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

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

[1]Mathematics Department, Karlsruhe Institute of Technology, Karlsruhe, Germany
[2]Mathematical Institute, Heidelberg University, Heidelberg, Germany
[3]Program for Mathematical Genomics, Department of Systems Biology, Columbia University, New York, NY, USA
[4]DataShape, Inria Sophia-Antipolis, Biot, France
[5]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.)
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 establish a fast and scalable early warning system based on Topological Data Analysis for the identification and surveillance of emerging adaptive mutations in large genomic datasets. Analyzing millions of SARS-CoV-2 genomes from GISAID, we demonstrate that topologically salient mutations are linked with an increase in infectivity or immune escape. We report on emerging potentially adaptive mutations as of January 2022, and pinpoint mutations in Variants of Concern that are likely due to convergent evolution. Our approach can improve the surveillance of mutations of concern, guide experimental studies, and aid vaccine development.

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 adaptive advantage [1, 2].

One approach to identify potential 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.

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. If a mutation gives some sort of advantage to the virus, we expect it to occur in several places independently and its frequency to increase with time. In this approach, one usually reconstructs an estimated phylogenetic tree and identifies mutations that occur in independent branches [7–12]. For instance, the `D614G` mutation in the Spike gene was identified early in the pandemic and is now found in virtually all virus isolates [13].

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 [14]. 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]. Phylogenetic approaches, however, become daunting as the number of sequences increases, and are computationally prohibitive when the number of genomes exceeds the tens of thousands [12, 15, 16]. In addition, the independent reemergence of mutations gives rise to homoplasies that confuse the generation of phylogenetic trees [17]. Moreover it has been observed that constructing a single optimal phylogeny for SARS-CoV-2 is generally problematic, as the number of sequences is large while the genetic diversity is low [15].

Here we establish a novel method based on Topological Data Analysis (TDA) that can efficiently identify emerging 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, 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.

Analyzing millions of genomic sequences shared via GISAID, the global data science initiative [14, 18], we first characterize the convergent evolution of the coronavirus SARS-CoV-2 in the early phase of the pandemic from December 2019 until February 2021. We demonstrate that our method can detect adaptive mutations at an early stage, often several weeks before they become recognizable by their prevalence in the population, and explain how this can serve as an early warning and surveillance system. We further report on potentially adaptive mutations in the current phase of the pandemic as of January 2022, and identify mutations in Variants of Concern, most notably in the Omicron variant, that are likely due to convergent evolution.

Results

Quantification of topological recurrence

Chan *et al.* [19] 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).

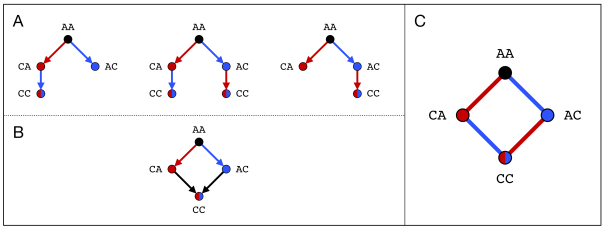


Figure 1. Reticulate events in molecular evolution create topological cycles. (A) and (B) show possible evolutionary histories on the level of individuals in the example of a genome with only two nucleotides. The coloring of the edges corresponds to the acquisition of a specific mutation, while the coloring 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).

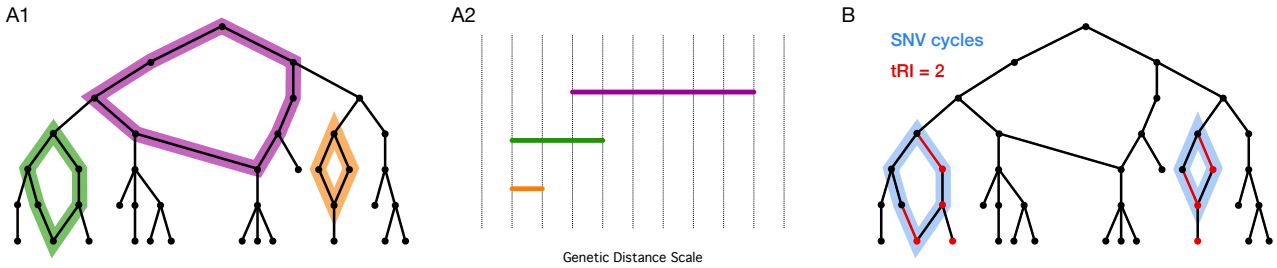


Figure 2. Topological data analysis quantifies convergent evolution. (A) Persistent homology detects reticulate events in viral evolution by means of a persistence barcode. Each bar in the barcode (A2) corresponds to a topological cycle in the reticulate phylogeny (A1). Bars at small genetic distance 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.

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` [20], 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).

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 [15]. 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 datasets with hundreds of thousands of sequences. A performance analysis showed that when applied to sample alignments of up to 5,000 SARS-CoV-2 sequences, the topological method was at least an order of magnitude faster than phylogeny-based methods [7, 8, 21] (see Figure 3). Moreover, the topological recurrence analysis of sequence alignments with millions of SARS-CoV-2 genomes was accomplished in less than a day (see Methods).

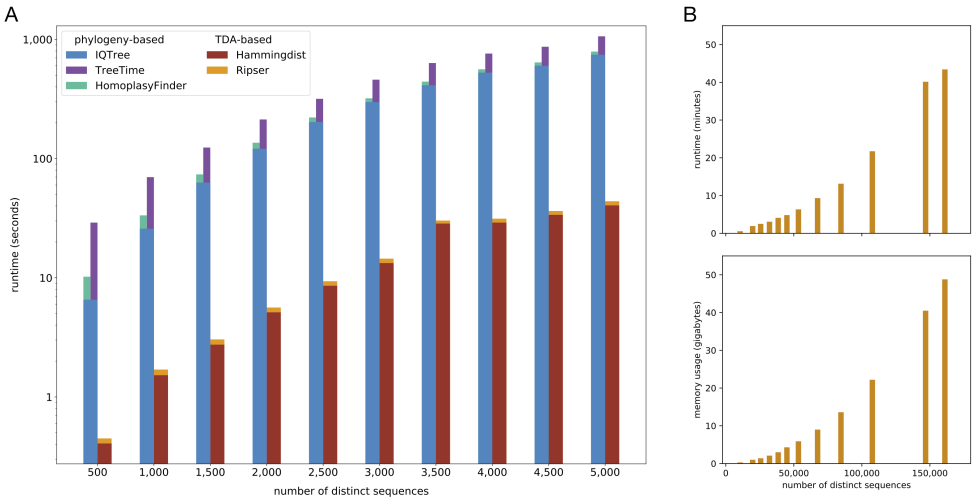


Figure 3. Performance analysis and comparison with phylogeny-based methods. (A) Basic runtime comparison between Topological Data Analysis (TDA)-based methods and standard phylogeny-based methods for random samples of up to 5,000 SARS-CoV-2 genomes. We used `IQTree` [21] to reconstruct phylogenetic trees. The subsequent homoplasy analysis was performed with `TreeTime` [7] and `HomoplasyFinder` [8]. For the TDA-approach we used `Hammingdist` [22] to generate genetic distance matrices and `Ripser` [20] for the subsequent computation of persistence barcodes (see Methods). (B) Runtime and memory usage for the computation of persistence barcodes with `Ripser` [20] 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 [14, 18] (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). In large genomic datasets, homoplasies may arise randomly, causing a certain amount of noise in persistent homology. 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 6 and 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 7). Here signals with tRI ≥ 2 are statistically significant ($p < 0.05$; see Methods). Most of the mutations with strong tRI signal, such as `ORF1ab:G11083T` (tRI $= 141$; see Figure 5), are known to be highly homoplasic and cause stability issues in the construction of phylogenetic trees [10, 15, 17, 23].

A1

Selected mutations across the whole genome (February 2021)

SNV	SAAV	gene	tRI	prevalence
G11083T[1,2]	L3606F	ORF1ab	141	5%
C21575T[1,2]	L5F	S	97	2%
C15324T[1,2]	synonymous	ORF1ab	38	3%
C3037T[1,2]	synonymous	ORF1ab	34	96%
A17615G	K1383R	ORF1ab	33	6%
C2453T	L730F	ORF1ab	31	2%
A10323G[1,2]	K3353R	ORF1ab	29	2%
A23403G[1,2]	D614G	S	29	96%
G25563T[1,2]	Q57H	ORF3a	28	26%
C27964T[2]	S24L	ORF8	27	10%

B1

Selected Spike mutations (February 2021)

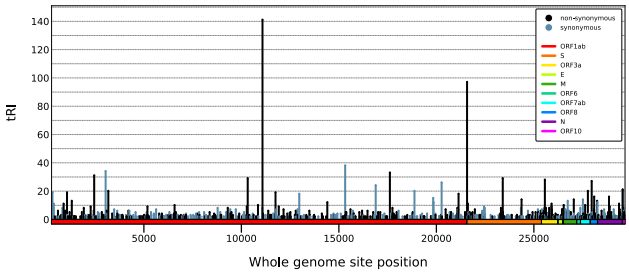
SNV	SAAV	tRI	prevalence
C22227T	A222V	776	19%
C21575T	L5F	466	2%
G22992A	S477N	433	4%
A23403G	D614G	400	96%
C21614T	L18F	362	8%
C21855T	S98F	342	2%
C22444T	synonymous	276	2%
G23593T	Q677H	251	1%
C23731T	synonymous	203	1%
C24334T	synonymous	193	4%
C22388T	synonymous	161	<1%
A22255T	synonymous	145	<1%
G25088T	V1176F	133	<1%
C21637T	synonymous	127	<1%
T24076C	synonymous	123	1%
C21846T	T95I	108	<1%

B2

Selected variable amino acid site positions (February 2021)

site position	total tRI per substitution
D936	Y:91, H:9, N:4, G:2, V:2, E:2
G75	V:43, R:12, S:4, D:2, A:2
G261	V:33, S:9, R:9, C:5, D:3
P26	S:46, L:34, H:8, T:3
D80	Y:60, N:10, A:8, G:3
H146	Y:22, Q:10, R:2, L:2
E156	D:5, Q:3, K:2, G:2
D215	H:41, Y:26, G:4, N:2
D427	Y:5, H:2, G:2, V:2
E484	K:49, G:10, Q:6, A:2
E554	D:6, Q:5, G:3, V:2
V622	F:44, I:8, L:3, A:3
P681	H:92, R:31, L:9, S:6
P1162	S:48, L:35, R:20, Q:3
D1199	Y:7, N:4, H:4, G:2
G1219	C:46, V:33, S:4, A:2

A2



B3

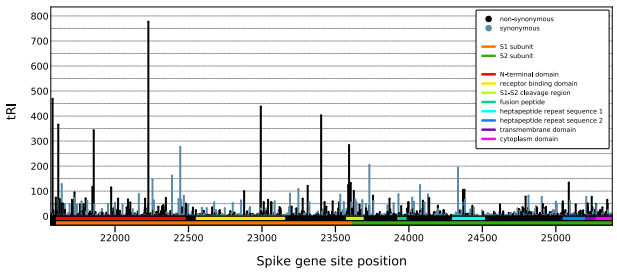


Figure 4. Topological signals in the first year of the pandemic (December 2019 until February 2021). (A) Topological signals of recurrent mutations across the whole genome. The table in (A1) lists the topological recurrence index (tRI) and the prevalence of mutations with strongest topological signal. Many of these have previously been reported to be highly homoplasic in Turakhia *et al.*[1] [17] and van Dorp *et al.*[2] [10]. For a complete list see Supplementary Information Table 7. (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. (B) Topological signals of recurrent mutations across the Spike gene. The table in (B1) lists the topological recurrence index (tRI) and the prevalence of Spike mutations with strongest topological signal. For a complete list see Supplementary Information Table 8. (B2) Values of the tRI for highly variable Spike amino acid site positions. For a complete list see Supplementary Information Table 9. (B3) 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.

Performing the topological recurrence analysis 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 8 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 [24], and in the Spike protein signal peptide as well as in the S1-S2 cleavage region [25] (see Figure 4). We found a strong signal (tRI $= 400$) for

S:D614G, which indicates that this mutation appeared independently at least 400 times during the pandemic. In fact, the mutation increases transmissibility [13, 26] and in vitro infectiousness [27–29]. We performed a comparative time series analysis (tRI vs. prevalence) which 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).

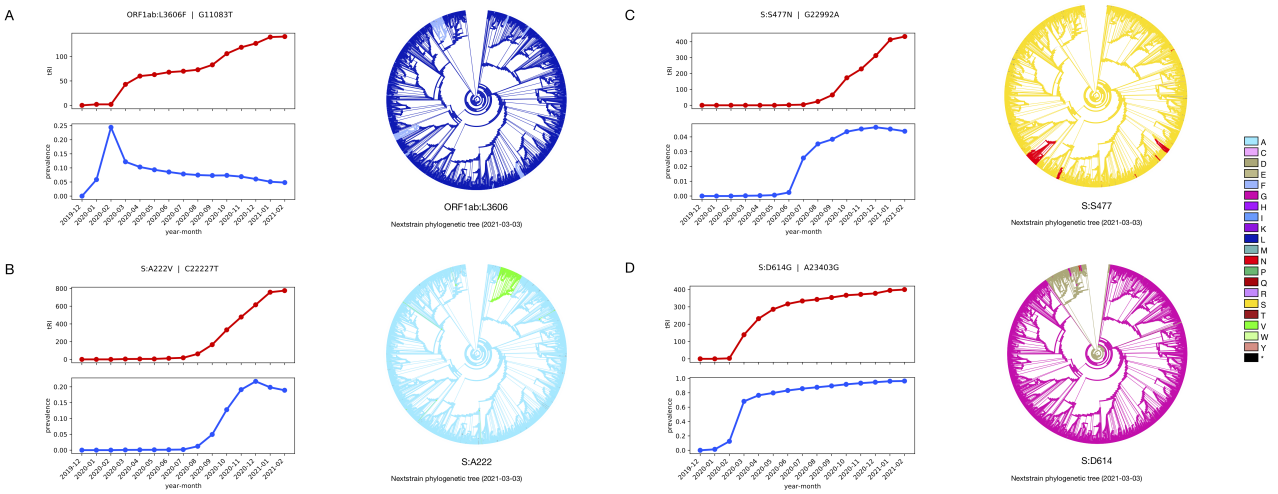


Figure 5. Comparative time series analysis and ancestral state reconstruction analysis for mutations with strongest topological signal in the first year of the pandemic (December 2019 until February 2021). (A) Topological footprint of a highly recurrent mutation in the example of the mutation ORF1ab:G11083T. 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 [16] of a curated subsample of 3,507 sequences from the GISAID dataset as of 3 March 2021 [14, 18].

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 [30]. While S:S477N is known to affect the binding affinity to the ACE2 receptor [3, 31] leading to a slight increase in fitness, there is no conclusive evidence yet whether or not S:A222V also results in higher transmissibility [32]. 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) [30], 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 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 [33–35]. 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 topologically salient mutations occur in lineages designated as VOI/VOC [30]. 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 provides evidence for a shift, beginning in late 2020, towards immune escape as the driving force behind adaptation during the first year of the pandemic.

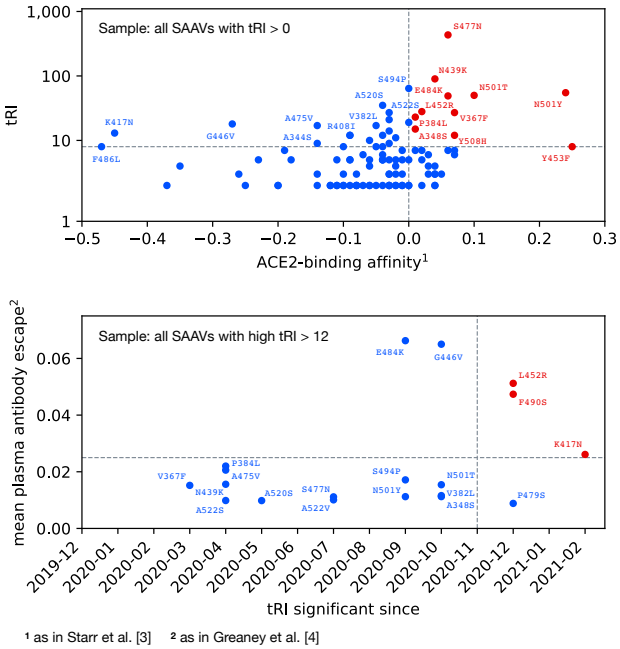


Figure 6. Topological recurrence correlates with increased binding affinity and immune evasion. There is a strong correlation between a significant topological recurrence index (tRI ≥ 8) and an increase in ACE2-binding affinity in the year from December 2019 until February 2021. Among highly topologically recurrent amino acid changes, the more recent ones show increased mean plasma antibody escape. Experimental data taken from Starr *et al.* [3, Table S2] and Greaney *et al.* [4, Table S3].

An early warning and surveillance tool

We found that the sensitivity of the topological recurrence index (tRI) is sufficient to detect signals of convergent evolution already at very low mutation frequencies, often several weeks before the respective mutation becomes recognizable by its prevalence in the population. More specifically, we observed that the onset of significant tRI signals precedes the actual rise in prevalence by several weeks for the adaptive mutations `S:L18F` [36], `K417N` [4], `L452R` [4, 37], `S:E484K` [4, 37], `S:N501Y` [3, 4] and `S:P681H` [38, 39] seen in the VOCs Alpha, Beta, Gamma, Delta and Omicron [30] (see Figure 7). This demonstrates that the topological recurrence index serves as an early indicator for emerging adaptive mutations.

In order to assess the reliability of this indicator, we retrospectively compared mutations that had been identified as potentially adaptive in February 2021 with the actual evolution of the virus, and the emergence of VOIs/VOCs in particular, during the later year 2021. 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, with a rising tendency in February 2021, while they are associated with an increased mean plasma antibody escape > 0.01 [4], 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, 37] and `S:G446S` [4, 40], which had not been observed in notable variants before, occurred in the Lambda [41, 42] and Omicron [43] variants, which were designated as VOI/VOC in June/November 2021 [30]. Similarly, the Spike substitutions `S:T95I`, `S:L452R`,

S:S477N, S:T478K, S:N679K, S:P681R and S:D796Y later occurred in the Delta, Kappa and Omicron variants, which were designated as VOC/VOI/VOC not until May/April/November 2021 [30]. While S:T478K had not yet reached a statistically significant tRI, the tRI of all of these mutations showed a rising tendency in February 2021 (see Table 1 and Supplementary Information Figure 3).

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 salient recurrent mutations during the pandemic (see Figure 7).

Updated analyses are available at https://tdalife.github.io/covtrec.

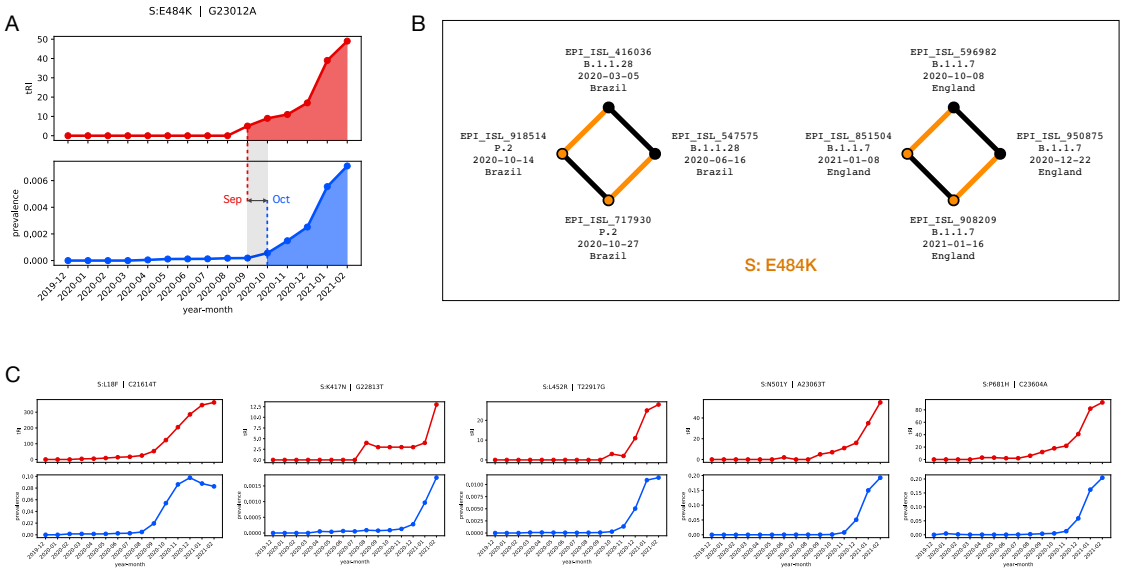


Figure 7. Early warning and surveillance for emerging adaptive Spike mutations. (A) Typical comparative time series pattern (tRI vs. prevalence) of an emerging adaptive mutation as observed in the escape mutation S:E484K. The tRI rises to a significant level (tRI = 5) in September 2020, while the prevalence 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. (B) 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. (C) Typical comparative time series pattern (tRI vs. prevalence) observed in emerging adaptive Spike mutations, with the onset of significant tRI signals preceding the actual rise in prevalence by several weeks.

Ongoing convergent evolution and the Omicron variant

We analyzed the ongoing evolution of the coronavirus by applying our early warning and surveillance pipeline to a curated alignment of 3,928,116 high-quality SARS-CoV-2 Spike gene sequences from GISAID [14, 18] that have been collected during the year from January 2021 until January 2022 (see Methods).

A total of 343 non-synonymous and 189 synonymous Spike mutations with significant tRI signal were found (see Figure 8, Supplementary Information Table 11 and Methods). These mutations are potentially adaptive in the current phase of the pandemic and might therefore appear in future variants. Here signals with tRI ≥ 88 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 8).

A Selected Spike mutations (January 2022)

SAAV	tRI	notable variants
G142D	19601	Delta, Omicron
T95I	7206	Omicron*, Iota, Mu
L5F	4894	Iota
D950N	4304	Delta, Mu
A222V	3475	
V1264L	2068	
S112L	1851	
L18F	1526	Gamma
Q677H	1368	Eta
V1104L	1305	
S98F	1252	
T19R	1160	Delta
K1191N	1150	
T572I	1063	
L452R	734	Delta, Kappa
S477N	695	Omicron
P681H	659	Alpha, Omicron, Mu
E484K	651	Alpha*, Gamma, Eta, Iota, Mu
A701V	609	Beta, Iota
N501Y	604	Alpha, Beta, Gamma, Omicron, Mu
A67V	586	Omicron*, Eta
H655Y	564	Gamma, Omicron

* mutation found in some sequences but not all

C Omicron Spike mutations (January 2022)

SAAV	tRI	SAAV	tRI
T19I*	**135**	**S477N**	**695**
L24S*	**89**	**T478K**	**341**
A67V*	**586**	**E484A**	**153**
T95I*	**7206**	Q493R	45
G142D	**19601**	G496S*	50
N211I*	5	Q498R	21
V213G*	0	**N501Y**	**604**
G339D*	**95**	Y505H	45
S371F*	2	T547K*	70
S371L*	0	**D614G**	**367**
S373P*	41	**H655Y**	**564**
S375F*	43	**N679K**	**147**
T376A*	4	**P681H**	**659**
S413R*	0	N764K*	41
D405N*	10	**D796Y**	**244**
R408S*	7	N856K*	10
K417N*	**402**	Q954H	3
N440K*	**227**	N969K	11
G446S*	**151**	L981F*	19

boldface highlights significant tRI

* mutation found in some sequences but not all

B RBD amino acid substitutions (January 2022)

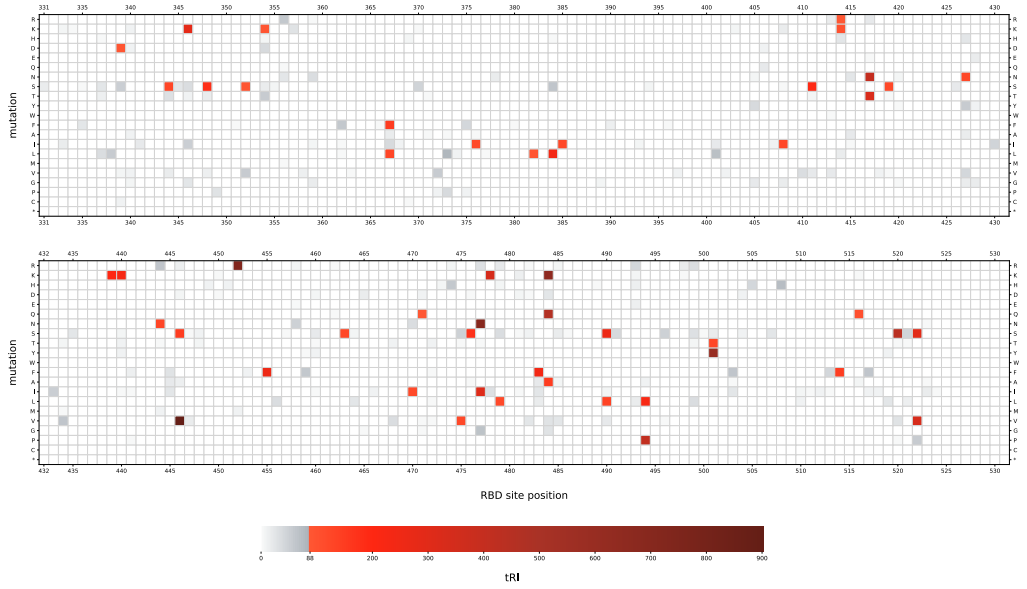


Figure 8. Spike gene mutations with strong topological signal of convergence as of January 2022 and the Omicron variant. (A) The table displays the top 14 amino acid substitutions on the Spike gene with significant topological recurrence index (tRI $\geq$ 88), plus a few more selected mutations. Among those 14 substitutions, 50% occur in lineages designated as VOI/VOC [30]. For the full table see Supplementary Information Table 11. (B) Heatmap of all amino acid variations across the RBD showing topological signals of convergence. Significant signals (tRI $\geq$ 88; 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) [30, 44]. 47% of these mutations (53%/56% in the sublineage BA.1/BA.2) are significantly topologically recurrent with tRI $\geq$ 88 (displayed in boldface) and are likely due to convergent evolution.

Of particular interest are the defining amino acid substitutions in the Omicron variant B.1.1.529 [43, 44], which was designated as VOC in November 2021 [30] and is rapidly spreading all over the world in January 2022 [45]. Our analysis revealed that 47% (18 out of 38) of the Spike amino acid substitutions seen in B.1.1.529—with 53% (16 out of 30) in the sublineage BA.1 and 56% (15 out of 27) in the sublineage BA.2—show a significant topological signal of recurrence (see Figure 8). This includes, among 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:N679` and `S:P681H` at the S1-S2 cleavage region [25, 46]. Our findings provide further insight into the possible origins of the Spike mutations observed in the Omicron variant [47]. In fact, all 18 substitutions with significant topological recurrence index tRI ≥ 88 likely arose from convergent evolution, while the emergence of the remaining substitutions showing weak topological signals may be due to other reasons.

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 [48]. 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. Our method does not rely on the prior reconstruction of an optimal phylogenetic tree and therefore (i) outperforms phylogeny-based methods [7, 8, 21] by at least an order of magnitude, and (ii) accomplishes the analysis of current datasets with millions of SARS-CoV-2 genomes within just one day. However, this gain in performance comes at the cost that, in contrast to phylogeny-based methods, our method will resolve homoplasies only on small genetic distance scales. We further remark that although persistent homology is robust with respect to noise in the genomic dataset [49], systematic sequencing errors [50, 51] might tamper topological inference to some extent.

Drawing from data from GISAID [14, 18], we applied our analysis pipeline to monitor the adaptive evolution of the coronavirus SARS-CoV-2 in two phases of the pandemic—the early phase from December 2019 until February 2021, and the current phase as of January 2022. The topological analysis of the early phase 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. 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. Including experimental data obtained by Starr *et al.* [3] and Greaney *et al.* [4] into our analysis, we inferred that on the receptor-binding domain topological recurrence correlates with host adaptation. Specifically, we observed a beginning shift in late 2020 from host receptor binding towards immune evasion as the main selective pressure. The topological analysis of the current phase of the pandemic revealed a total of 532 topologically recurrent mutations on the Spike gene.

We demonstrated that our method (i) is sufficiently sensitive to detect emerging adaptive mutations already at an early stage when mutation frequencies are still very low, and (ii) provides geographic and temporal information about virus isolates involved in concrete reticulate events. We identified potentially adaptive mutations on the Spike gene that showed significant topological signals of recurrence with a rising tendency in February 2021. These mutations are likely to appear in future variants, and in fact, several of these candidate mutations have later

received special attention as features of the Delta, Lambda and Omicron Variants of Interest/Concern [30], 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, at least half of the Spike gene mutations observed in the Omicron variant are likely due to convergent evolution. Updated analyses are available at `https://tdalife.github.io/covtrec`.

Based on these insights, we propose persistent homology as an early warning system for the emergence of new adaptive mutations in the ongoing COVID-19 pandemic and foresee this capability also in future pandemics of various pathogens. Basically, our method allows for a targeted mapping of recurrent mutations in any region of the genome, which can guide and motivate the experimental study of adaptive effects of specific mutations also in genes other than the Spike gene. Our method could in particular aid the development and rapid adaptation of vaccines for new emerging variants [2, 52]. 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 use two separate datasets of SARS-CoV-2 genome sequences provided by the GISAID EpiCoV Database [14, 18]. 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 (Version 3.8.31)` [53], 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 (character "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 alignment `msa_0117.fasta` which we downloaded from GISAID on 19 January 2022. It comprises 6,475,061 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 (Version 7.497)` [54]. Sequences with collection date between 1 January 2021 and 17 January 2022 were selected, then truncated to the Spike gene (reference site positions 21,563 to 25,384), and finally sequences containing any characters other than A, C, T or G were removed. This resulted in an alignment comprising 3,928,116 complete SARS-CoV-2 Spike genes of length 4,669nt. A list of accession numbers of all sequences in these two alignments, along with an acknowledgement of the contributions of both the submitting and the originating laboratories, can be retrieved through the Data Acknowledgement Locator at `https://www.gisaid.org` with IDs `EPI_SET_20220127bo` and `EPI_SET_20220124he`.

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.

Distance matrices

We used `Hammingdist (Version 0.13.0)` [22] 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 Spike gene alignment covering the year from January 2021 until January 2022 contained 291,141 genetically distinct sequences. We computed the genetic distance matrix for this alignment.

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).

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 5).

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

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 [20]. 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 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* [56, 57], 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 (Version 1.0)` [58] 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 6). 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 6). 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` [22] and `Ripser` [20]. 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.

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 13). 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 [59, 60]. 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 12). 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 13. 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 6). 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 6).

Topological recurrence analysis

We performed topological recurrence analyses for the whole genome and for the Spike gene. Notably, the analysis can also be carried out for single genes, based on alignments of appropriately truncated genome sequences. In this case, evolutionary processes outside the specific gene are ignored, leading to the creation of more topological cycles and hence to a more detailed picture of the ongoing convergent evolution in the respective gene.

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 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 ($p < 0.05$).

An analogous analysis was carried out for the Spike gene alignment covering the year from January 2021 until January 2022. In that case, a tRI ≥ 88 is significant ($p < 0.05$).

Performance analysis

We performed a basic runtime comparison between Topological Data Analysis (TDA)-based methods and standard phylogeny-based methods for random samples of up to 5,000 SARS-CoV-2 genomes drawn from the GISAID alignment covering the first year of the pandemic. We used `IQTree` [21] to reconstruct phylogenetic trees (with default settings and fast search option). The subsequent homoplasy analysis was performed with `TreeTime` [7] and `HomoplasyFinder` [8] (with default settings). For the TDA-approach we used `Hammingdist` [22] (with OpenMP multithreading disabled) to generate genetic distance matrices, and `Ripser` [20] (with scale parameter threshold set to 2) for the subsequent computation of persistence barcodes. All computations were carried out on one kernel of an Intel Xeon E7-4870 processor. 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` [22] (with OpenMP multithreading enabled) on a virtual 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` [20] (with scale parameter threshold set to 2) on an Intel Xeon Gold 6230R processor. 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).

Similarly, for the Spike gene alignment with 3,928,116 sequences covering the year from Janaury 2021 until January 2022, the computation of the genetic distance matrix for 291,141 genetically distinct sequences took 4 hours, and the computation of the corresponding persistence barcodes was completed within 2.2 hours.

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 (Version 3.61)` [61]. As input we used a curated alignment of 3,507 genome sequences and its corresponding Maximum-Likelihood tree, downloaded from Nextstrain [16] 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.

Data availability

The SARS-CoV-2 genome data used in this work are available from the GISAID EpiCov Database [14, 18] at `https://www.gisaid.org` and can be retrieved through the Data Acknowledgement Locator with IDs `EPI_SET_20220127bo` and `EPI_SET_20220124he`. Experimental data on viral phenotypes by Starr *et al.* and Greaney *et al.* is available from [3, Table S2] and [4, Table S3].

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.

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. Sagulenko, P., Puller, V. & Neher, R. A. TreeTime: Maximum-Likelihood Phylodynamic Analysis. *Virus Evolution* **4** (2018). doi:10.1093/ve/vex042.
8. 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.

9. 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.

10. 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.

11. 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.

12. 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.

13. 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.

14. 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.

15. 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.

16. 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.

17. 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.

18. 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.

19. 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.

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

21. 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.

22. Keegan, L. & Kempf, D. *Hammingdist: A Fast Tool to Calculate Hamming Distances* version 0.13.0. 2021. https://github.com/ssciwr/hammingdist visited on 2021-12-01.

23. 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.

24. 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.

25. 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.

26. 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.

27. 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.

28. 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.

29. 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.

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

31. 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.

32. 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.

33. 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.

34. 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.

35. 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.

36. 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.

37. 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.

38. 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.

39. 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.

40. 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.

41. 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.

42. 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.

43. 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.

44. 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.

45. *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.

46. Willett, B. J., Grove, J., MacLean, O., *et al.* The hyper-transmissible SARS-CoV-2 Omicron variant exhibits significant antigenic change, vaccine escape and a switch in cell entry mechanism. *medRxiv* (2022). doi:10.1101/2022.01.03.21268111.

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

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

49. 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.

50. 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.

51. 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.

52. Noh, J. Y., Jeong, H. W. & Shin, E.-C. SARS-CoV-2 Mutations, Vaccines, and Immunity: Implication of Variants of Concern. *Signal Transduction and Targeted Therapy* **6,** 203 (2021). doi:10.1038/s41392-021-00623-2.

53. 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.

54. 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.

55. 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.

56. 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.

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

58. 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.

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

60. 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.

61. 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.

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

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 [14, 18], on which this research is based. Acknowledgement tables can be retrieved through the Data Acknowledgement Locator at https://www.gisaid.org with IDs EPI_SET_20220127bo and EPI_SET_20220124he. 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 [62]. The authors further acknowledge support 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. A.O. acknowledges funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 281869850 (RTG 2229).

Author contributions

M.B., L.H., A.O., J.P.G., R.R. 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., 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.

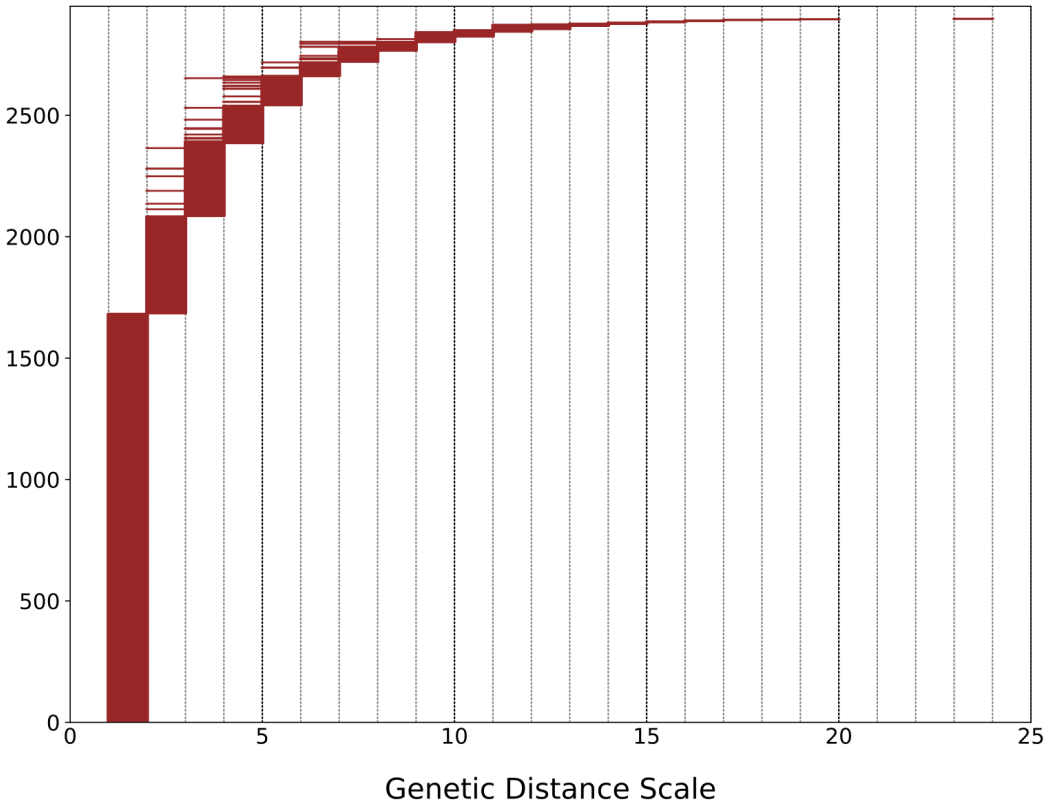
Competing interests

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.

Supplementary Information



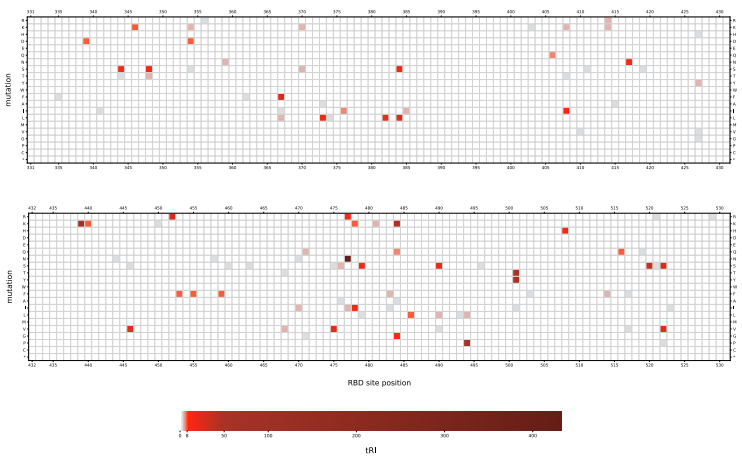
Supplementary Information Figure 1. Persistent homology of the GISAID dataset. Persistence barcode representing the persistent homology in dimension one of the GISAID [14, 18] dataset covering the first year of the pandemic from December 2019 until February 2021, comprising 161,024 genetically distinct high-quality SARS-CoV-2 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.

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

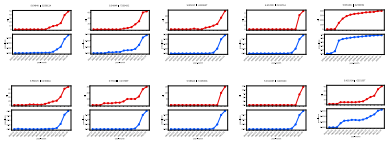
B



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 [44], 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.

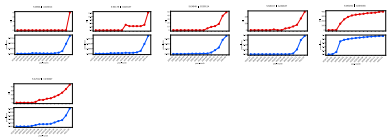
Alpha

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



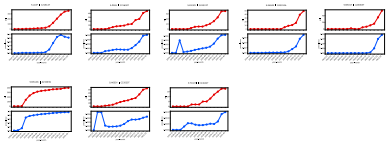
Beta

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



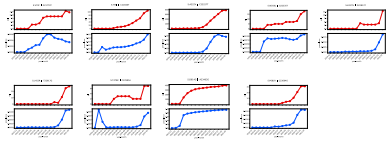
Gamma

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

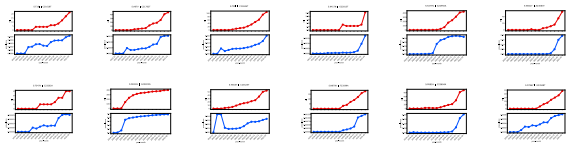
Spike mutation	tRI
T19R	0
V70F*	12
T95I	108
G142D	2
R158G	0
A222V*	776
W258L*	18
K417N*	13
L452R	28
T478K	7
D614G	400
P681R	31
D950N	6



Omicron

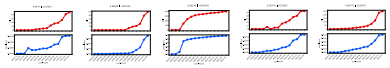
Spike mutation	tRI
T19I*	22
L24S*	4
A67V*	46
T95I*	108
G142D	2
N211I*	0
V213G*	0
G339D*	7
S371F*	0
S371L*	0
S373P*	0
S375F*	0
T376A*	0
S413R*	0
D405N*	0
R408S*	0
K417N*	13
N440K*	7
G446S*	2

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



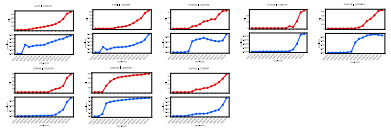
Eta

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



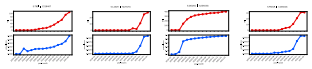
Iota

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



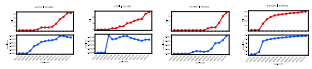
Kappa

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



Lambda

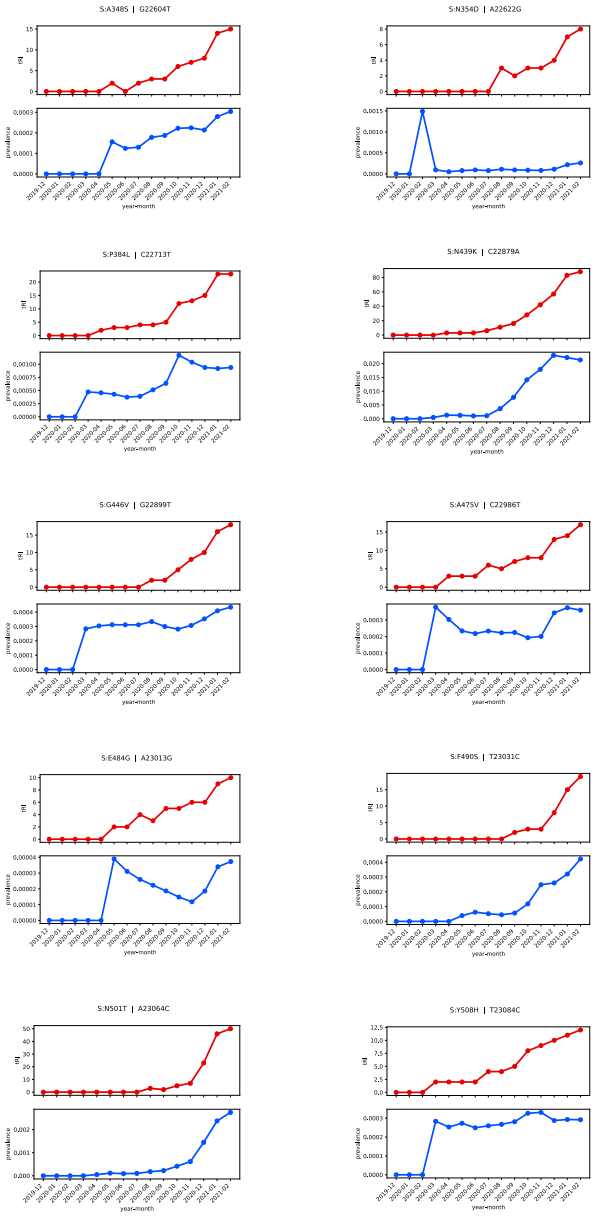
Spike mutation	tRI
G75V	43
T76I	28
L452Q	0
F490S	19
D614G	400
T859N	5



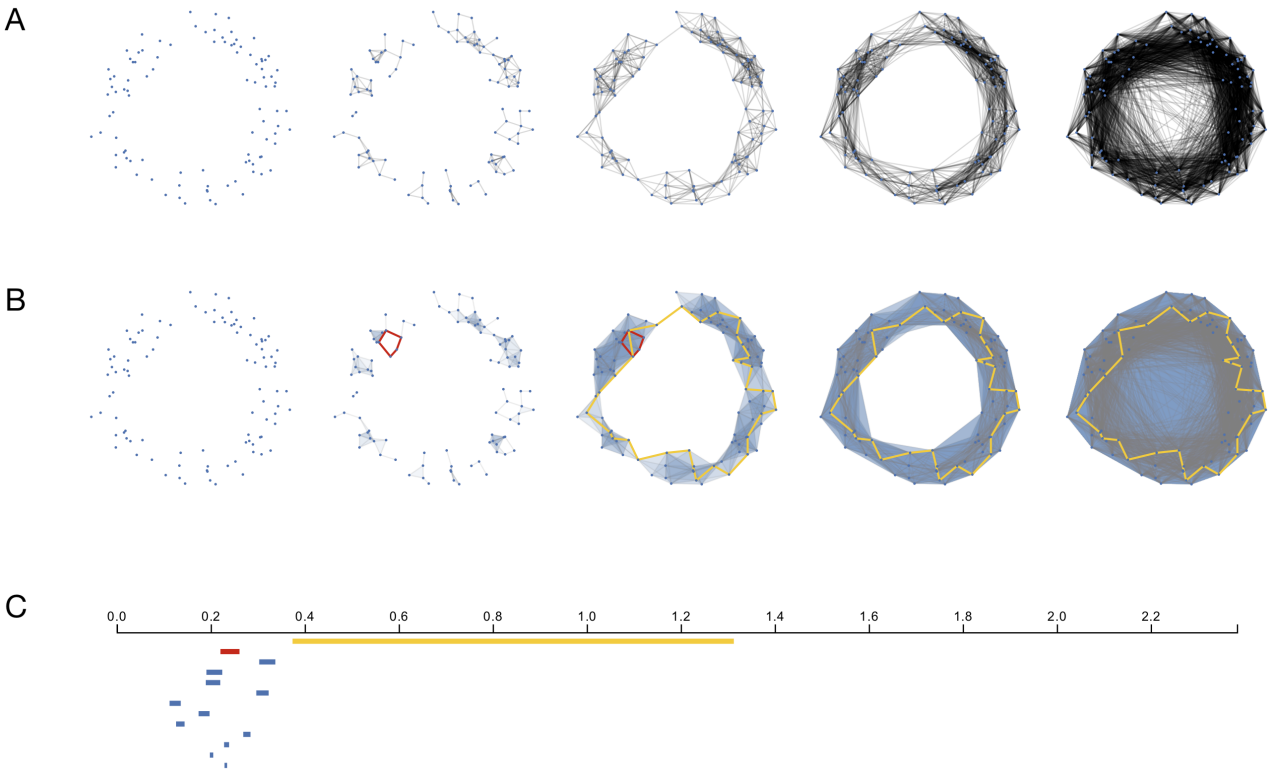
* mutation found in some sequences but not all

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 [30].

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

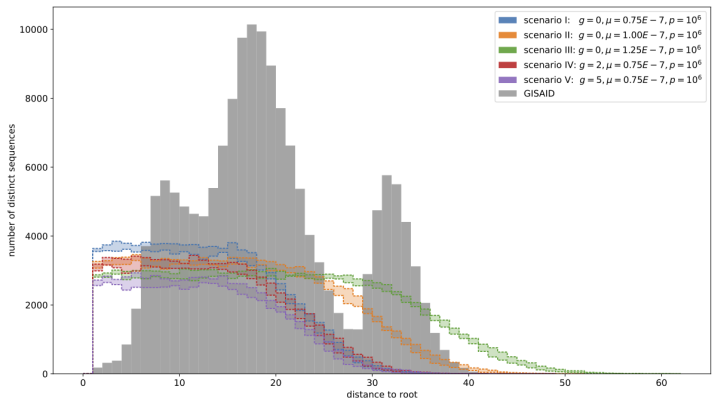


Supplementary Information Figure 4. Early warning for emerging escape mutations on the receptor-binding domain. Comparative time series analysis charts (tRI vs. prevalence) for mutations on the receptor-binding domain that showed a significant tRI signal with rising tendency in February 2021 and are 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, 40] and `S:F490S` [4, 37] later appeared in the Lambda [41, 42] and Omicron [43] variants, which were designated as VOI/VOC in June/November 2021 [30].

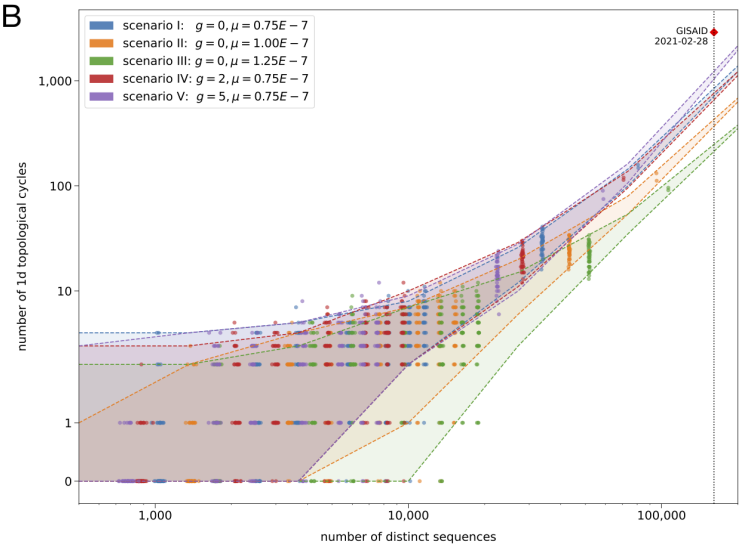


Supplementary Information Figure 5. The Vietoris–Rips filtration of a point cloud. Each point represents a sample, and we display the geometric graphs (A), the resulting Vietoris–Rips complexes at different scales (B) and the persistence barcode in dimension one (C). If one only chooses one scale, one 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.

A



B



C

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 6. The number of topological features in the GISAID dataset covering the first year of the pandemic (December 2019 until February 2021) is statistically significant. (A) 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. (B) Simulations were generated with `SANTA-SIM` [58] 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 extrapolation in the worst case scenario V predicts that less than 40% of all topological cycles (1,171 topological cycles) appear randomly in the phylogeny. (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 $c = 10^6$.

Supplementary Information Table 7. External spreadsheet containing full results of the topological recurrence analysis for the whole genome in the first year of the pandemic (December 2019 until February 2021) . The table lists mutations together with their topological recurrence index (tRI) and prevalence. All mutations with statistically significant tRI ≥ 2 are included.

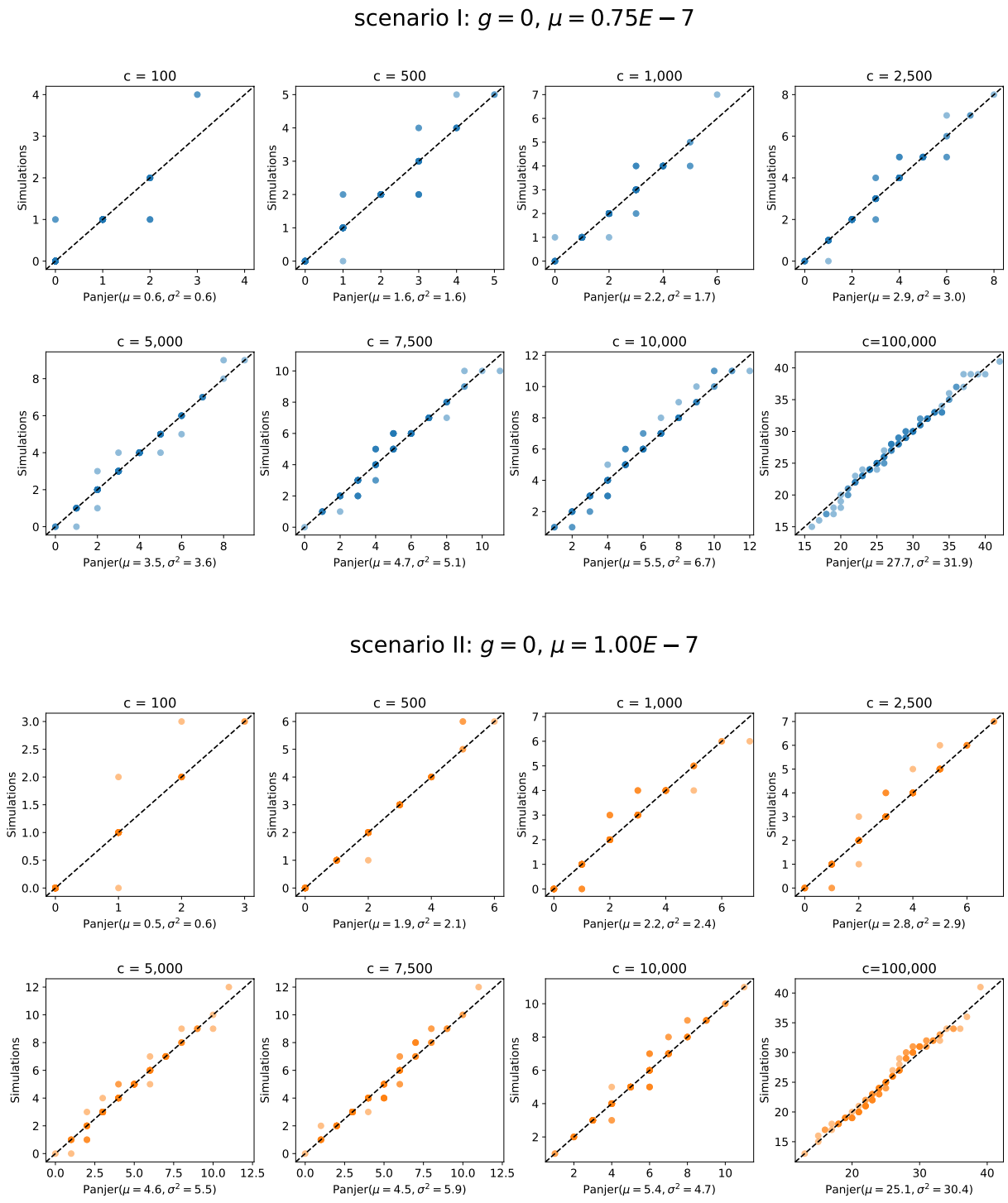
Supplementary Information Table 8. External spreadsheet containing full results of the topological recurrence analysis for the Spike gene in the first year of the pandemic (December 2019 until February 2021). The table lists mutations together with their topological recurrence index (tRI) and prevalence. All mutations with tRI ≥ 2 are included, but only a tRI ≥ 8 is statistically significant.

Supplementary Information Table 9. External spreadsheet containing a complete list of the topological recurrence index (tRI) for all variable amino acid site positions on the Spike gene in the first year of the pandemic (December 2019 until February 2021). All variable site positions with tRI ≥ 2 are included, but only a tRI ≥ 8 is statistically significant.

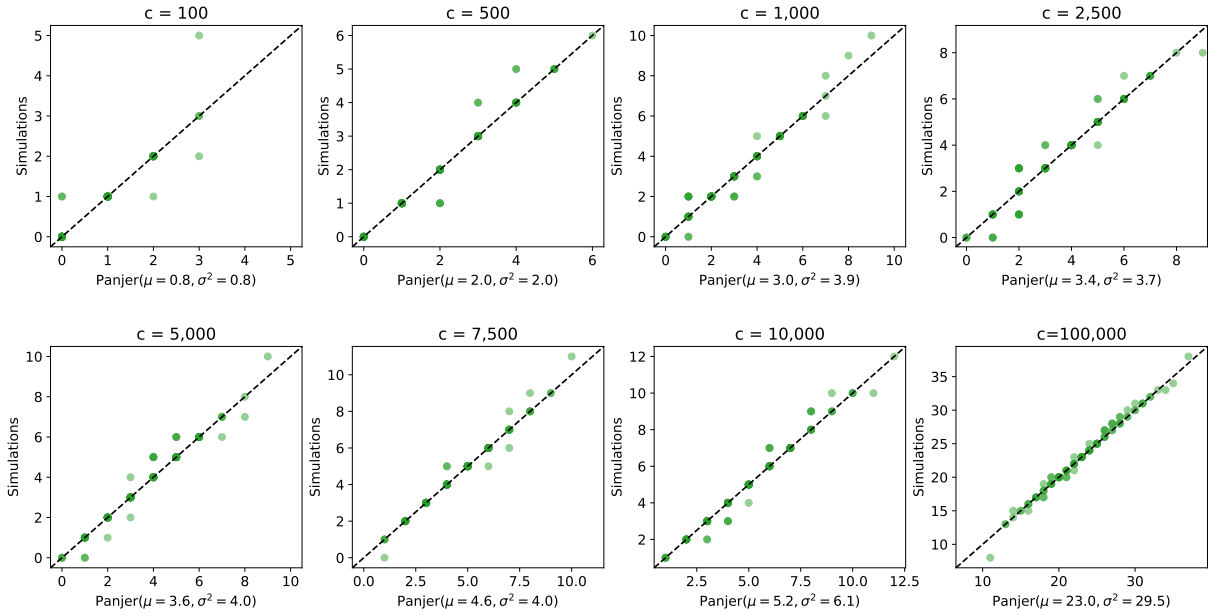
Supplementary Information Table 10. External spreadsheet containing a sublist of the list in Supplementary Information Table 8 featuring all mutations on the receptor-binding domain together with their topological recurrence index (tRI) and prevalence.

Supplementary Information Table 11. 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 ≥ 88 is statistically significant.

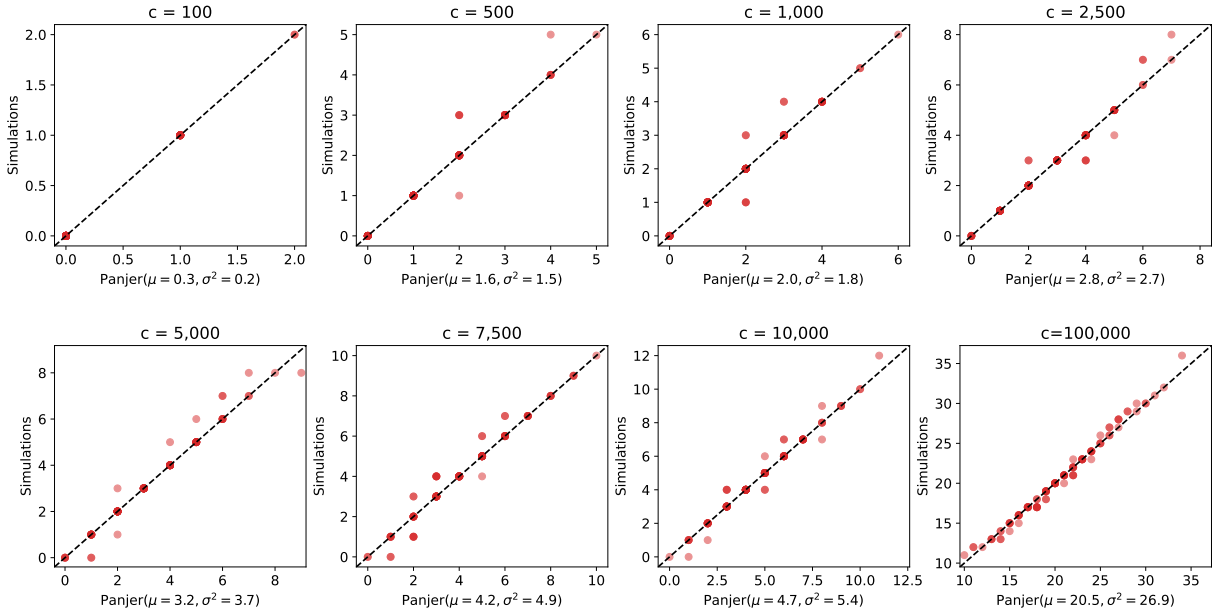
Supplementary Information Figure 12. Quantile-quantile analysis of Panjer distribution versus observed number of one-dimensional cycles in simulated phylogenies. For each value of carrying population c we determined the mean and variation of the observations and used these as parameters for the Panjer distribution.



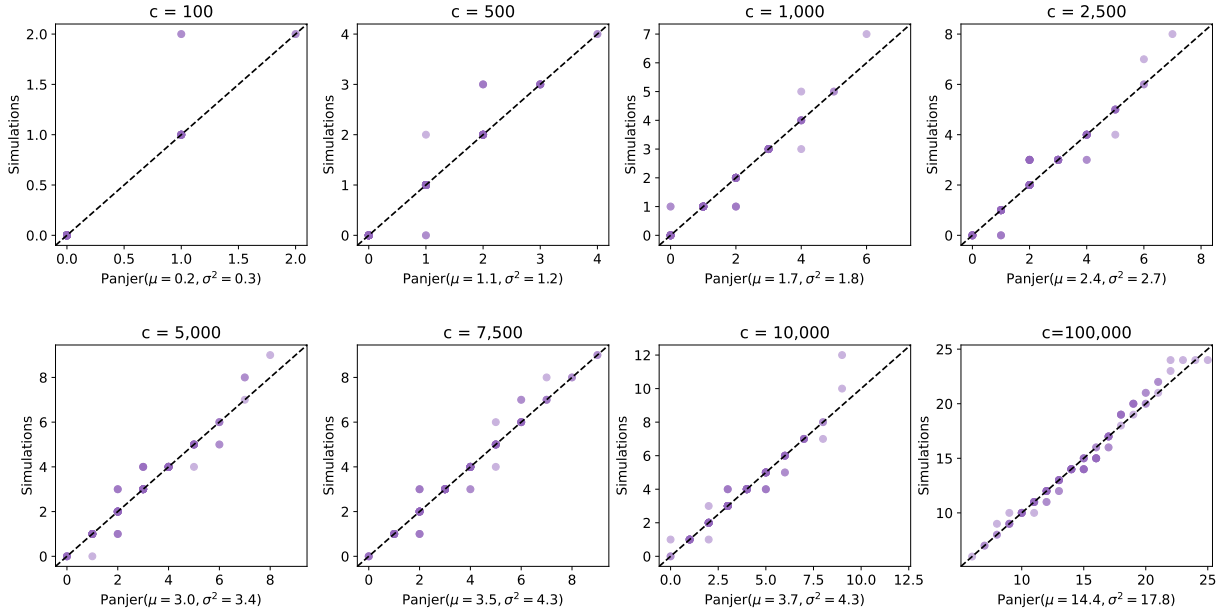
scenario III: $g = 0$, $\mu = 1.25E-7$



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

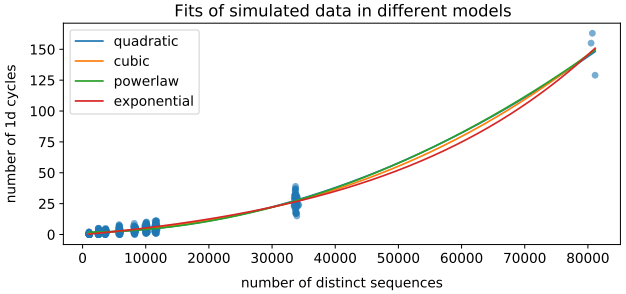


scenario V: $g = 5$, $\mu = 0.75E - 7$



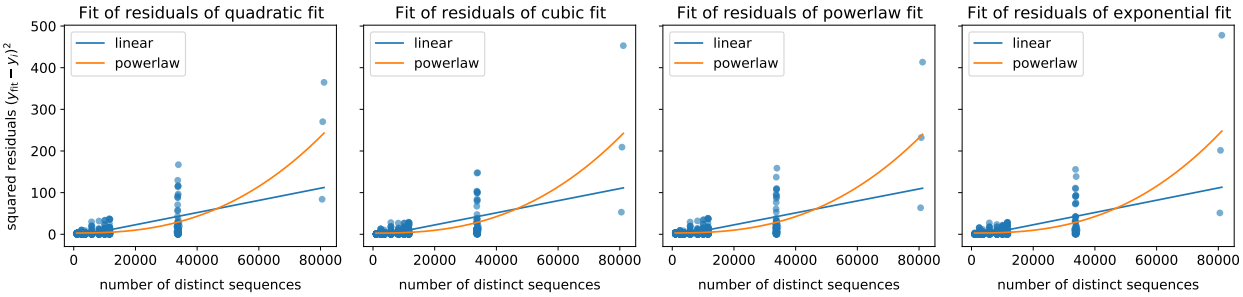
Supplementary Information Figure 13. Analysis of models that extrapolate the simulated data. For each scenario we fit quadratic, cubic, powerlaw, and exponential models to the observed number of one-dimensional cycles in simulations. Then we fit a linear and powerlaw model to the corresponding residuals 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: $g = 0, \mu = 0.75E-7$

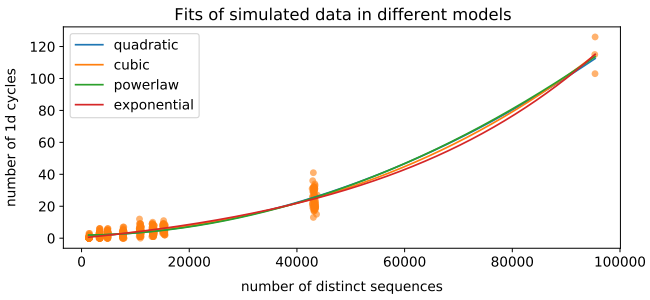


Log-Likelihood of validation data

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

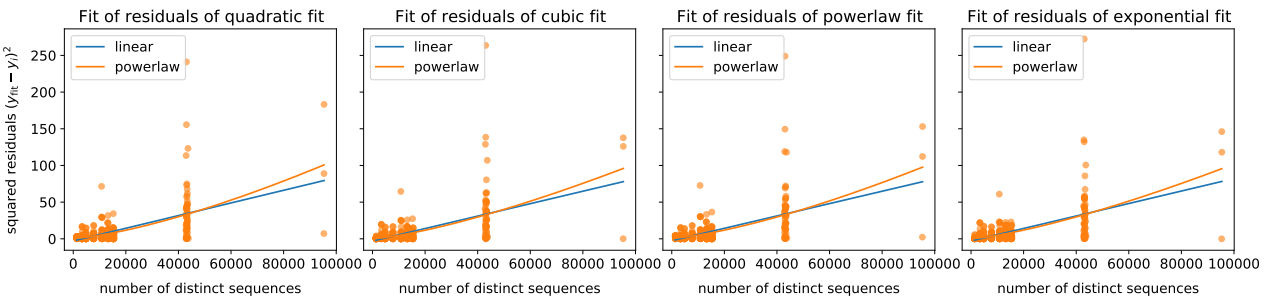


scenario II: $g = 0, \mu = 1.00E-7$

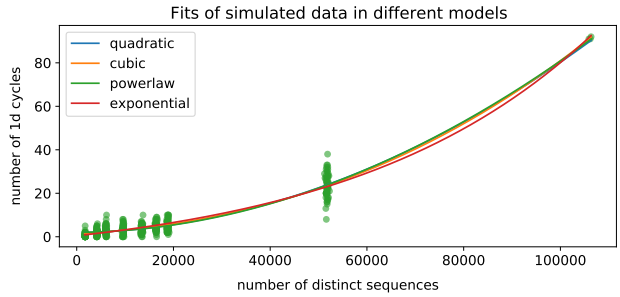


Log-Likelihood of validation data

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

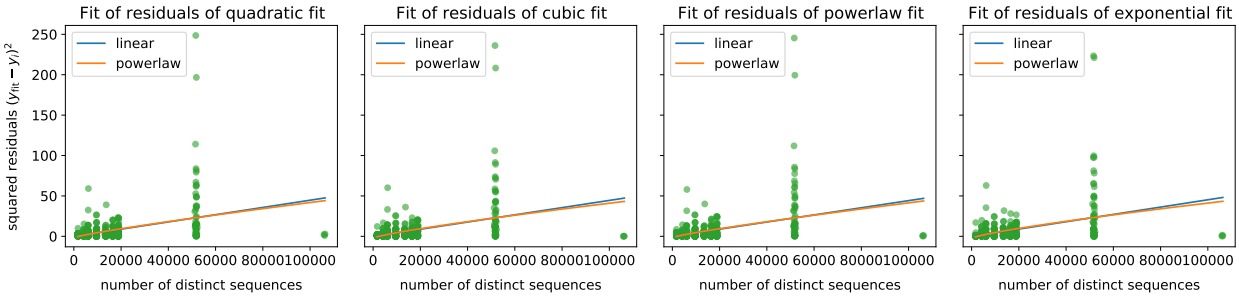


scenario III: $g = 0, \mu = 1.25E-7$

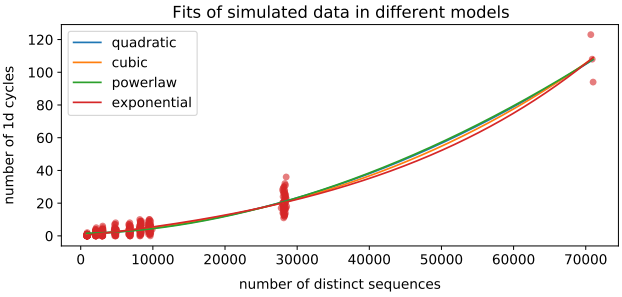


Log-Likelihood of validation data

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

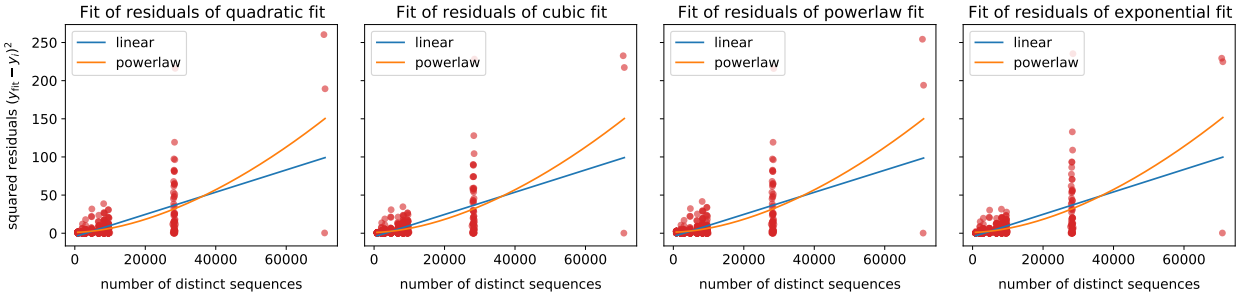


scenario IV: $g = 2, \mu = 0.75E-7$

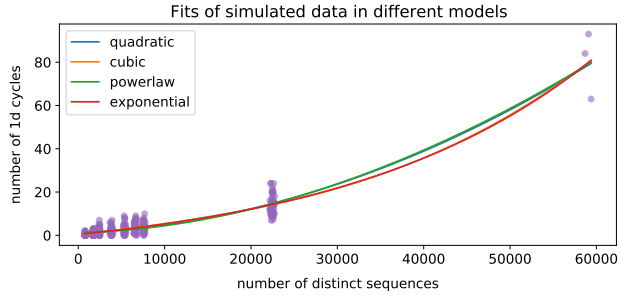


Log-Likelihood of validation data

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



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