcia, M. G., *et al.* Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2. *Science Advances* **6** (2020). doi:10.1126/sciadv.abb5813.

49. Mourier, T., Sadykov, M., Carr, M. J., *et al.* Host-directed editing of the SARS-CoV-2 genome. *Biochemical and Biophysical Research Communications* **538,** 35–39 (2021). doi:10.1016/j.bbrc.2020.10.092.

50. Graudenzi, A., Maspero, D., Angaroni, F., *et al.* Mutational signatures and heterogeneous host response revealed via large-scale characterization of SARS-CoV-2 genomic diversity. *iScience* **24,** 102116 (2021). doi:10.1016/j.isci.2021.102116.

51. Alexandrov, L. B., Kim, J., Haradhvala, N. J., *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578,** 94–101 (2020). doi:10.1038/s41586-020-1943-3.

52. 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* **5** (ed Schwemmle, M.) (2020). doi:10.1128/msphere.00408-20.

53. Ratcliff, J. & Simmonds, P. Potential APOBEC-mediated RNA editing of the genomes of SARS-CoV-2 and other coronaviruses and its impact on their longer term evolution. *Virology* **556,** 62–72 (2021). doi:10.1016/j.virol.2020.12.018.

54. Zhao, L., Hall, M., de Cesare, M., *et al.* The mutational spectrum of SARS-CoV-2 genomic and antigenomic RNA. *Proceedings of the Royal Society B: Biological Sciences* **289** (2022). doi:10.1098/rspb.2022.1747.

medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

55. Goymer, P. Synonymous mutations break their silence. *Nature Reviews Genetics* **8,** 92–92 (2007). doi:10.1038/nrg2056.

56. Ramazzotti, D., Angaroni, F., Maspero, D., *et al.* Large-scale analysis of SARS-CoV-2 synonymous mutations reveals the adaptation to the human codon usage during the virus evolution. *Virus Evolution* **8** (2022). doi:10.1093/ve/veac026.

57. Mogro, E. G., Bottero, D. & Lozano, M. J. Analysis of SARS-CoV-2 synonymous codon usage evolution throughout the COVID-19 pandemic. *Virology* **568,** 56–71 (2022). doi:10.1016/j.virol.2022.01.011.

58. Wu, X., Shan, K.-j., Zan, F., *et al.* Optimization and Deoptimization of Codons in SARS-CoV-2 and Related Implications for Vaccine Development. *Advanced Science* (2023). doi:10.1002/advs.202205445.

59. Bai, H., Ata, G., Sun, Q., *et al.* Natural selection pressure exerted on "Silent" mutations during the evolution of SARS-CoV-2: Evidence from codon usage and RNA structure. *Virus Research* **323,** 198966 (2023). doi:10.1016/j.virusres.2022.198966.

60. Bloom, J. D. & Neher, R. A. *Distribution of fitness effects of mutations* Github. https://jbloomlab.github.io/SARS2-mut-fitness/gisaid_2023-03-29/effects_histogram.html visited on 2023-07-29.

61. Huang, Y., Yang, C., Xu, X.-f., *et al.* Structural and Functional Properties of SARS-CoV-2 Spike Protein: Potential Antivirus Drug Development for COVID-19. *Acta Pharmacologica Sinica* **41,** 1141–1149 (2020). doi:10.1038/s41401-020-0485-4.

62. Piccoli, L., Park, Y.-J., Tortorici, M. A., *et al.* Mapping Neutralizing and Immunodominant Sites on the SARS-CoV-2 Spike Receptor-Binding Domain by Structure-Guided High-Resolution Serology. *Cell* **183,** 1024–1042.e21 (2020). doi:10.1016/j.cell.2020.09.037.

63. Barnes, C. O., Jette, C. A., Abernathy, M. E., *et al.* SARS-CoV-2 neutralizing antibody structures inform therapeutic strategies. *Nature* **588,** 682–687 (2020). doi:10.1038/s41586-020-2852-1.

64. Dejnirattisai, W., Zhou, D., Ginn, H. M., *et al.* The antigenic anatomy of SARS-CoV-2 receptor binding domain. *Cell* **184,** 2183–2200.e22 (2021). doi:10.1016/j.cell.2021.02.032.

65. Chen, Y., Zhao, X., Zhou, H., *et al.* Broadly neutralizing antibodies to SARS-CoV-2 and other human coronaviruses. *Nature Reviews Immunology* **23,** 189–199 (2022). doi:10.1038/s41577-022-00784-3.

66. Johnson, B. A., Xie, X., Bailey, A. L., *et al.* Loss of furin cleavage site attenuates SARS-CoV-2 pathogenesis. *Nature* **591,** 293–299 (2021). doi:10.1038/s41586-021-03237-4.

67. Li, Q., Wu, J., Nie, J., *et al.* The Impact of Mutations in SARS-CoV-2 Spike on Viral Infectivity and Antigenicity. *Cell* **182,** 1284–1294.e9 (2020). doi:10.1016/j.cell.2020.07.012.

68. Plante, J. A., Liu, Y., Liu, J., *et al.* Spike Mutation D614G Alters SARS-CoV-2 Fitness. *Nature* **592,** 116–121 (2021). doi:10.1038/s41586-020-2895-3.

69. Hou, Y. J., Chiba, S., Halfmann, P., *et al.* SARS-CoV-2 D614G Variant Exhibits Efficient Replication Ex Vivo and Transmission in Vivo. *Science,* eabe8499 (2020). doi:10.1126/science.abe8499.

70. Yurkovetskiy, L., Wang, X., Pascal, K. E., *et al.* Structural and Functional Analysis of the D614G SARS-CoV-2 Spike Protein Variant. *Cell* **183,** 739–751.e8 (2020). doi:10.1016/j.cell.2020.09.032.

71. *Tracking SARS-CoV-2 Variants* World Health Organization. https://www.who.int/activities/tracking-SARS-CoV-2-variants visited on 2022-01-25.

medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

72. Singh, A., Steinkellner, G., Köchl, K., *et al.* Serine 477 Plays a Crucial Role in the Interaction of the SARS-CoV-2 Spike Protein with the Human Receptor ACE2. *Scientific Reports* **11,** 4320 (2021). doi:10.1038/s41598-021-83761-5.

73. Hodcroft, E. B., Zuber, M., Nadeau, S., *et al.* Spread of a SARS-CoV-2 Variant through Europe in the Summer of 2020. *Nature,* 1–9 (2021). doi:10.1038/s41586-021-03677-y.

74. Welkers, M. R. A., Han, A. X., Reusken, C. B. E. M. & Eggink, D. Possible Host-Adaptation of SARS-CoV-2 Due to Improved ACE2 Receptor Binding in Mink. *Virus Evolution* (2020). doi:10.1093/ve/veaa094.

75. Oude Munnink, B. B., Sikkema, R. S., Nieuwenhuijse, D. F., *et al.* Transmission of SARS-CoV-2 on Mink Farms between Humans and Mink and Back to Humans. *Science* **371,** 172–177 (2021). doi:10.1126/science.abe5901.

76. Van Dorp, L., Tan, C. C., Lam, S. D., *et al.* Recurrent Mutations in SARS-CoV-2 Genomes Isolated from Mink Point to Rapid Host-Adaptation. *bioRxiv* (2020). doi:10.1101/2020.11.16.384743.

77. Carabelli, A. M., Peacock, T. P., Thorne, L. G., *et al.* SARS-CoV-2 variant biology: immune escape, transmission and fitness. *Nature Reviews Microbiology* (2023). doi:10.1038/s41579-022-00841-7.

78. Peacock, T. P., Goldhill, D. H., Zhou, J., *et al.* The furin cleavage site in the SARS-CoV-2 spike protein is required for transmission in ferrets. *Nature Microbiology* **6,** 899–909 (2021). doi:10.1038/s41564-021-00908-w.

79. McCallum, M., De Marco, A., Lempp, F. A., *et al.* N-terminal domain antigenic mapping reveals a site of vulnerability for SARS-CoV-2. *Cell* **184,** 2332–2347.e16 (2021). doi:10.1016/j.cell.2021.03.028.

80. Liu, Z., VanBlargan, L. A., Bloyet, L.-M., *et al.* Identification of SARS-CoV-2 spike mutations that attenuate monoclonal and serum antibody neutralization. *Cell Host & Microbe* **29,** 477–488.e4 (2021). doi:10.1016/j.chom.2021.01.014.

81. Haynes, W. A., Kamath, K., Lucas, C., *et al.* Impact of B.1.1.7 Variant Mutations on Antibody Recognition of Linear SARS-CoV-2 Epitopes. *medRxiv* (2021). doi:10.1101/2021.01.06.20248960.

82. Lista, M. J., Winstone, H., Wilson, H. D., *et al.* The P681H mutation in the Spike glycoprotein confers Type I interferon resistance in the SARS-CoV-2 alpha (B.1.1.7) variant. *bioRxiv* (2021). doi:10.1101/2021.11.09.467693.

83. Thomson, E. C., Rosen, L. E., Shepherd, J. G., *et al.* Circulating SARS-CoV-2 spike N439K variants maintain fitness while evading antibody-mediated immunity. *Cell* **184,** 1171–1187.e20 (2021). doi:10.1016/j.cell.2021.01.037.

84. Liu, L., Iketani, S., Guo, Y., *et al.* Striking Antibody Evasion Manifested by the Omicron Variant of SARS-CoV-2. *bioRxiv* (2021). doi:10.1101/2021.12.14.472719.

85. Romero, P. E., Dávila-Barclay, A., Salvatierra, G., *et al.* The Emergence of Sars-CoV-2 Variant Lambda (C.37) in South America. *Microbiology Spectrum* **9** (ed Mostafa, H. H.) e00789–21 (2021). doi:10.1128/Spectrum.00789-21.

86. Kimura, I., Kosugi, Y., Wu, J., *et al.* The SARS-CoV-2 Lambda variant exhibits enhanced infectivity and immune resistance. *Cell Reports* **38,** 110218 (2022). doi:10.1016/j.celrep.2021.110218.

87. Viana, R., Moyo, S., Amoako, D. G., *et al.* Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa. *Nature* (2022). doi:10.1038/s41586-022-04411-y.

medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

88. Rambaut, A., Holmes, E. C., O'Toole, Á., *et al.* A Dynamic Nomenclature Proposal for SARS-CoV-2 Lineages to Assist Genomic Epidemiology. *Nature Microbiology* **5,** 1403–1407 (2020). doi:10.1038/s41564-020-0770-5.

89. *COVID-19 Weekly Epidemiological Update, Edition 76* World Health Organization. https://www.who.int/publications/m/item/weekly-epidemiological-update-on-covid-19---25-january-2022 visited on 2022-01-25.

90. Chen, C., Nadeau, S., Yared, M., *et al.* CoV-Spectrum: analysis of globally shared SARS-CoV-2 data to identify and characterize new variants. *Bioinformatics* **38** (ed Alkan, C.) 1735–1737 (2021). doi:10.1093/bioinformatics/btab856.

91. Chen, C., Nadeau, S., Yared, M., *et al.* *covSPECTRUM* Computational Evolution Group at ETH Zurich. https://cov-spectrum.org visited on 2023-07-25.

92. Willett, B. J., Grove, J., MacLean, O. A., *et al.* SARS-CoV-2 Omicron is an immune escape variant with an altered cell entry pathway. *Nature Microbiology* **7,** 1161–1179 (2022). doi:10.1038/s41564-022-01143-7.

93. Where did 'weird' Omicron come from? *Science* **374,** 1179 (2021). doi:10.1126/science.acx9754.

94. Callaway, E. Beyond Omicron: what's next for COVID's viral evolution. *Nature* **600,** 204–207 (2021). doi:10.1038/d41586-021-03619-8.

95. Kistler, K. E., Huddleston, J. & Bedford, T. Rapid and parallel adaptive mutations in spike S1 drive clade success in SARS-CoV-2. *Cell Host & Microbe* **30,** 545–555.e4 (2022). doi:10.1016/j.chom.2022.03.018.

96. Rodriguez-Rivas, J., Croce, G., Muscat, M. & Weigt, M. Epistatic models predict mutable sites in SARS-CoV-2 proteins and epitopes. *Proceedings of the National Academy of Sciences* **119** (2022). doi:10.1073/pnas.2113118119.

97. Rochman, N. D., Faure, G., Wolf, Y. I., *et al.* Epistasis at the SARS-CoV-2 Receptor-Binding Domain Interface and the Propitiously Boring Implications for Vaccine Escape. *mBio* **13** (ed Diamond, M. S.) (2022). doi:10.1128/mbio.00135-22.

98. Starr, T. N., Greaney, A. J., Hannon, W. W., *et al.* Shifting mutational constraints in the SARS-CoV-2 receptor-binding domain during viral evolution. *Science* **377,** 420–424 (2022). doi:10.1126/science.abo7896.

99. Starr, T. N., Greaney, A. J., Stewart, C. M., *et al.* Deep mutational scans for ACE2 binding, RBD expression, and antibody escape in the SARS-CoV-2 Omicron BA.1 and BA.2 receptor-binding domains. *PLOS Pathogens* **18** (ed Mok, C. K. P.) e1010951 (2022). doi:10.1371/journal.ppat.1010951.

100. Moulana, A., Dupic, T., Phillips, A. M., *et al.* Compensatory epistasis maintains ACE2 affinity in SARS-CoV-2 Omicron BA.1. *Nature Communications* **13** (2022). doi:10.1038/s41467-022-34506-z.

101. Neverov, A. D., Fedonin, G., Popova, A., *et al.* Coordinated evolution at amino acid sites of SARS-CoV-2 spike. *eLife* **12** (2023). doi:10.7554/elife.82516.

102. Jankowiak, M., Obermeyer, F. H. & Lemieux, J. E. Inferring selection effects in SARS-CoV-2 with Bayesian Viral Allele Selection. *PLOS Genetics* **18** (ed Gojobori, T.) e1010540 (2022). doi:10.1371/journal.pgen.1010540.

103. Lythgoe, K. A., Hall, M., Ferretti, L., *et al.* Shared SARS-CoV-2 Diversity Suggests Localised Transmission of Minority Variants. *bioRxiv* (2020). doi:10.1101/2020.05.28.118992.

104. Lythgoe, K. A., Hall, M., Ferretti, L., *et al.* SARS-CoV-2 within-host diversity and transmission. *Science* **372,** eabg0821 (2021). doi:10.1126/science.abg0821.

medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

105. Cohen-Steiner, D., Edelsbrunner, H. & Harer, J. Stability of Persistence Diagrams. *Discrete & Computational Geometry* **37,** 103–120 (2007). doi:10.1007/s00454-006-1276-5.

106. Edgar, R. C. MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput. *Nucleic Acids Research* **32,** 1792–1797 (2004). doi:10.1093/nar/gkh340.

107. Katoh, K. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* **30,** 3059–3066 (2002). doi:10.1093/nar/gkf436.

108. Bloom, J. D. & Neher, R. A. *Fitness effects of SARS-CoV-2 amino-acid mutations estimated from observed versus expected mutation counts* Github. https://github.com/jbloomlab/SARS2-mut-fitness/tree/main/results_gisaid_2023-03-29/aa_fitness visited on 2023-06-27.

109. COSMIC. *COSMIC Mutational Signatures* Wellcome Sanger Institute. https://cog.sanger.ac.uk/cosmic-signatures-production/documents/COSMIC_v3.3.1_SBS_GRCh37.txt visited on 2023-07-29.

110. Bleher, M., Hahn, L., Neumann, M., *et al.* MuRiT: efficient computation of pathwise persistence barcodes in multi-filtered flag complexes via Vietoris-Rips transformations. *arXiv* (2022). doi:10.48550/arXiv.2207.03394.

111. Gromov, M. in *Essays in Group Theory* (ed Gersten, S. M.) 75–263 (Springer, New York, NY, 1987). doi:10.1007/978-1-4613-9586-7_3.

112. Edelsbrunner, H. & Ölsböck, K. Holes and Dependences in an Ordered Complex. *Computer Aided Geometric Design* **73,** 1–15 (2019). doi:10.1016/j.cagd.2019.06.003.

113. Zomorodian, A. & Carlsson, G. Computing Persistent Homology. *Discrete & Computational Geometry* **33,** 249–274 (2005). doi:10.1007/s00454-004-1146-y.

114. Jariani, A., Warth, C., Deforche, K., *et al.* SANTA-SIM: Simulating Viral Sequence Evolution Dynamics under Selection and Recombination. *Virus Evolution* **5** (2019). doi:10.1093/ve/vez003.

115. Panjer, H. H. Recursive Evaluation of a Family of Compound Distributions. *ASTIN Bulletin* **12,** 22–26 (1981). doi:10.1017/S0515036100006796.

116. Sundt, B. & Jewell, W. S. Further Results on Recursive Evaluation of Compound Distributions. *ASTIN Bulletin: The Journal of the IAA* **12,** 27–39 (1981). doi:10.1017/S0515036100006802.

117. Maddison, W. P. & Maddison, D. *Mesquite: A Modular System for Evolutionary Analysis.* version 3.61. 2019. http://www.mesquiteproject.org visited on 2021-06-01.

118. Hanussek, M. *VALET* 2021. https://github.com/MaximilianHanussek/VALET visited on 2021-06-01.

119. Sanderson, T. & Kramer, A. *Cov2Tree* UCSC. https://cov2tree.org visited on 2023-07-10.

Acknowledgements

The authors 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 [18, 36], on which this research is based. An acknowledgement table can be retrieved from https://doi.org/10.55876/gis8.230731by. The authors acknowledge the use of de.NBI Cloud and the support by the High Performance and Cloud Computing Group at the Zentrum für Datenverarbeitung of the University of Tübingen and the German Federal Ministry of Education and Research (BMBF) through grant no 031 A535A. They thank M. Hanussek for IT support and early access to VALET [118]. The authors further acknowledge support

medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

from the Interdisciplinary Center for Scientific Computing at Heidelberg University and the development work of the Scientific Software Center of Heidelberg University carried out by L. Keegan and D. Kempf. This research was supported by the DFG Collaborative Research Center SFB/TRR 109 "Discretization in Geometry and Dynamics". M.B. was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy EXC 2181/1 - 390900948 (the Heidelberg STRUCTURES Excellence Cluster). L.H. thanks the Evangelisches Studienwerk Villigst for their support. M.N. was supported by the Vector Stiftung ("Topological Genomics"). A.O. acknowledges funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 281869850 (RTG 2229). This research was funded by the Federal Ministry of Education and Research (BMBF) and the Baden-Württemberg Ministry of Science as part of the Excellence Strategy of the German Federal and State Governments (KIT Centers, "Topological Genomics"), and by the Vector Stiftung ("Topological Genomics").

Author contributions

M.B., L.H., M.N., J.P.G., R.R., A.O. designed the study; M.B., L.H., A.O. curated data; M.B., M.C., L.H., A.O., J.P.G. performed computational analyses; U.B., M.B., L.H., M.N., A.O. developed and implemented software; M.B., L.H., A.O. acquired computing resources; M.B., L.H., A.O. drafted the manuscript; all authors contributed to the final version of the paper.

Competing interests

R.R. is a founder of Genotwin, he is member of the Scientific Advisory Board of AimedBio and consults for Arquimea Research. The other authors declare no competing interests.

Additional information

Supplementary Information is available for this paper.
Correspondence and requests for materials should be addressed to M.B., L.H. or A.O.

medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Supplementary Information

Supplementary Information Table 1. External spreadsheet containing full results of the topological recurrence analysis for the whole genome as of February 2021. The table lists mutations together with their topological recurrence index (tRI) and prevalence. All mutations with statistically significant tRI ≥ 2 are included. The table is augmented by filtration rules (columns "FILTER" and "INFO") taken from Turakhia *et al.* [33, 46] and De Maio *et al.* [47] for highly homoplasic sites that are normally removed or masked in SARS-CoV-2 phylogenetic analyses.

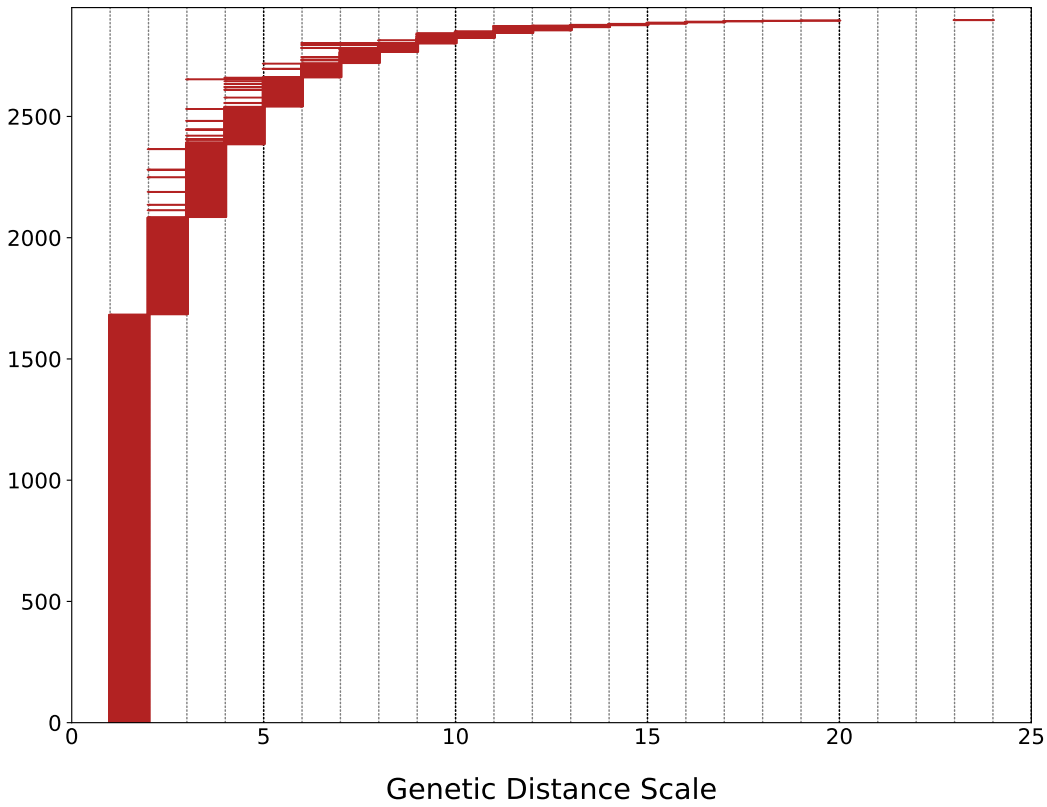
Supplementary Information Table 2. External spreadsheet containing full results of the topological recurrence analysis for the Spike gene as of February 2021. The table lists mutations together with their topological recurrence index (tRI) and prevalence. All mutations with tRI ≥ 2 are included, but only a tRI ≥ 8 is statistically significant. The table is augmented by filtration rules (columns "FILTER" and "INFO") taken from Turakhia *et al.* [33, 46] and De Maio *et al.* [47] for highly homoplasic sites that are normally removed or masked in SARS-CoV-2 phylogenetic analyses.

Supplementary Information Table 3. External spreadsheet containing a sublist of the list in Supplementary Information Table 2 featuring all mutations on the receptor-binding domain together with their topological recurrence index (tRI) and prevalence.

Supplementary Information Table 4. External spreadsheet containing full results of the topological recurrence analysis for the Spike gene as of January 2022. The table lists mutations together with their topological recurrence index (tRI). All mutations with tRI ≥ 2 are included, but only a tRI ≥ 76 is statistically significant. The table is augmented by filtration rules (columns "FILTER" and "INFO") taken from Turakhia *et al.* [33, 46] and De Maio *et al.* [47] for highly homoplasic sites that are normally removed or masked in SARS-CoV-2 phylogenetic analyses.

Supplementary Information Table 5. External spreadsheet containing full results of the topological recurrence analysis for the Spike gene as of March 2023. The table lists mutations together with their topological recurrence index (tRI). All mutations with tRI ≥ 2 are included, but only a tRI ≥ 75 is statistically significant. The table is augmented by filtration rules (columns "FILTER" and "INFO") taken from Turakhia *et al.* [33, 46] and De Maio *et al.* [47] for highly homoplasic sites that are normally removed or masked in SARS-CoV-2 phylogenetic analyses.

medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .



Supplementary Information Figure 1. Persistent homology of the GISAID dataset. Persistence barcode representing the persistent homology in dimension one of the GISAID dataset [18, 36] covering the first year of the pandemic from December 2019 until February 2021, comprising 161,024 genetically distinct high-quality SARS-CoV-2 whole genomes (see Methods). Each of the 2,899 bars in the barcode corresponds to a topological cycle in the reticulate phylogeny. The rich topology of the dataset indicates a multitude of reticulate events that shaped the evolution of the virus in the course of the pandemic. A total of 1,684 bars, which is 58% of all bars, concentrate at small genetic distance scales ≤ 2 and are therefore expected to be associated mainly with homoplasic events.

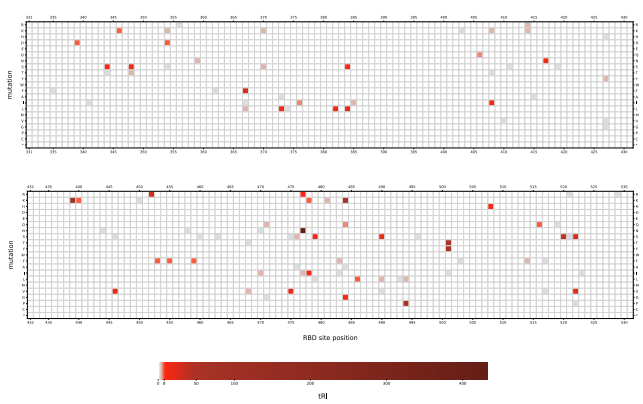
medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

A

SAAV	tRI		significant since	prevalence		notable lineages	ACE2-binding affinity[1]	mean plasma antibody escape[2]
S477N	433	↗	2020-07	4%	→	B.1.160, B.1.526*, B.1.1.529	0.06	0.011
N439K	88	↗	2020-04	2%	→		0.04	0.016
S494P	64	↗	2020-09	<1%	↗	B.1.1.7*	0	0.017
N501Y	55	↗	2020-09	19%	↗	B.1.1.7, P.1, B.1.351, B.1.1.529	0.24	0.011
N501T	50	↗	2020-10	<1%	↗		0.1	0.015
E484K	49	↗	2020-09	<1%	↗	B.1.1.7*, P.1, P.2, P.3, B.1.351, B.1.525, B.1.526*	0.06	0.066
A520S	35	↗	2020-05	<1%	→		-0.04	0.0098
L452R	28	↗	2020-12	1%	↗	B.1.427, B.1.429, B.1.617,	0.02	0.051
V367F	27	→	2020-03	<1%	→		0.07	0.015
A522S	27	↗	2020-04	<1%	→		-0.03	0.0099
P384L	23	↗	2020-04	<1%	→		0.01	0.022
A522V	21	→	2020-07	<1%	→		-0.03	0.010
F490S	19	↗	2020-12	<1%	↗	C.37	0	0.047
G446V	18	↗	2020-10	<1%	↗		-0.27	0.065
A475V	17	↗	2020-04	<1%	→		-0.14	0.021
A348S	15	↗	2020-10	<1%	↗		0.01	0.011
V382L	14	→	2020-10	<1%	→		-0.05	0.012
P479S	14	↗	2020-12	<1%	→		-0.03	0.0089
K417N	13	↗	2021-02	<1%	↗	B.1.351, B.1.671.2*, B.1.1.529*	-0.45	0.026
P384S	12	→	2020-12	<1%	→		-0.09	0.018
R408I	12	↗	2020-11	<1%	→		-0.09	–
T478I	12	→	2020-12	<1%	→		-0.04	0.0082
Y508H	12	↗	2020-07	<1%	→		0.07	0.017
S373L	11	→	2020-08	<1%	→		-0.02	0.011
E484G	10	↗	2021-01	<1%	↗		-0.06	0.065
A344S	9	→	2020-06	<1%	↘		-0.14	0.0078
S477R	9	↗	2021-02	<1%	↗		-0.03	0.0089
N354D	8	↗	2021-02	<1%	→		-0.04	0.024
Y453F	8	→	2020-06	<1%	↘	Mink, B.1.1.298	0.25	0.015
S459F	8	→	2021-01	<1%	→		-0.1	0.0073
F486L	8	→	2020-05	<1%	↘	Mink	-0.47	0.039
E516Q	8	→	2021-02	<1%	↗		-0.05	–
T478K	7	↗	–	<1%	↗	B.1.617.2, B.1.1.529	0.02	0.0088
E484Q	6	→	–	<1%	→	B.1.617.1, B.1.617.3	0.03	0.062

[1] as in Starr et al. [3] [2] as in Greaney et al. [4] * mutation found in some sequences but not all

B

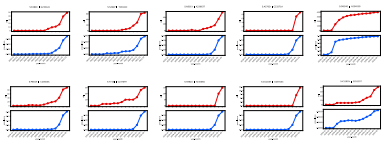


Supplementary Information Figure 2. Topological signals of amino acid changes on the receptor-binding domain during the first year of the pandemic (December 2019 until February 2021). (A) Table of all amino acid changes with statistically significant topological recurrence index (tRI ≥ 8), plus two more selected mutations. The table provides the tRI together with its tendency, the initial acquisition date of a significant tRI signal ($p < 0.05$), the prevalence together with its tendency, notable Pango lineages containing the mutation [88], ACE2-binding affinity as in Starr *et al.* [3, Table S2], and mean plasma antibody escape as in Greaney *et al.* [4, Table S3] (see Methods). Mutations with rising tRI signal and mean plasma antibody escape > 0.01 (shaded) potentially confer a fitness advantage to the virus and are therefore candidates that might appear in future new variants. (B) Heatmap of all amino acid variations across the RBD showing any topological signal of convergence. There is a distinct accumulation of signals in the receptor-binding motif, while other regions on the RBD, notably residues 390-435, show only few signals of convergence.

medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

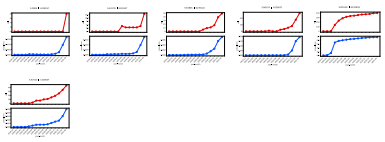
Alpha

Spike mutation	tRI
E484K*	49
S494P*	64
N501Y	55
A570D	9
D614G	400
P681H	92
T716I	30
S982A	15
D1118H	12
K1191N*	41



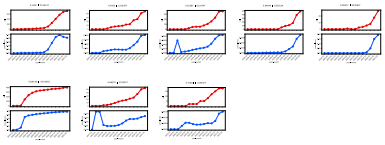
Beta

Spike mutation	tRI
D80A	0
D215G	4
K417N	13
E484K	49
N501Y	55
D614G	400
A701V	45



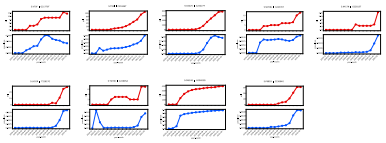
Gamma

Spike mutation	tRI
L18F	362
T20N	0
P26S	46
D138Y	71
R190S	0
K417T	0
E484K	49
N501Y	55
D614G	400
H655Y	54
T1027I	18



Delta

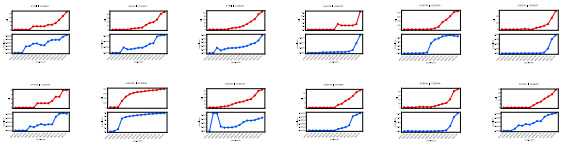
Spike mutation	tRI
T19R	0
V70F*	12
T95I	108
G142D	2
R158G	0
A222V*	776
W258L*	18
K417N*	13
L452R	28
T478K	7
D614G	400
P681R	31
D950N	6



Omicron

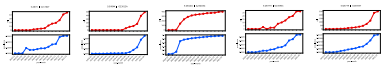
Spike mutation	tRI
T19I*	22
L24S*	4
A67V*	46
T95I*	108
G142D	2
N211I*	0
V213G*	0
G339D*	7
S371F*	0
S371L*	0
S373P*	0
S375F*	0
T376A*	0
S413R*	0
D405N*	0
R408S*	0
K417N*	13
N440K*	7
G446S*	2

Spike mutation	tRI
S477N	433
T478K	7
E484A	2
Q493R	0
G496S*	2
Q498R	0
N501Y	55
Y505H	0
T547K*	8
D614G	400
H655Y	54
N679K	18
P681H	92
N764K*	0
D796Y	24
N856K*	0
Q954H	0
N969K	0
L981F*	0



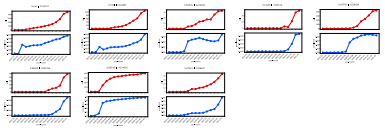
Eta

Spike mutation	tRI
A67V	46
E484K	49
D614G	400
Q677H	27
F888L	0



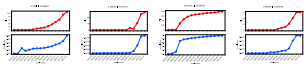
Iota

Spike mutation	tRI
L5F	466
D80G*	3
T95I	108
D253G	33
L452R*	28
S477N*	433
E484K	49
D614G	400
A701V	45
T859N*	5
D950H*	0
Q957R*	2



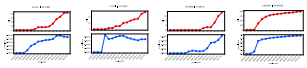
Kappa

Spike mutation	tRI
T95I*	108
G142D	2
E154K	2
L452R	28
E484Q	6
D614G	400
P681R	31
Q1071H	3



Lambda

Spike mutation	tRI
G75V	43
T76I	28
L452Q	0
F490S	19
D614G	400
T859N	5



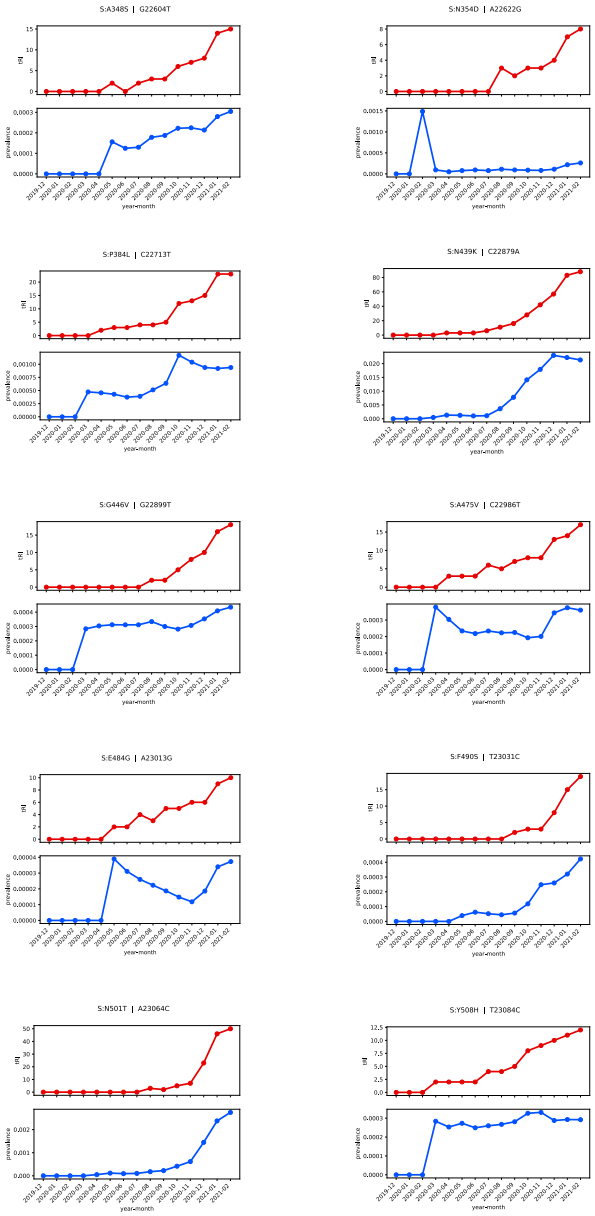
boldface highlights significant tRI

* mutation found in some sequences but not all

Supplementary Information Figure 3. Topological signals of Spike mutations in Variants of Interest/Concern. Comparative time series analysis charts (tRI vs. prevalence) as of February 2021 for amino acid changes on the Spike gene seen in notable lineages that have been designated as VOIs/VOCs [71]. Significant tRI values are highlighted in boldface.

medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Spike mutation	tRI
A348S	15
N354D	8
P384L	23
N439K	88
G446V	18
A475V	17
E484G	10
F490S	19
N501T	50
Y508H	12



Supplementary Information Figure 4. Surveillance of emerging escape mutations on the receptor-binding domain. Comparative time series analysis charts (tRI vs. prevalence) for mutations on the receptor-binding domain that showed a significant tRI signal with rising tendency in February 2021 and are associated with an increased mean plasma antibody escape > 0.01 [4]. These mutations had low prevalence $< 5\%$ and were not seen in any VOI/VOC as of February 2021 (see Supplementary Information Figure 2), but mutations at corresponding residues are likely to confer a fitness advantage to the virus and might therefore appear in future variants. In fact, the immune escape mutations `S:G446S` [4, 84] and `S:F490S` [4, 80] later appeared in the Lambda [85, 86] and Omicron [87] variants, which were designated as VOI/VOC in June/November 2021 [71].

medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

A Alpha variant (B.1.1.7) Spike gene mutations — Date reported: November 2020

amino acid site	tRI per time bucket														
	2019-12	2020-01	2020-02	2020-03	2020-04	2020-05	2020-06	2020-07	2020-08	2020-09	2020-10	2020-11	2020-12	2021-01	2021-02
N501	0	0	0	0	0	0	2	0	3	7	**12**	**18**	**42**	**84**	**110**
A570	0	0	0	0	0	0	0	0	0	2	4	5	11	18	21
D614	0	0	**3**	**139**	**232**	**286**	**317**	**334**	**343**	**354**	**367**	**372**	**378**	**395**	**402**
P681	0	0	0	0	**3**	**7**	**7**	**8**	**15**	**26**	**33**	**45**	**74**	**128**	**138**
T716	0	0	0	**3**	**3**	3	4	4	6	**10**	10	10	14	**26**	**30**
S982	0	0	0	0	0	0	0	0	0	0	0	0	0	10	15
D1118	0	0	0	0	0	2	**5**	5	5	5	8	**15**	**17**	**27**	**33**

B Beta variant (B.1.351) Spike gene mutations — Date reported: July 2020

amino acid site	tRI per time bucket														
	2019-12	2020-01	2020-02	2020-03	2020-04	2020-05	2020-06	2020-07	2020-08	2020-09	2020-10	2020-11	2020-12	2021-01	2021-02
L18	0	0	0	**4**	**5**	**9**	**14**	**17**	**25**	**53**	**125**	**207**	**291**	**354**	**371**
D80	0	0	0	0	2	**5**	**5**	**12**	**24**	**36**	**43**	**53**	**64**	**88**	**97**
D215	0	0	0	0	0	**6**	**11**	**13**	**14**	**17**	**31**	**60**	**71**	**82**	**82**
R246	0	0	0	0	0	0	0	0	0	0	0	0	2	6	6
K417	0	0	0	0	0	0	0	0	4	3	3	3	3	4	13
E484	0	0	0	0	0	2	2	4	**7**	**15**	**20**	**23**	**29**	**59**	**70**
N501	0	0	0	0	0	0	2	0	3	7	**12**	**18**	**42**	**84**	**110**
D614	0	0	**3**	**139**	**232**	**286**	**317**	**334**	**343**	**354**	**367**	**372**	**378**	**395**	**402**
A701	0	0	0	0	0	2	**11**	**11**	**15**	**16**	**23**	**28**	**36**	**45**	**57**

C Gamma variant (P.1) Spike gene mutations — Date reported: January 2021

amino acid site	tRI per time bucket														
	2019-12	2020-01	2020-02	2020-03	2020-04	2020-05	2020-06	2020-07	2020-08	2020-09	2020-10	2020-11	2020-12	2021-01	2021-02
L18	0	0	0	**4**	**5**	**9**	**14**	**17**	**25**	**53**	**125**	**207**	**291**	**354**	**371**
T20	0	0	0	**2**	**4**	**6**	**12**	**19**	**25**	**33**	**41**	**50**	**62**	**92**	**103**
P26	0	0	0	0	**5**	**9**	**12**	**17**	**20**	**26**	**31**	**50**	**56**	**86**	**93**
D138	0	0	0	0	2	**6**	**11**	**17**	**18**	**25**	**35**	**48**	**68**	**107**	**119**
R190	0	0	0	0	0	0	0	2	3	3	6	7	11	14	15
K417	0	0	0	0	0	0	0	0	4	3	3	3	3	4	13
E484	0	0	0	0	0	2	2	4	**7**	**15**	**20**	**23**	**29**	**59**	**70**
N501	0	0	0	0	0	0	2	0	3	7	**12**	**18**	**42**	**84**	**110**
D614	0	0	**3**	**139**	**232**	**286**	**317**	**334**	**343**	**354**	**367**	**372**	**378**	**395**	**402**
H655	0	0	0	**2**	**3**	**5**	**9**	**13**	**16**	**18**	**22**	**25**	**36**	**50**	**54**
T1027	0	0	0	0	0	2	2	2	5	5	8	12	15	18	18
V1176	0	0	0	**7**	**14**	**20**	**37**	**52**	**63**	**71**	**81**	**103**	**110**	**125**	**133**

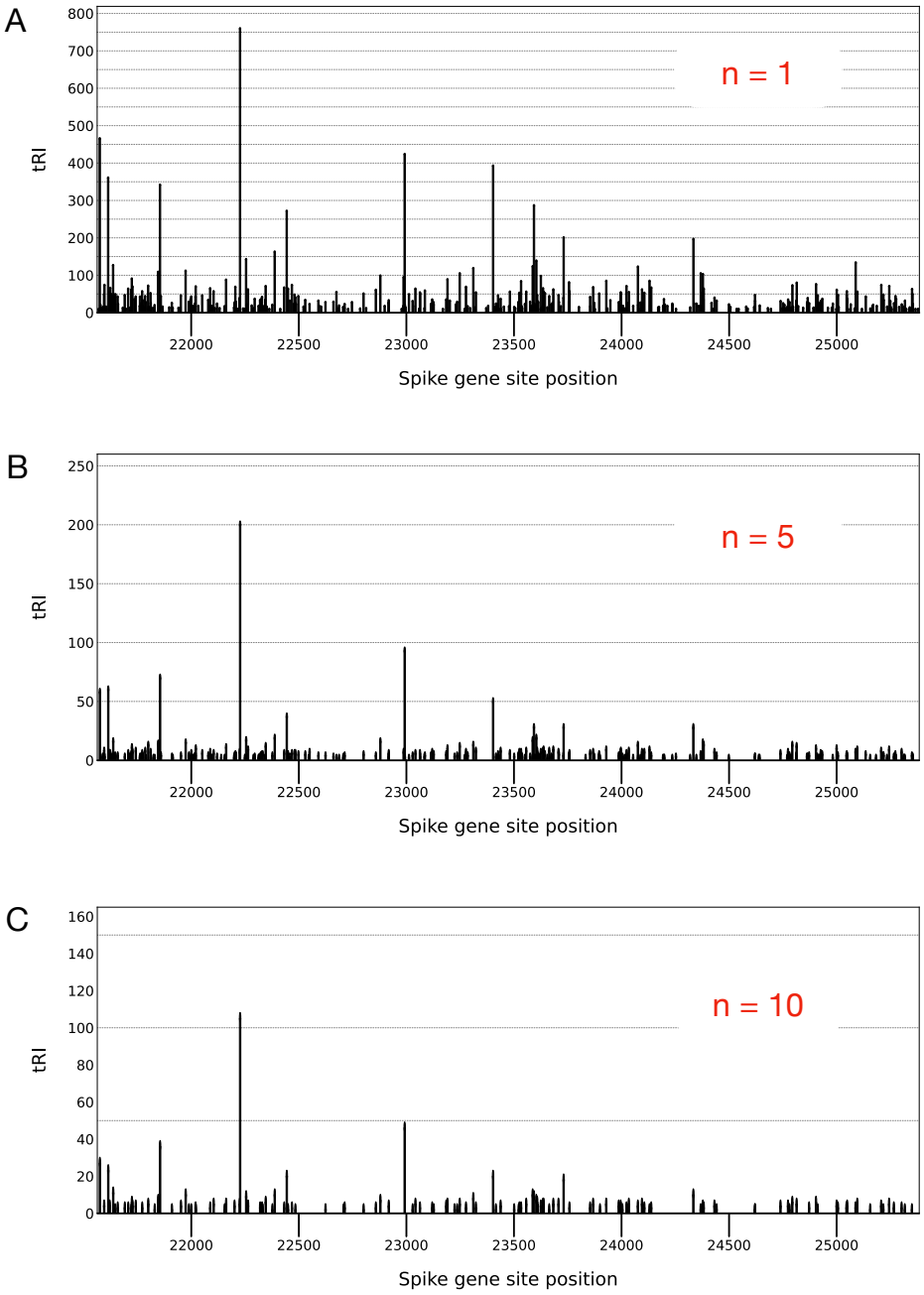
D Delta variant (B.1.617.2) Spike gene mutations — Date reported: December 2020

amino acid site	tRI per time bucket														
	2019-12	2020-01	2020-02	2020-03	2020-04	2020-05	2020-06	2020-07	2020-08	2020-09	2020-10	2020-11	2020-12	2021-01	2021-02
T19	0	0	0	0	0	**4**	4	4	4	6	6	8	14	21	**26**
G142	0	0	0	0	**3**	3	**5**	5	6	8	**12**	**14**	13	19	21
R158	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
L452	0	0	0	0	2	3	3	**6**	**8**	**16**	**21**	**27**	**39**	**61**	**65**
T478	0	0	0	0	0	2	3	3	3	5	5	6	9	19	21
D614	0	0	**3**	**139**	**232**	**286**	**317**	**334**	**343**	**354**	**367**	**372**	**378**	**395**	**402**
P681	0	0	0	0	**3**	**7**	**7**	**8**	**15**	**26**	**33**	**45**	**74**	**128**	**138**
D950	0	0	0	0	0	0	0	0	0	0	0	4	7	9	12

boldface highlights significant tRI

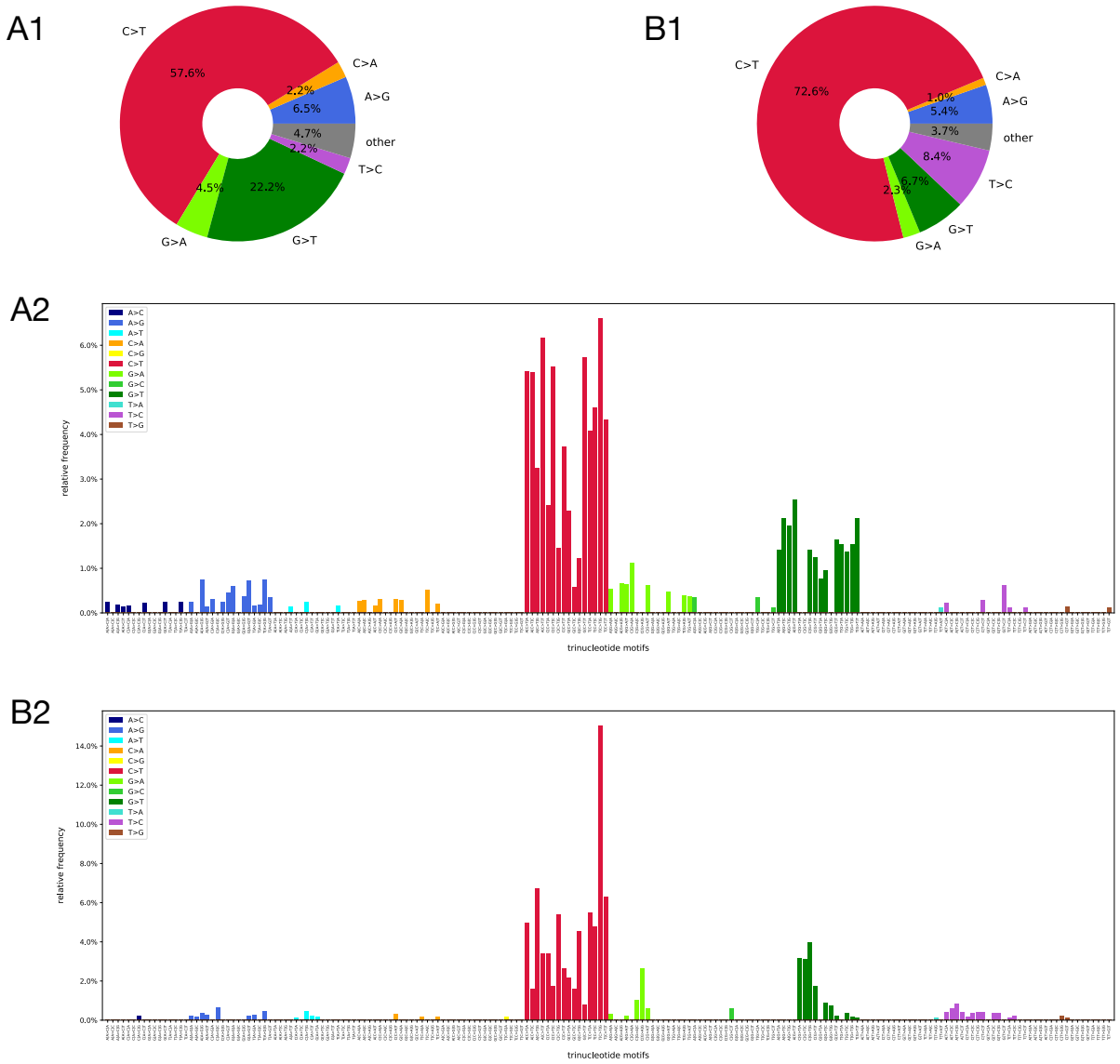
Supplementary Information Figure 5. Topological signals over time for Spike gene amino acid changes in the VOCs Alpha, Beta, Gamma and Delta. For each time bucket ranging from December 2019 until February 2021, the tables display the topological recurrence index (tRI) for all defining variable Spike gene amino acid sites in the given VOC. Significant tRI values are highlighted in boldface. For all four VOCs, a substantial number of defining variable amino acid sites showed a significant topological signal already months before the variant was first reported.

medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .



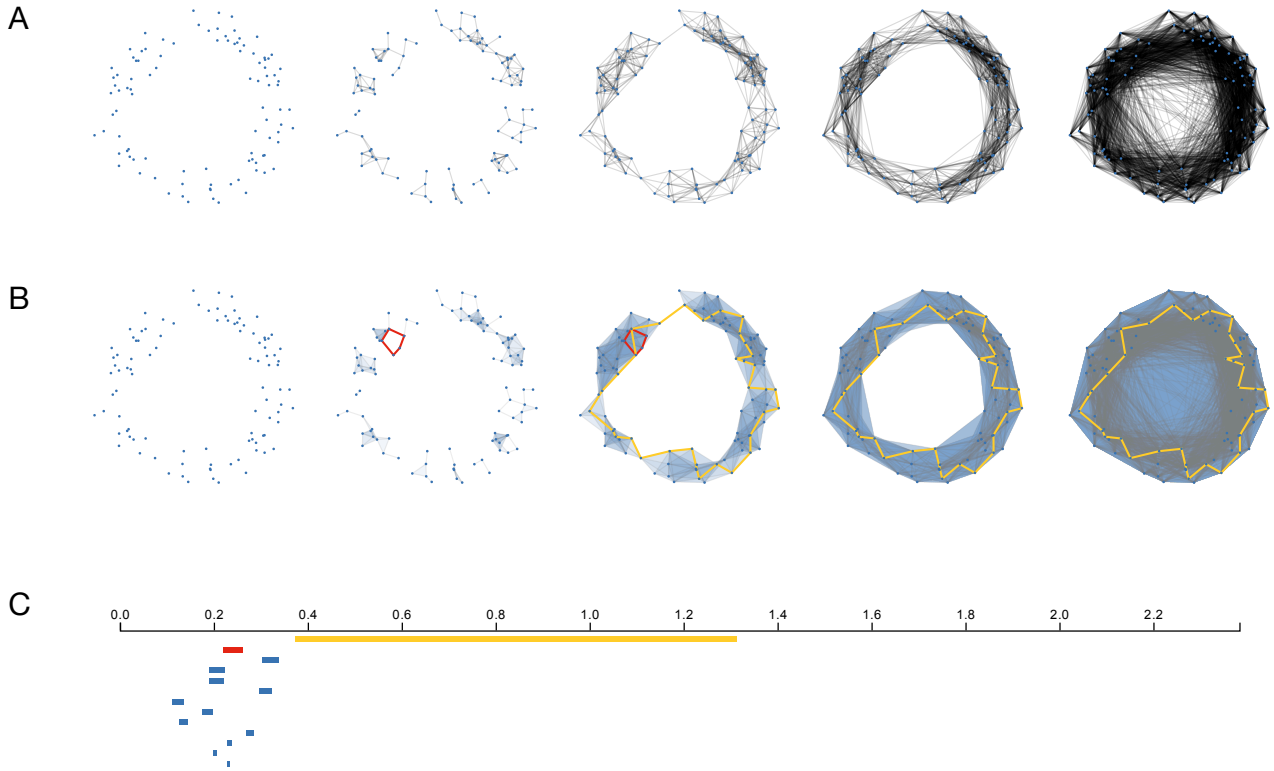
Supplementary Information Figure 6. Change of topological signals under additional filtration by multiple identical genomes. The panels display topological signals along the Spike gene as of February 2021 for different filtration parameters n. Sequence alignments were filtered by collecting only those sequences that appear in at least n identical copies in the alignment. (A) The tRI landscape for $n = 1$, based on the whole alignment with no additional filtration as it is used in the analyses in this study. (B) and (C) display the tRI landscape after additional filtration with parameters $n = 5$ and $n = 10$. The qualitative structure of the tRI landscape does not change: while the overall strength of tRI signals diminishes, characteristic and significant signals persist.

medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .



Supplementary Information Figure 7. Mutational patterns in topological signals of recurrence. Frequency analysis of nucleotide substitutions across the whole genome showing significant topological signal during the first year of the pandemic (December 2019 until February 2021). The panels display total frequencies of single nucleotide variations for non-synonymous mutations (A1) and synonymous mutations (B1). There is an excess in C-to-T transitions which suggests that the topological recurrence index also captures the action of APOBEC-mediated host RNA editing processes. Relative frequencies of trinucleotide motifs centered at single nucleotide variations for non-synonymous variations (A2) and synonymous variations (B2) (see Methods). Statistical analyses based on Supplementary Information Table 1.

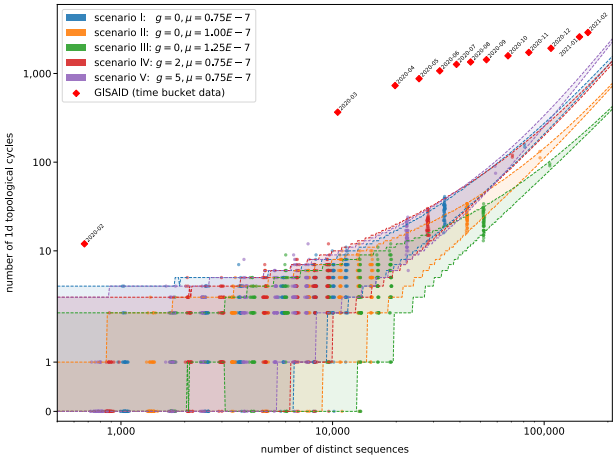
medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .



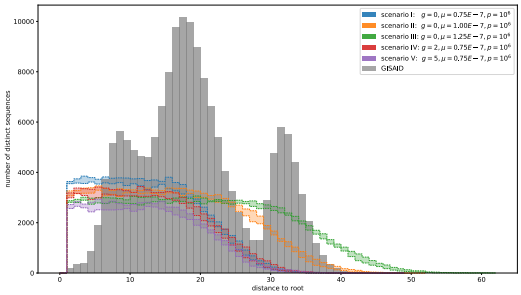
Supplementary Information Figure 8. Vietoris–Rips filtration of a point cloud. Each point represents a sample, and we display the geometric graphs (A), the resulting Vietoris–Rips complexes at different scales (B) and the persistence barcode in dimension one (C). If one only chooses one scale, one might either see nothing, or detect the small red cycle but miss the large yellow one, or vice versa. A solution to handle this issue is to characterize each cycle with its scale of appearance and disappearance: the red cycle induces a red bar in the barcode, and similarly for the yellow cycle.

medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

A Count of topological cycles in simulated data vs. GISAID data



B Genetic distance from reference sequence in simulated data vs. GISAID data



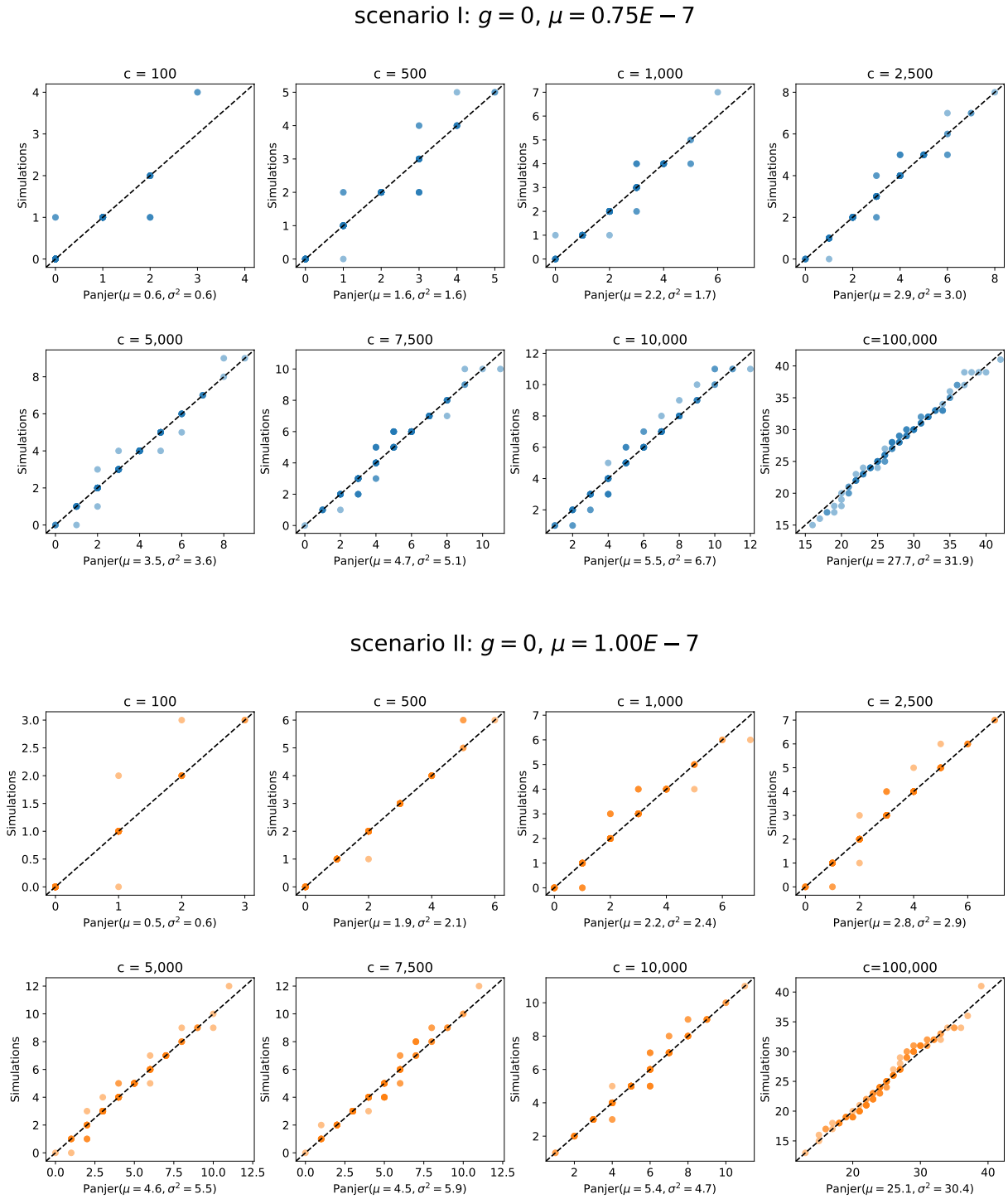
C Parameters for simulated data

scenario	mutation rate μ [10^{-7} subst./(site×gen.)]	growth rate g [1/gen.]	initial population p / carrying capacity c	expected no. cycles (95% PI) sample size 161,000 sequences
I	0.75	0	$p = c =$ 1e2, 5e2, 1e3, 2.5e3, 5e3, 7.5e3, 1e4, 1e5, 1e6	[692, 814]
II	1	0		[362, 408]
III	1.25	0		[211, 237]
IV	0.75	2	$p = 100$ / $c =$ 1e2, 5e2, 1e3, 2.5e3, 5e3, 7.5e3, 1e4, 1e5, 1e6	[646, 727]
V	0.75	5		[1023, 1171]

Supplementary Information Figure 9. The number of topological features in the GISAID whole genome alignment covering the first year of the pandemic (December 2019 until February 2021) is statistically significant. (A) Simulations were generated with `SANTA-SIM` [114] for five distinct scenarios with varying growth rate g and mutation rate μ. The 95% prediction intervals for the number of one-dimensional cycles in each scenario are based on the extrapolation of a Panjer distribution for an increasing number of distinct sequences in the simulated phylogenies (see Methods). For each scenario, the validation dataset shown in the plot is well-described by the corresponding prediction intervals. The simulations demonstrate that for each time bucket, the number of topological cycles observed in GISAID data is significantly larger than the expected number of topological cycles that are due to noise (i.e. random topological cycles in simulated data). As of February 2021, scenario V suggests that less than 40% (1,171) of all cycles (2,899) in the GISAID alignment are due to noise. (B) Comparison of genetic distances to the root in simulated data vs. distances to the Wuhan/Hu-1 reference sequence `EPI_ISL_402125` in the GISAID dataset. Scenarios I, IV and V with low mutation rate systematically underestimate the maximal distance, while the highest mutation rate in scenario III yields larger distances. The mutation rate of scenario II describes the maximal distance and overall diversity well. Differences to the GISAID data are expected to be due to real-world effects like variation of population growth, belated up-take in sequencing efforts, and enhanced spread of certain variants. (C) Parameters and prediction intervals in scenarios I-V. Scenarios I-III vary over a range of mutation rates that roughly capture the diversity of the GISAID dataset. Scenarios IV and V probe the influence of logistic population growth. For all scenarios we produced 100 simulations for each of the values of the carrying population $c \leq 10^5$, and five simulations for $c = 10^6$.

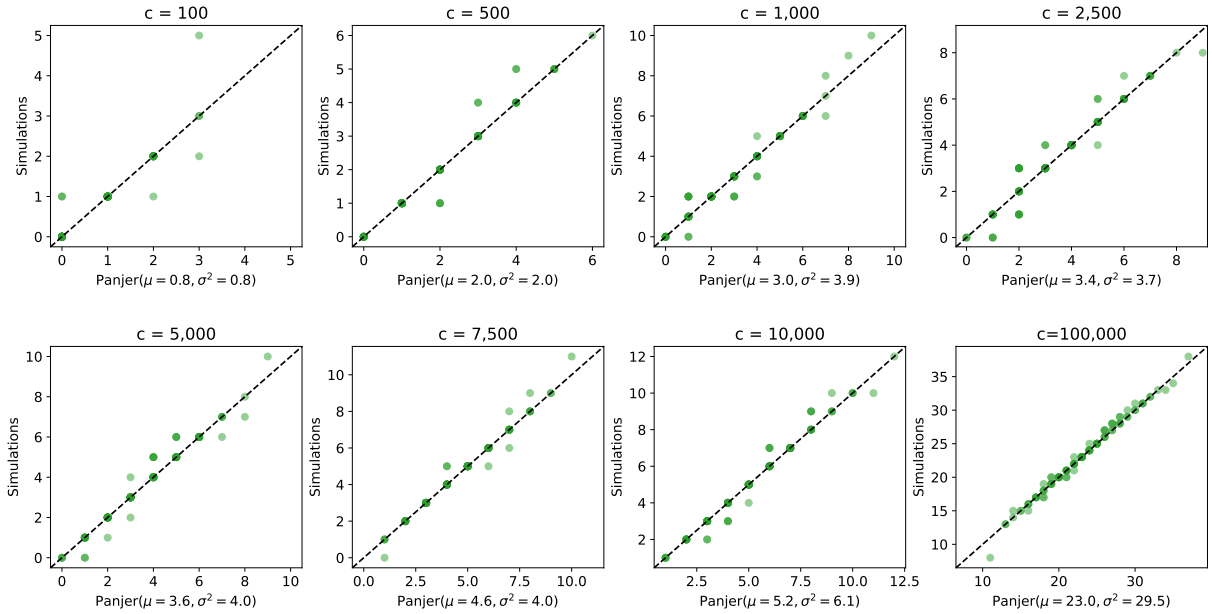
medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Supplementary Information Figure 10. Quantile-quantile analysis of Panjer distribution versus observed number of one-dimensional cycles in simulated phylogenies. For each value of carrying population c we determined the mean and variation of the observations and used these as parameters for the Panjer distribution.

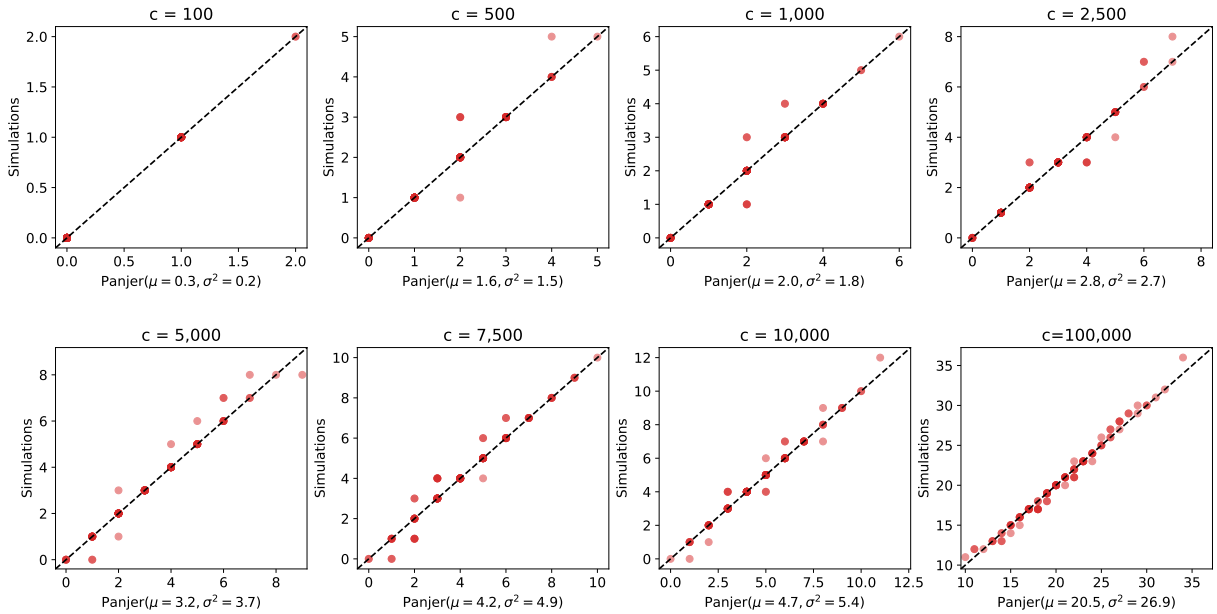


medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

scenario III: $g = 0$, $\mu = 1.25E - 7$

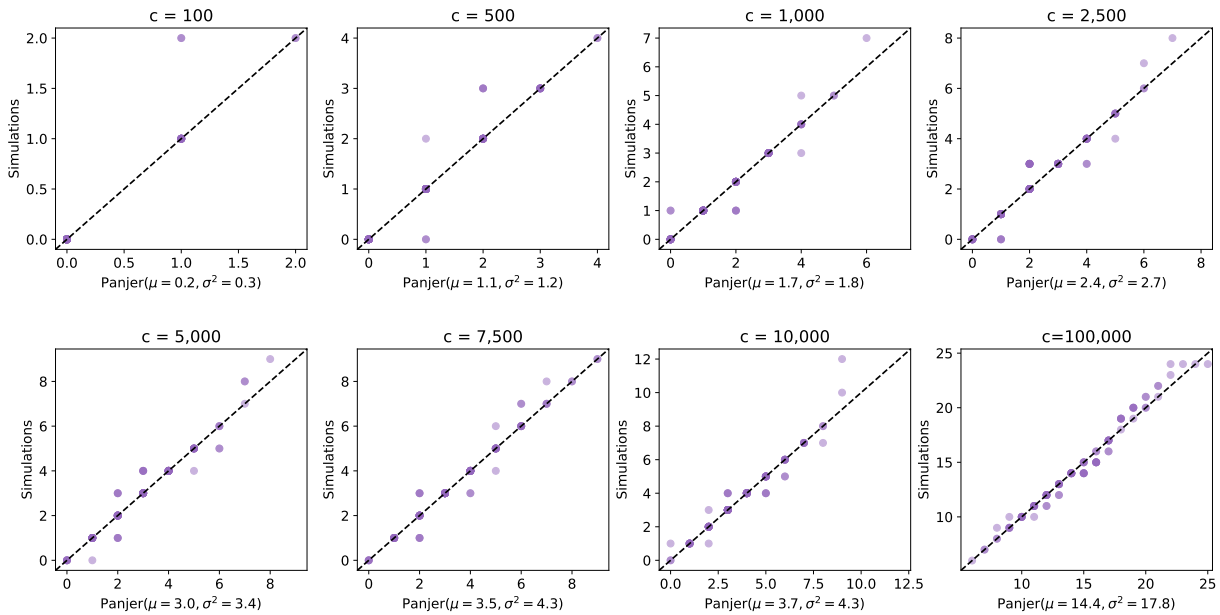


scenario IV: $g = 2$, $\mu = 0.75E - 7$



medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

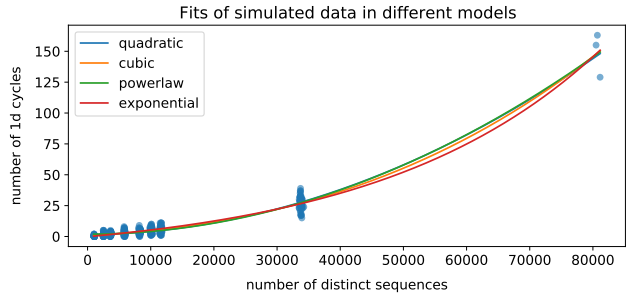
scenario V: $g = 5$, $\mu = 0.75E - 7$



medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

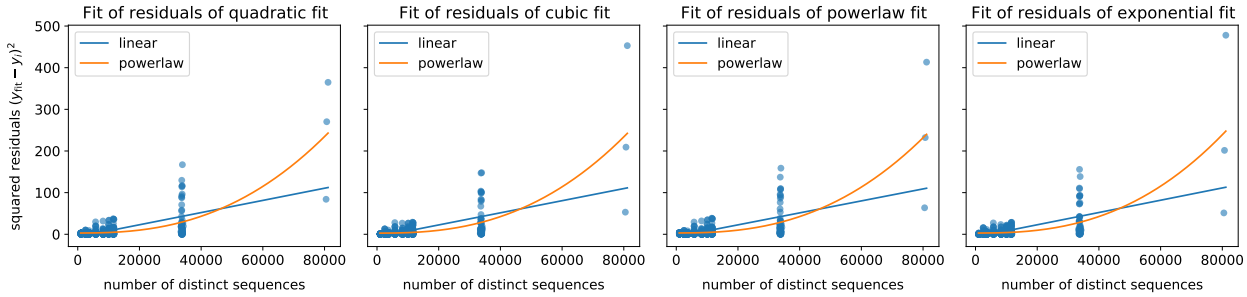
Supplementary Information Figure 11. Analysis of models that extrapolate the simulated data. For each scenario we fit quadratic, cubic, powerlaw, and exponential models to the observed number of one-dimensional cycles in simulations. Then we fit a linear and powerlaw model to the corresponding residuals as an estimate for the variance of the data. The quality of each model is evaluated through the log-likelihood to observe the validation dataset given a certain model.

scenario I: $g = 0, \mu = 0.75E-7$

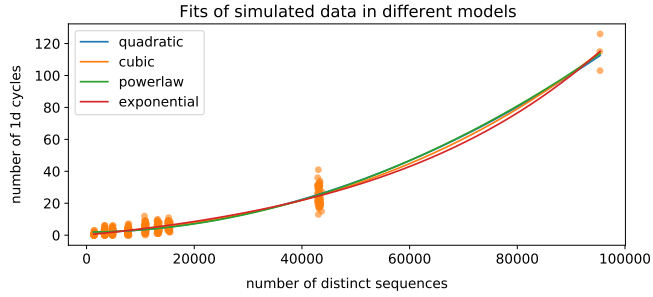


Log-Likelihood of validation data

	linear	powerlaw
quadratic	-1082	-665
cubic	-1299	-663
powerlaw	-1132	-669
exponential	-1062	-707

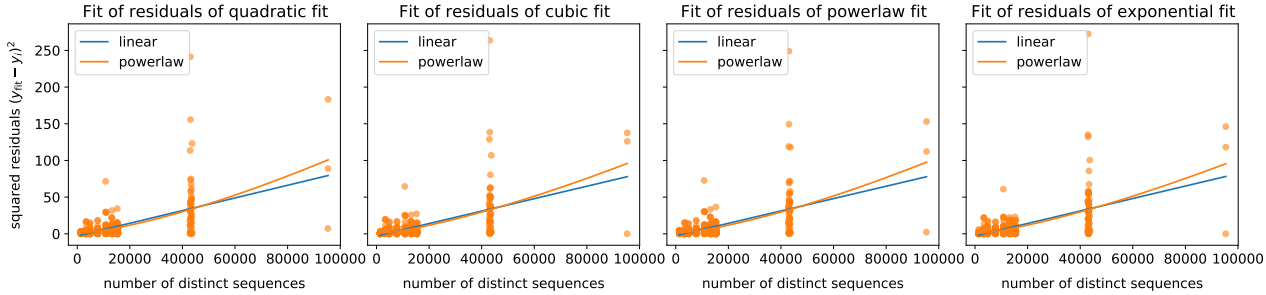


scenario II: $g = 0, \mu = 1.00E-7$



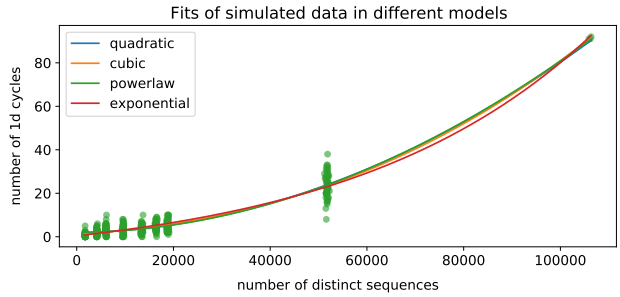
Log-Likelihood of validation data

	linear	powerlaw
quadratic	-1023	-685
cubic	-831	-693
powerlaw	-1002	-703
exponential	-902	-653



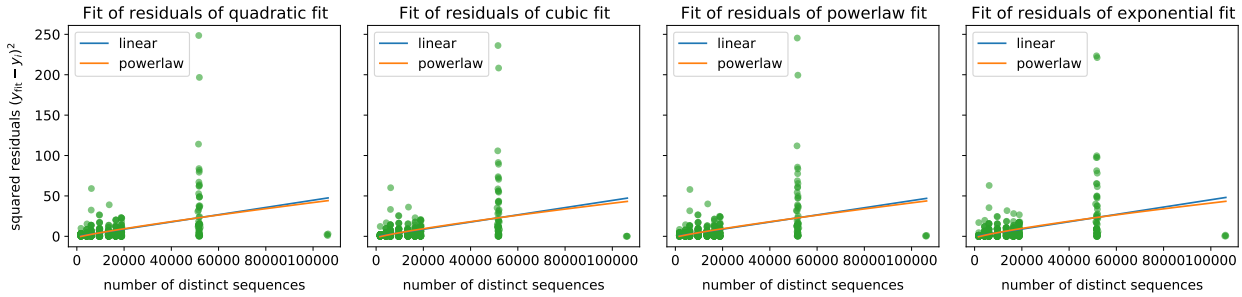
medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

scenario III: $g = 0, \mu = 1.25E - 7$

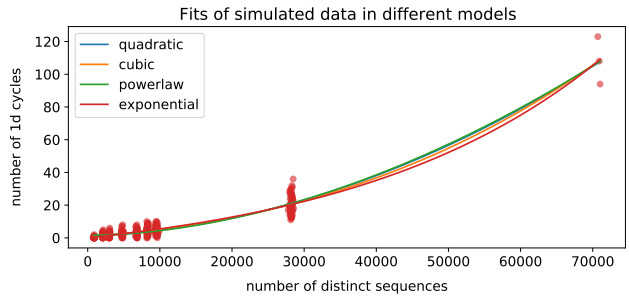


Log-Likelihood of validation data

	linear	powerlaw
quadratic	-713	-797
cubic	-701	-714
powerlaw	-728	-730
exponential	-689	-741

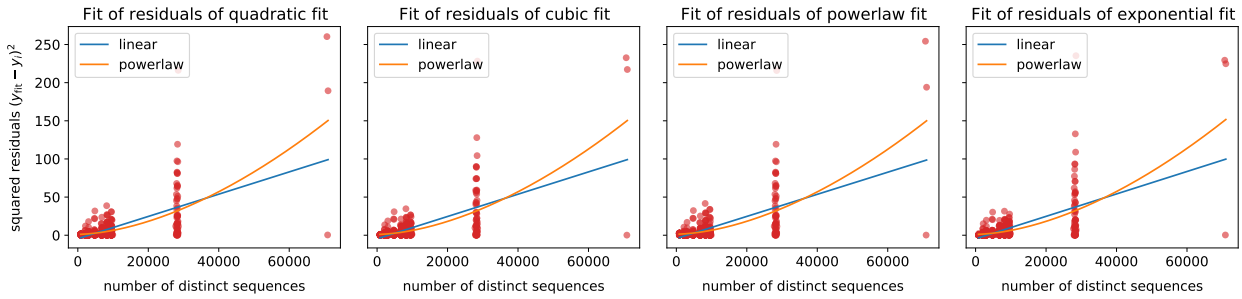


scenario IV: $g = 2, \mu = 0.75E - 7$



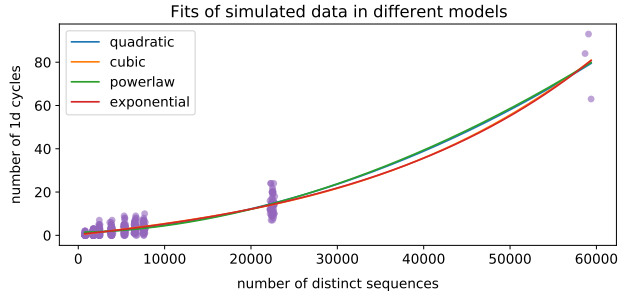
Log-Likelihood of validation data

	linear	powerlaw
quadratic	-1340	-638
cubic	-981	-625
powerlaw	-1178	-656
exponential	-1018	-626



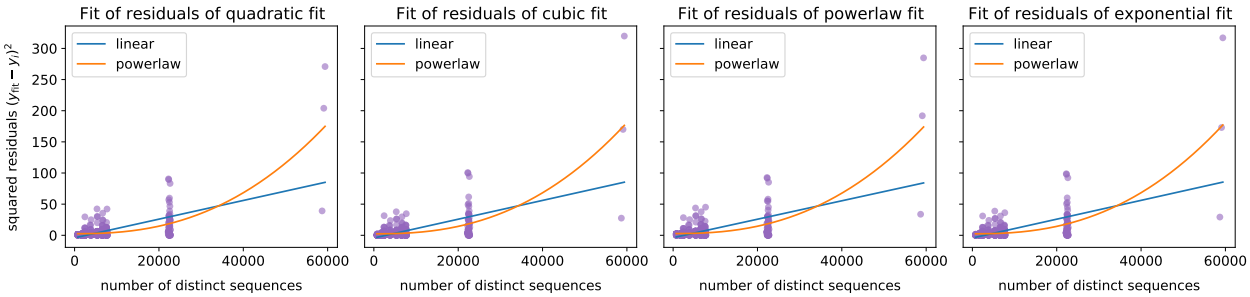
medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

scenario V: $g = 5, \mu = 0.75E - 7$



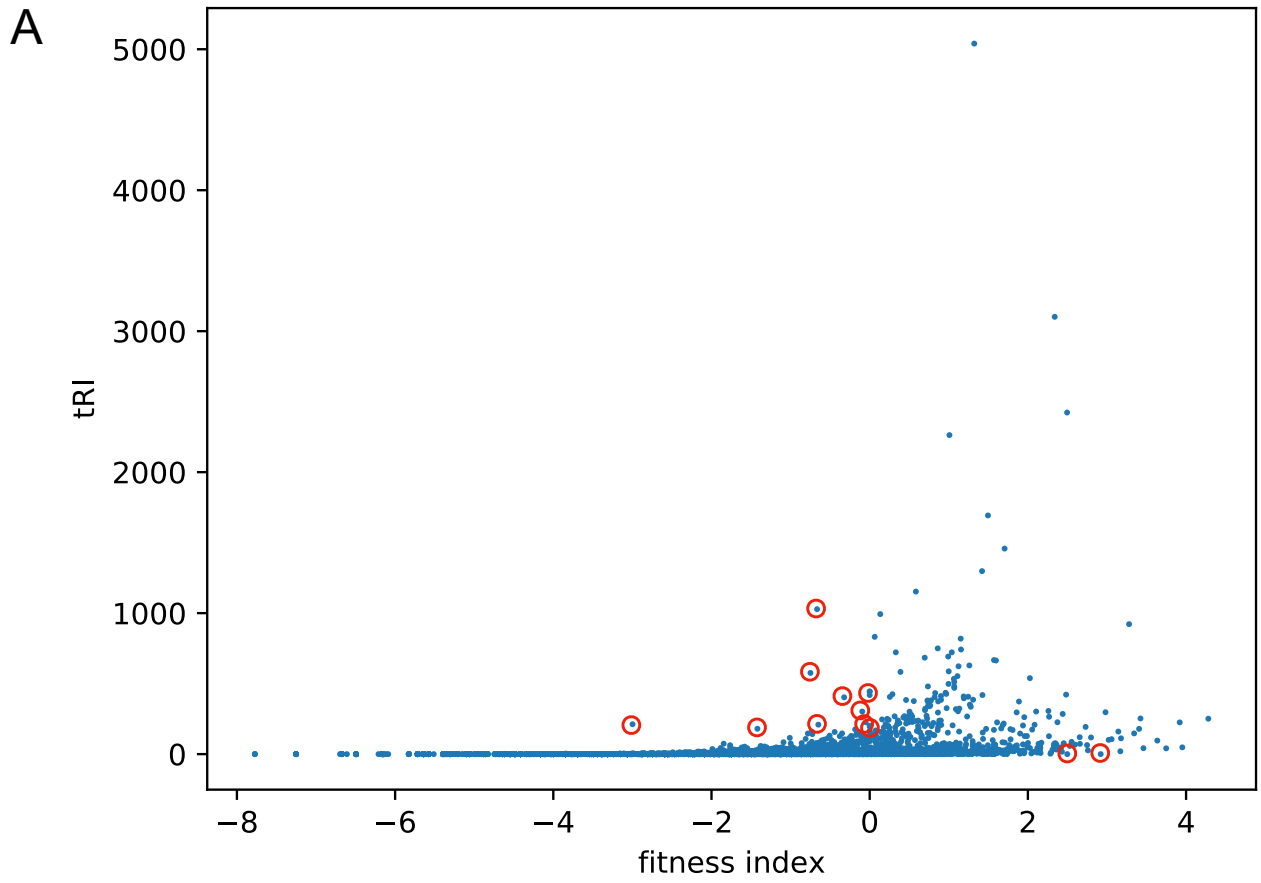
Log-Likelihood of validation data

	linear	**powerlaw**
quadratic	-1088	-602
cubic	-875	-593
powerlaw	-1198	-615
exponential	-909	-593



medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Supplementary Information Figure 12. Correlation between Topological Data Analysis-based and phylogeny-based signals of convergent evolution. Topological recurrence index (tRI) vs. fitness index [7] for Spike gene amino acid changes as of March 2023. (A) We found a strong correlation between positive tRI and positive fitness index (Fisher's exact test, $p < 10e{-}10$) with a Spearman correlation coefficient of 0.63 ($p < 10e{-}99$). Significant outliers, for which tRI and fitness index show divergent tendencies, are marked with red circles and are discussed in panel (B). The topological recurrence analysis is based on GISAID data [18, 36] covering the whole pandemic from December 2019 until March 2023, comprising 13,766,674 genetically distinct high-quality SARS-CoV-2 Spike genes (see Supplementary Information Table 5 and Methods). The data on fitness effects (fitness index) is taken from Bloom & Neher [108]. The statistical correlation analysis was performed for Spike gene amino acid changes listed in the file `aa_fitness.csv` (see Methods). (B) Assessment of the potential adaptiveness for mutations marked with a red circle in (A) for which tRI and fitness index show significantly divergent tendencies. Among these, all mutations flagged with significant tRI ≥ 75 show signs of adaptation such as frequent occurrence in independent lineages, occurrence in Variants of Concern or experimental evidence. The assessment is based on phylogenetic data taken from Cov2Tree [119] and epidemiological data taken from covSPECTRUM [91], as well as experimental results from the literature.



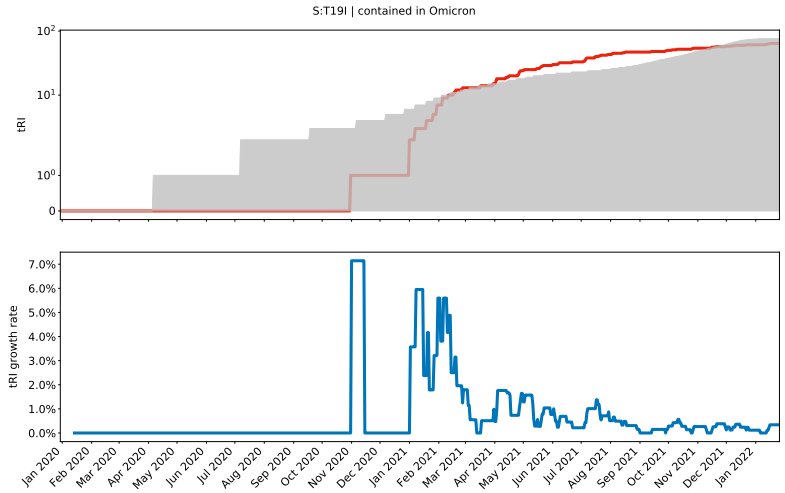
medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

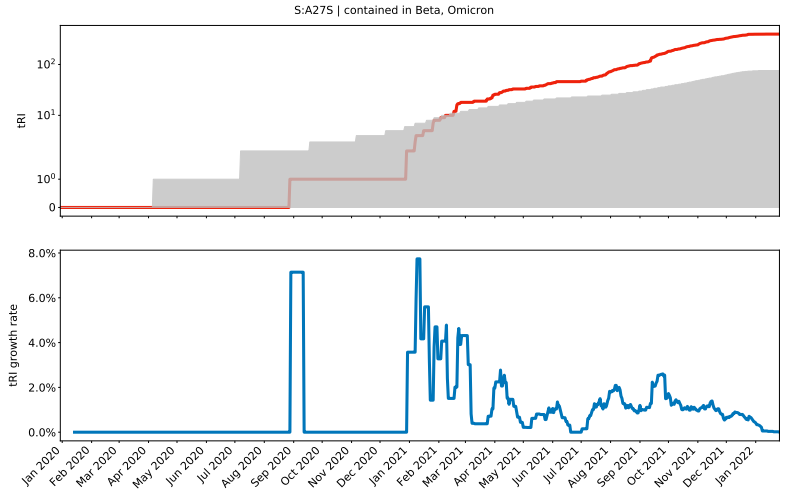
B

SAAV	tRI	fitness index	number of occurrences in independent lineages with at least one descendant (cov2tree.org as of 19 July 2023)	assessment of potential adaptiveness
S112L	1028	-0.66541	119	found in AY.25 (Delta) variant; in NTD region of Spike gene
P251L	576	-0.74907	118	found in AY.98.1 (Delta) and B.1.1.249 variants; in NTD region of Spike gene
T19I	444	0.0	243	found in BA.2 and BA.5 (Omicron) variants
L452R	444	0.0	257	found in AY.4 (Delta) variant; experimental evidence for increased infectivity and viral replication, see https://doi.org/10.1016/j.chom.2021.06.006
D614G	419	0.0	40	found in all major variants; experimental evidence for increased infectivity, see https://doi.org/10.1016/j.cell.2020.06.043
T19R	403	-0.32301	208	found in AY.4 (Delta) variant; in NTD region of Spike gene
D1259Y	302	-0.093773	148	found in AY.4.5 (Delta) variant
G1167V	226	-0.049625	146	found in B.1.617.2 and AY.41 (Delta) variants
S371F	212	-2.9983	156	found in Omicron variant; in RBD region of Spike gene; experimental evidence for antibody escape, see https://doi.org/10.1038/s41586-022-04594-4
T719I	209	-0.64916	121	found in AY.42.1 (Delta) variant
S1252F	201	-7.7807E-05	255	found in AY.4 (Delta) and BA.1 and BA.2 (Omicron) variants
T478K	195	0.0	137	found in AY.4 (Delta) and BA.2 and BA.5 (Omicron) variants; experimental evidence for slight antibody escape, see https://doi.org/10.1016/j.chom.2021.02.003
L1141F	180	-1.4197	48	found in AY.4 (Delta) variant; flagged as fitness increasing by PyR0, see https://doi.org/10.1126/science.abm1208
L110V	0	2.9208	19	not found in any notable variant; global prevalence < 1%; in NTD region of Spike gene
Q1054E	0	2.4978	27	not found in any notable variant; global prevalence < 1%

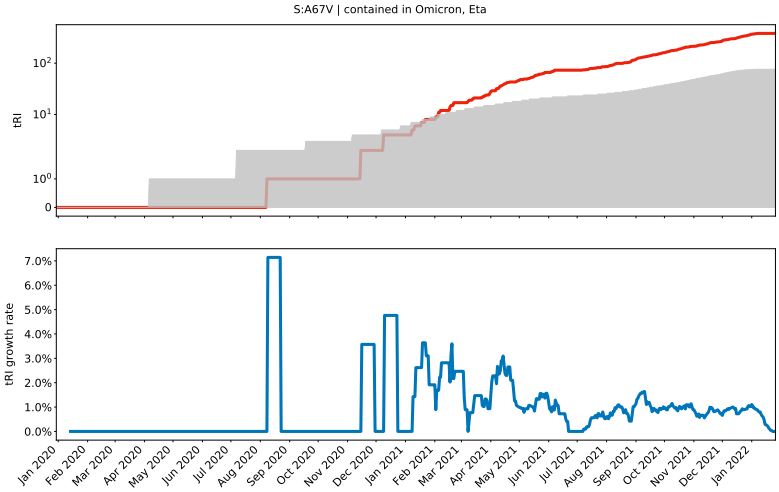
medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

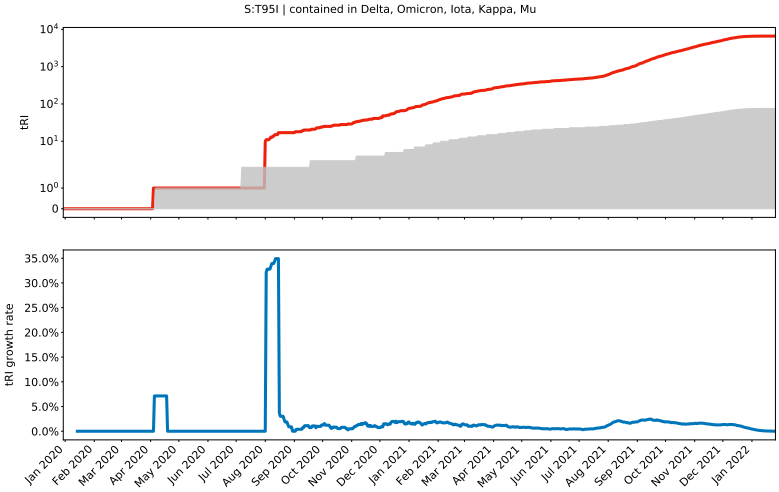
Supplementary Information Figure 13. Longitudinal analysis of topological signals of adaptation for the Omicron variant. Time series analysis charts for Omicron Spike gene amino acid changes with positive topological signal listed in panel (C) of Figure 9. Each chart shows the topological recurrence index (tRI, in red) and the tRI growth rate (14 days moving average, in blue) from January 2020 until January 2022 at daily resolution, where in the upper diagram the shaded region marks the level of significance (see Methods). The tRI growth rate is tampered by computational artifacts at the time when the tRI first attains a positive value, and towards the end of the covered period in January 2022 when the sampling rate in the underlying sequence alignment may still be low.

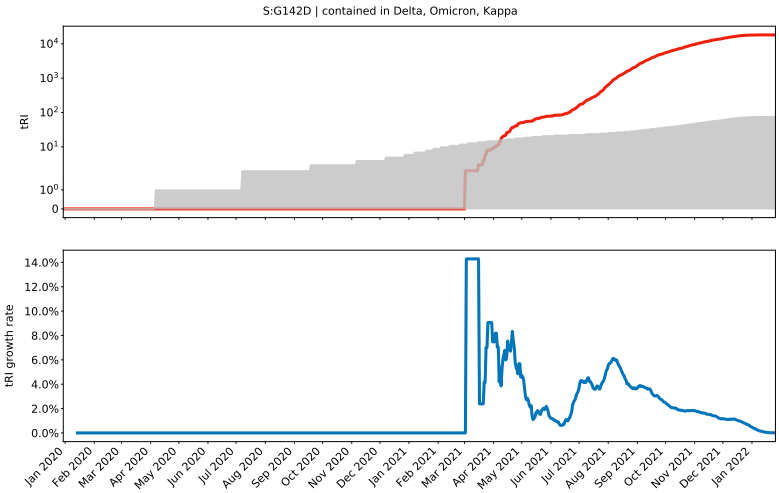




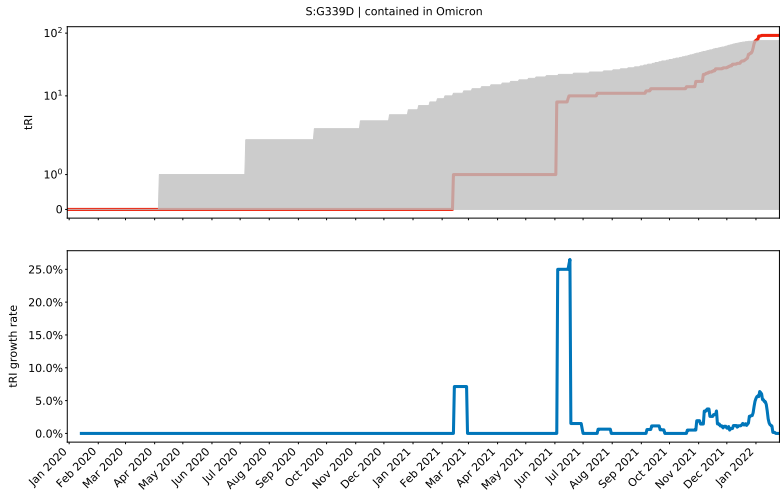
medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

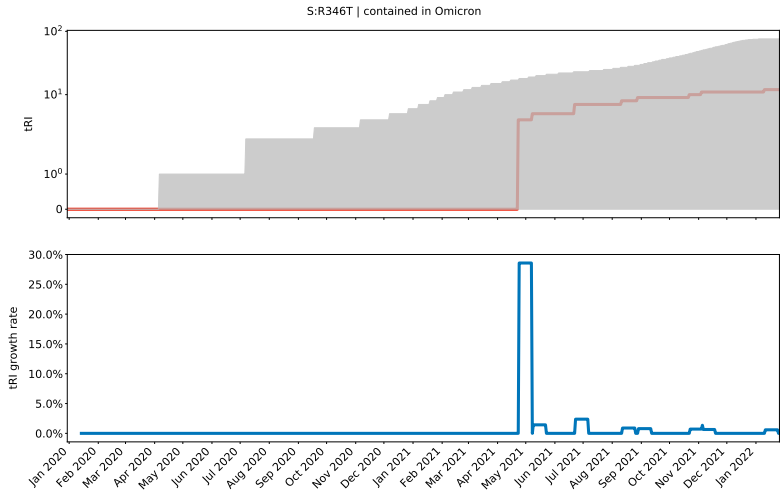


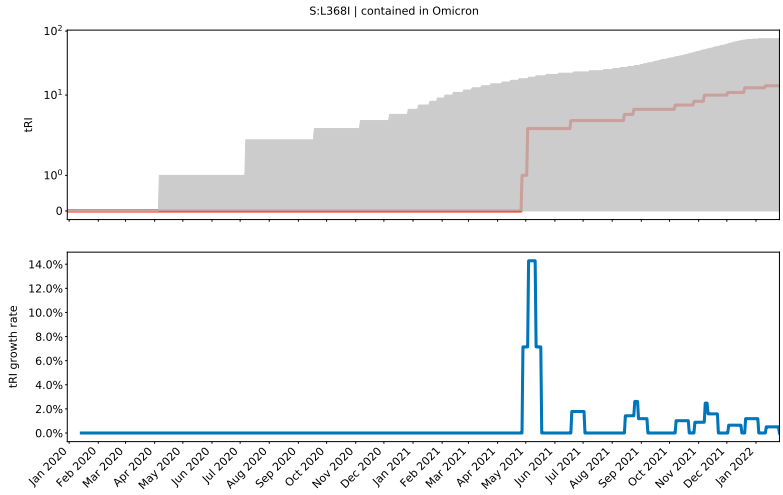




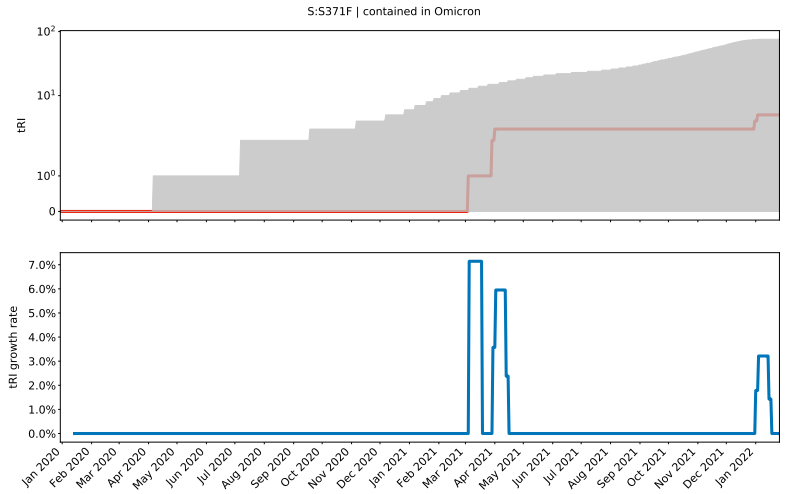
medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

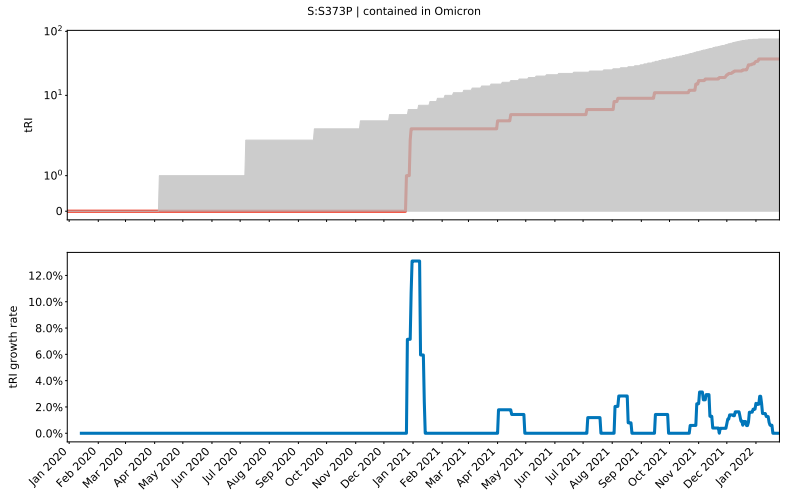


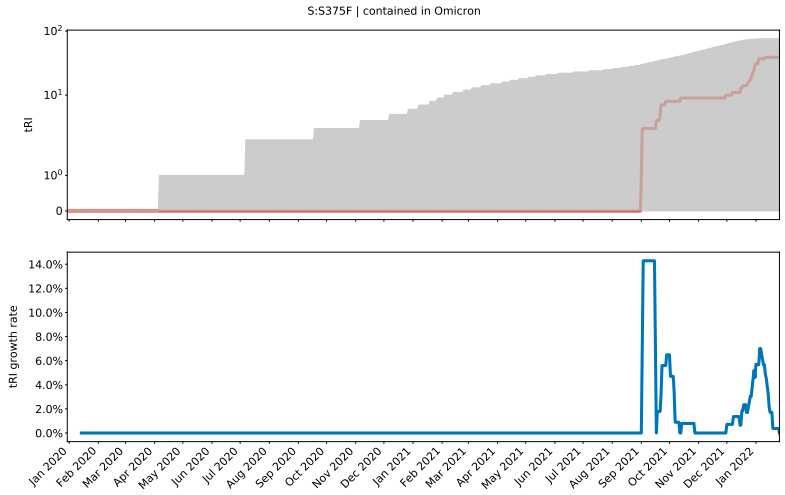




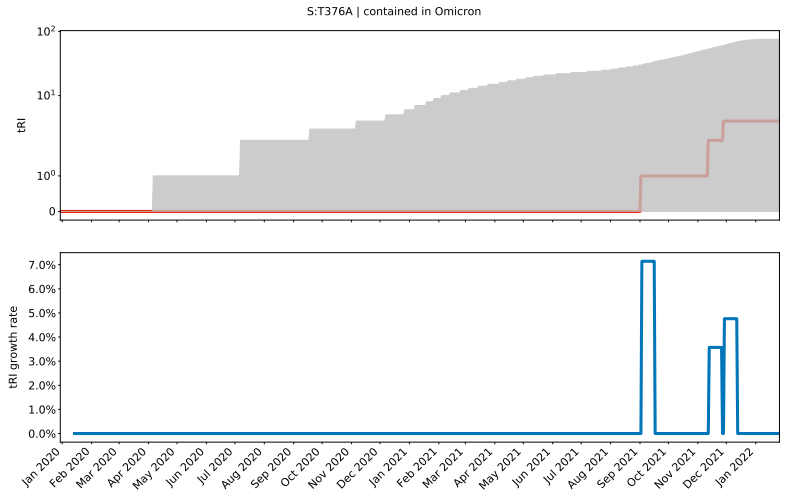
medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

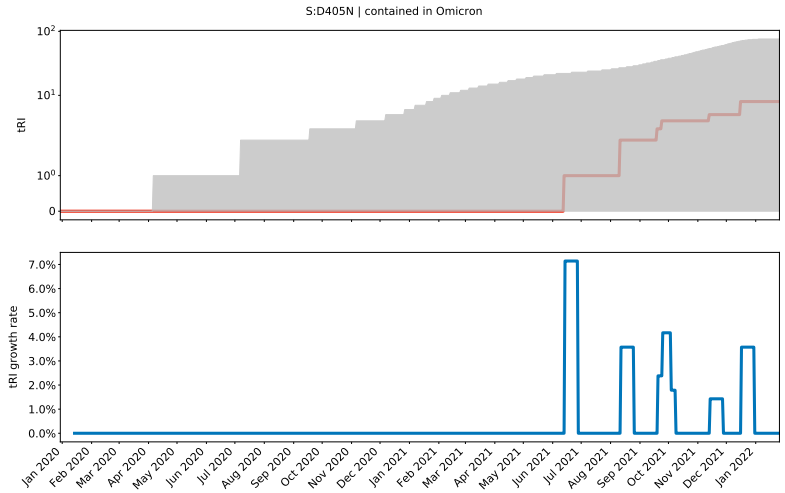


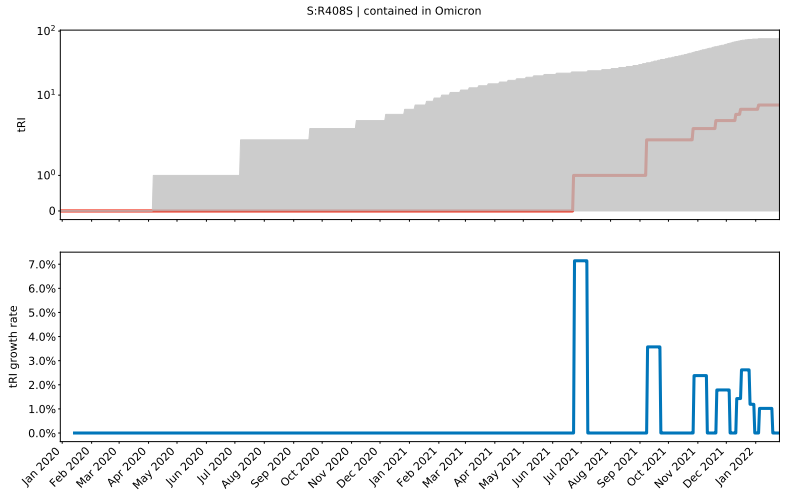




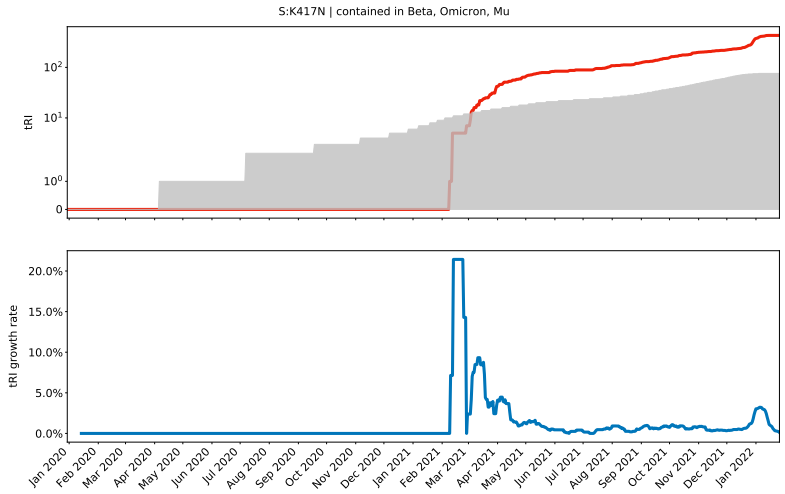
medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

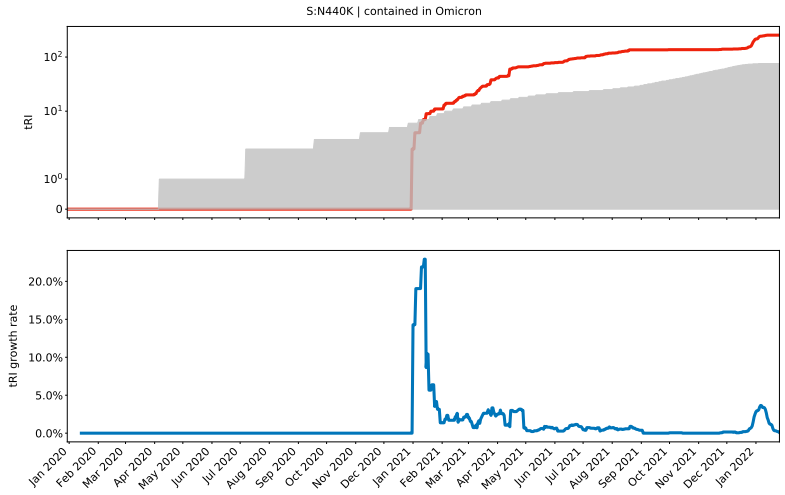


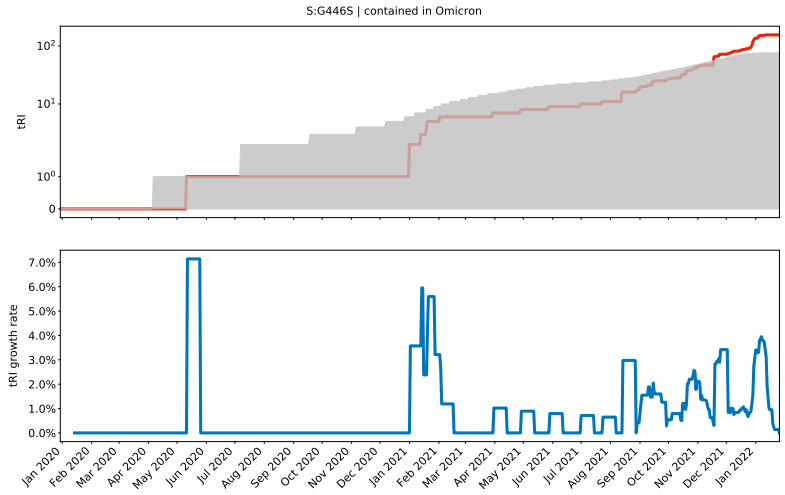




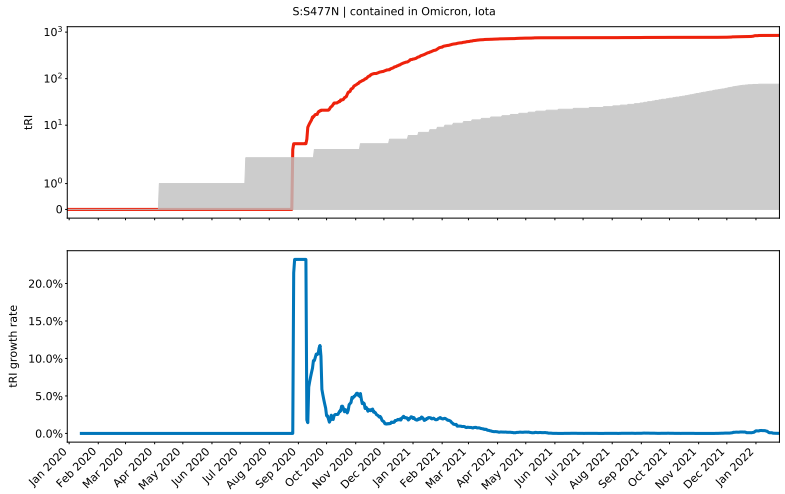
medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

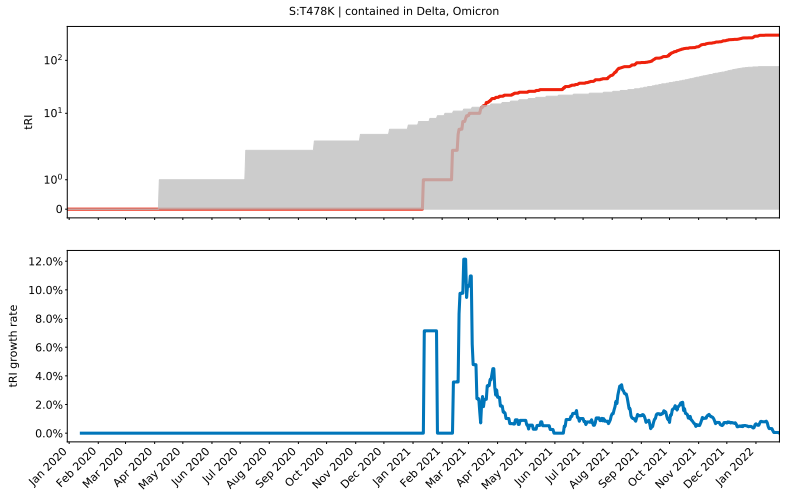


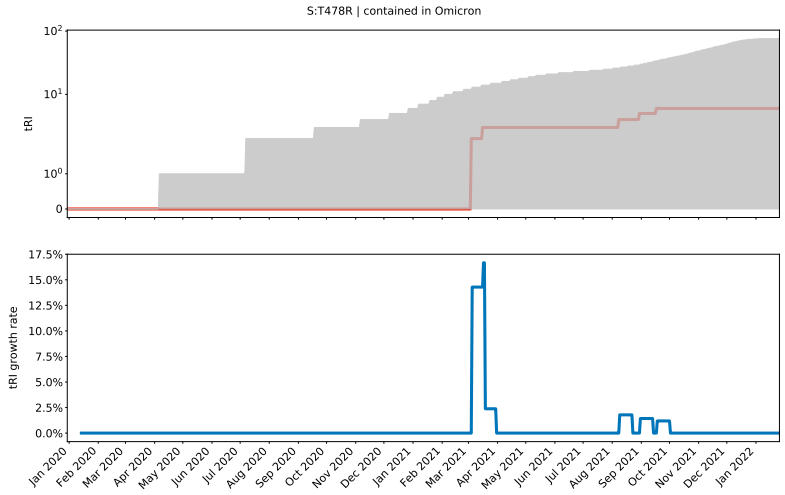




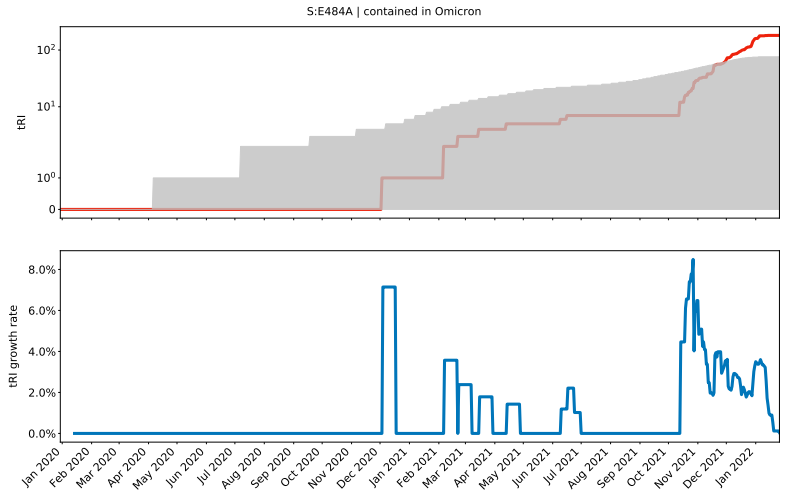
medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

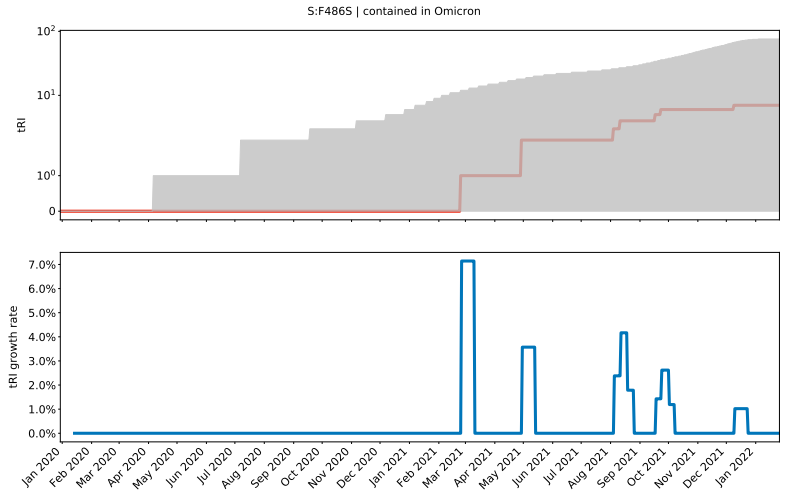


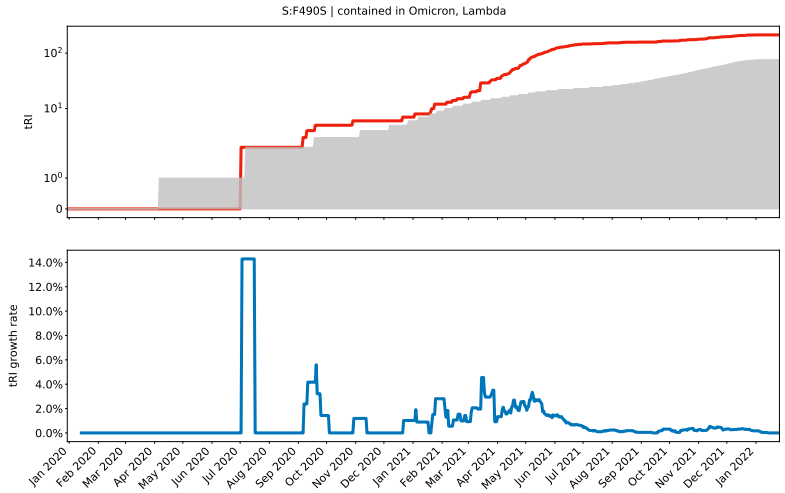




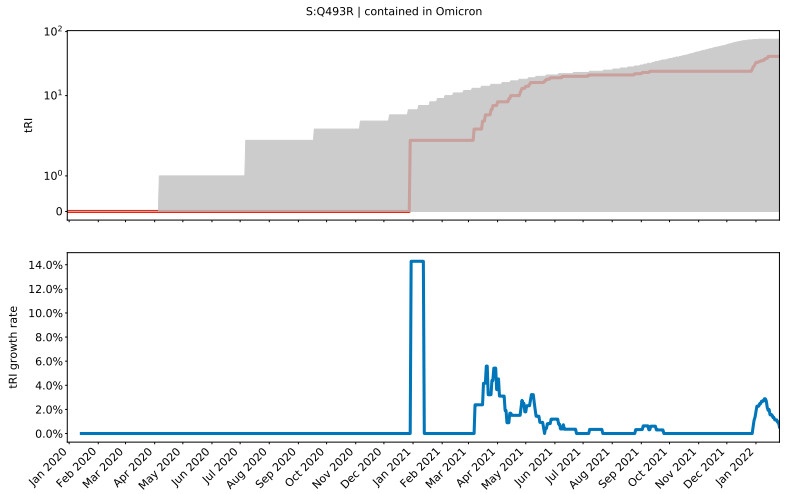
medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

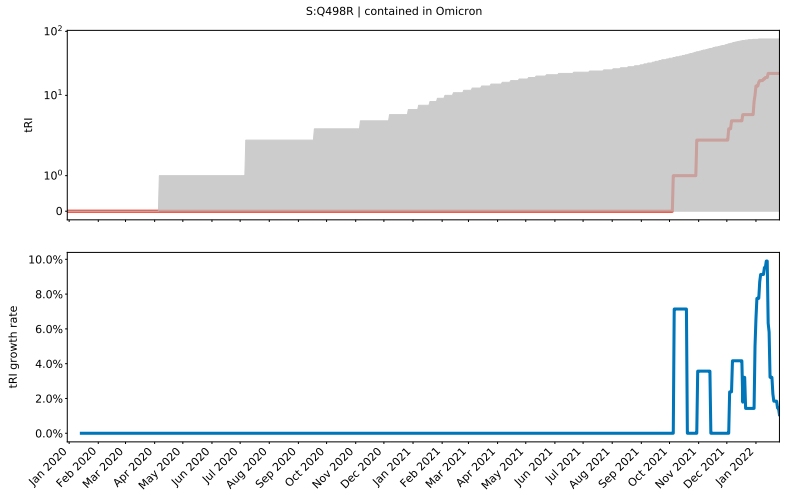


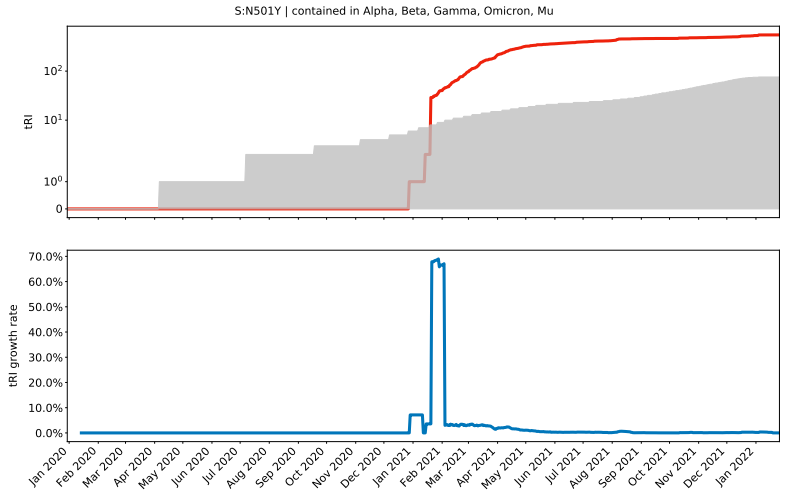




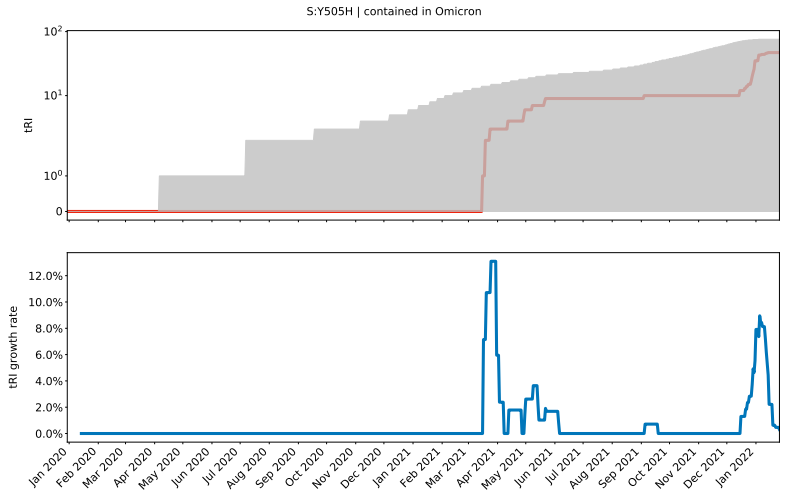
medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

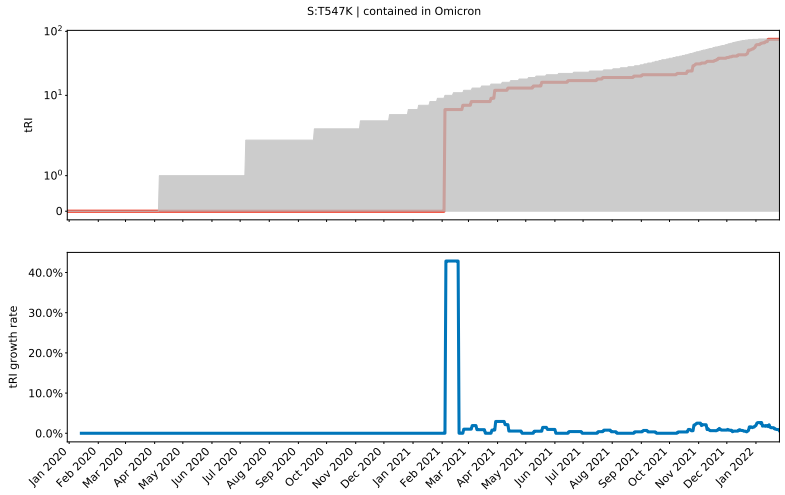


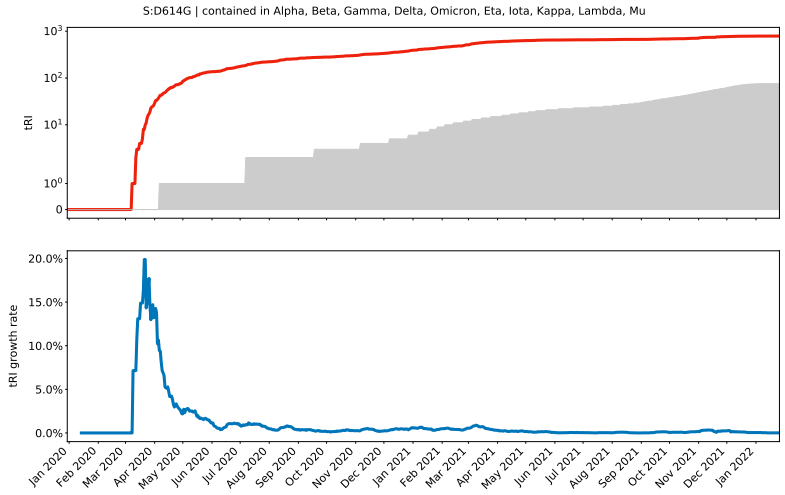




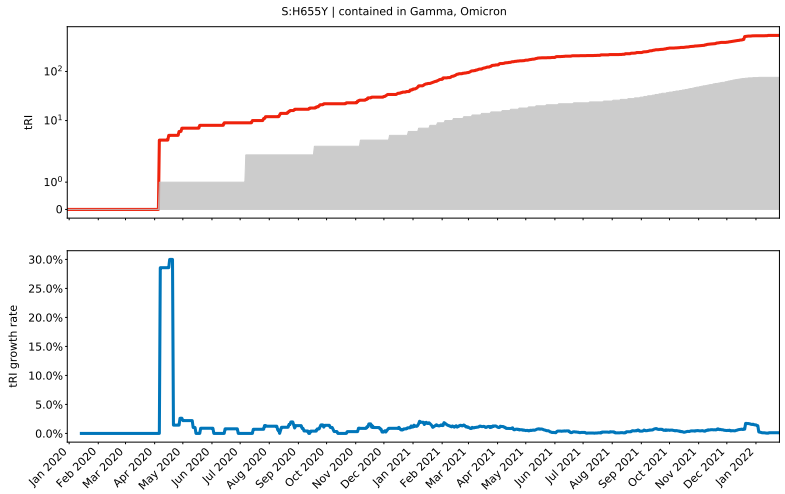
medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

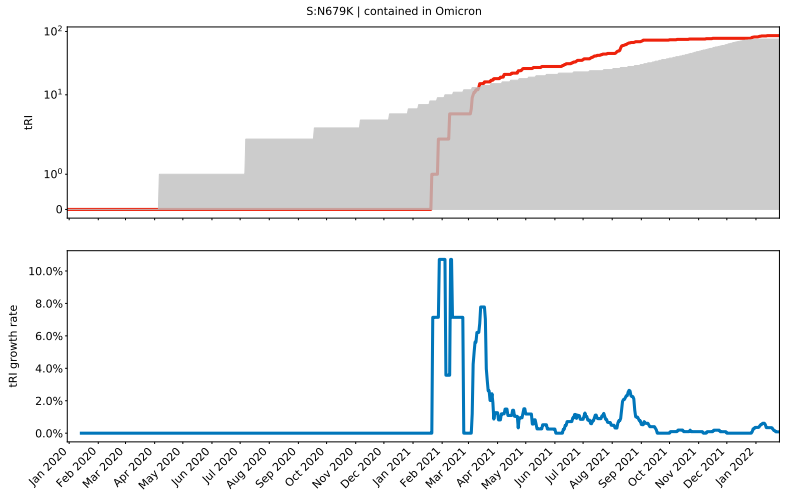


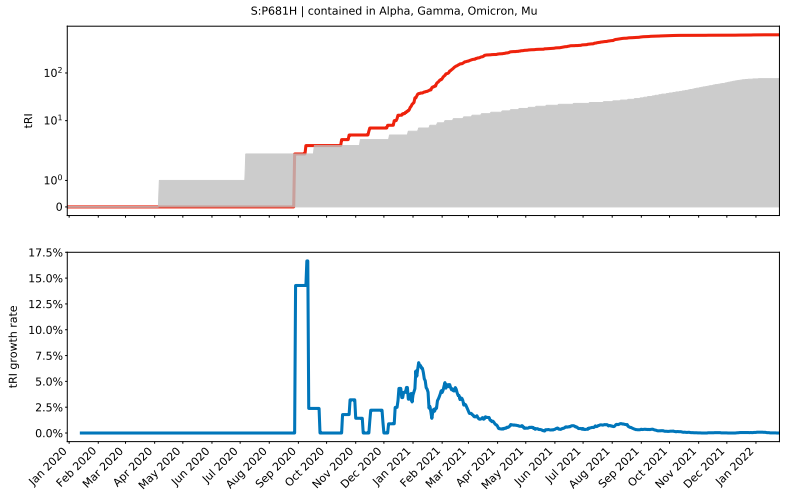




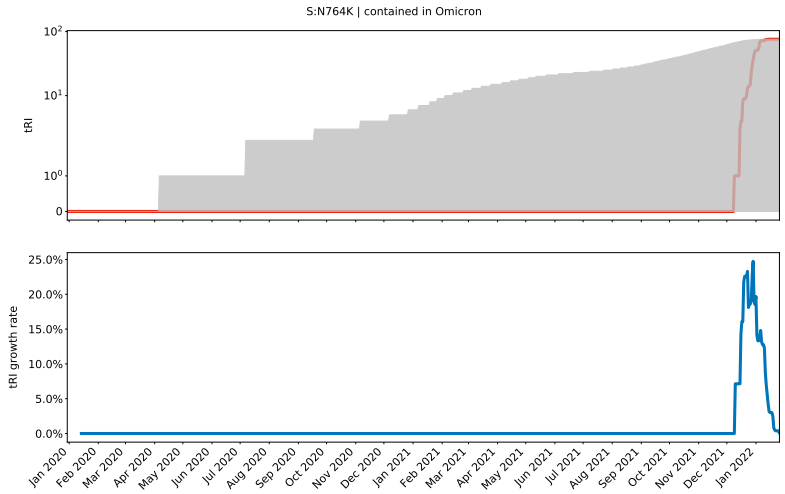
medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

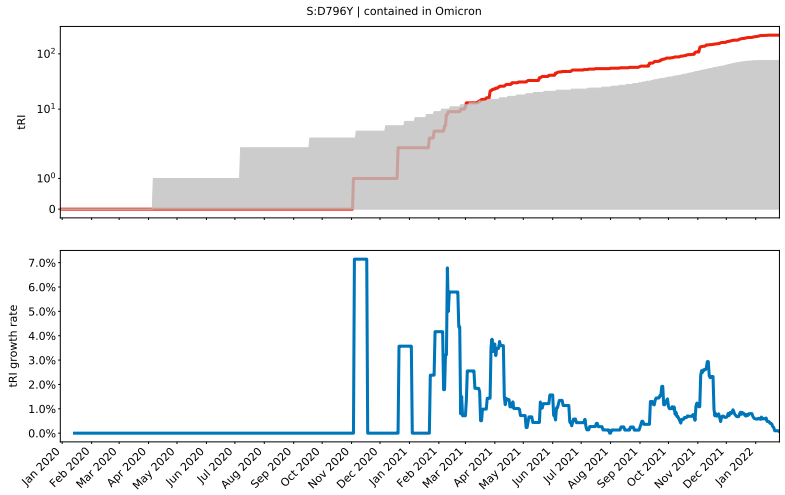


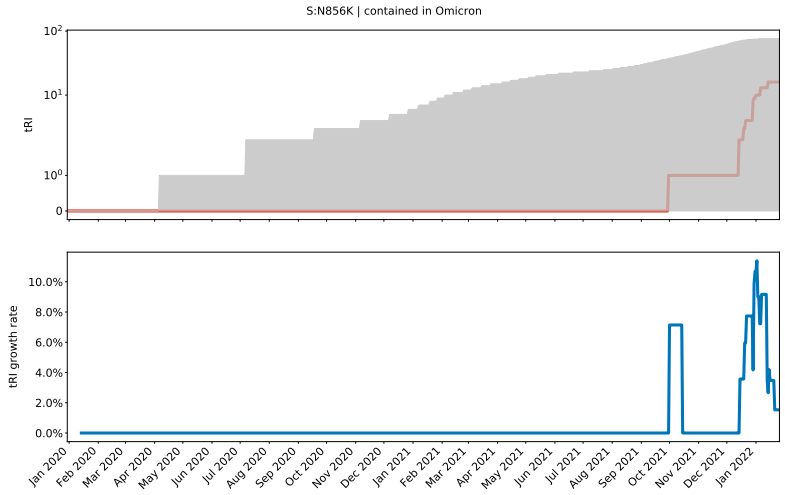
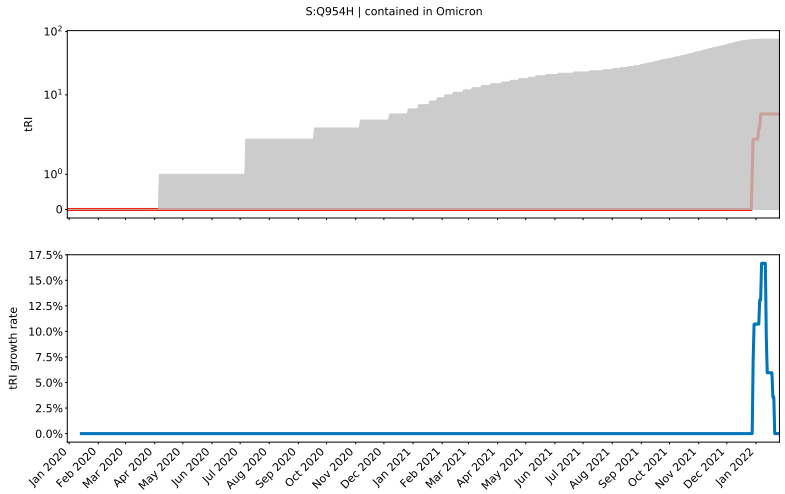
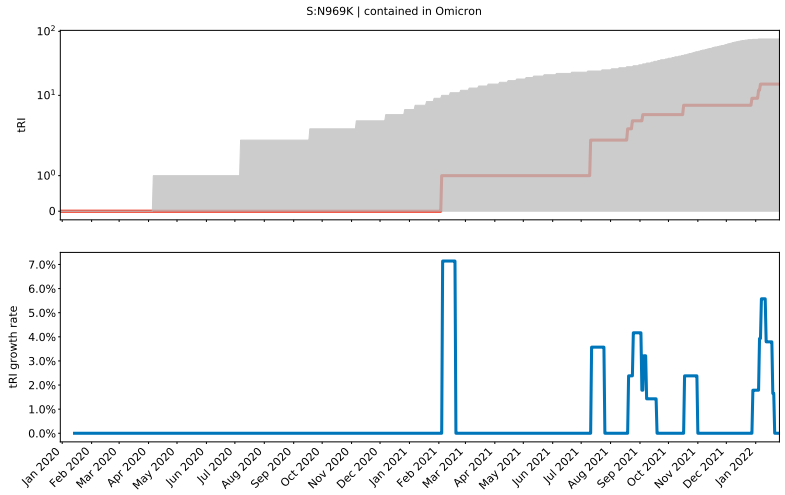
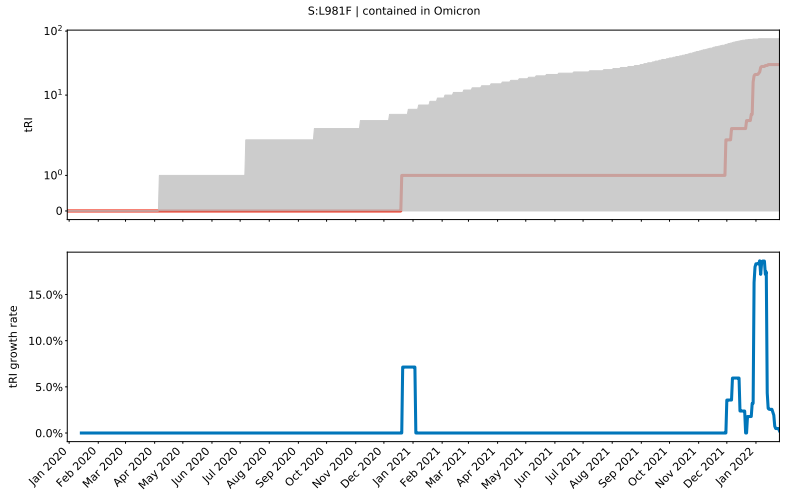




medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .







medRxiv preprint doi: https://doi.org/10.1101/2021.06.10.21258550; this version posted August 28, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .