Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

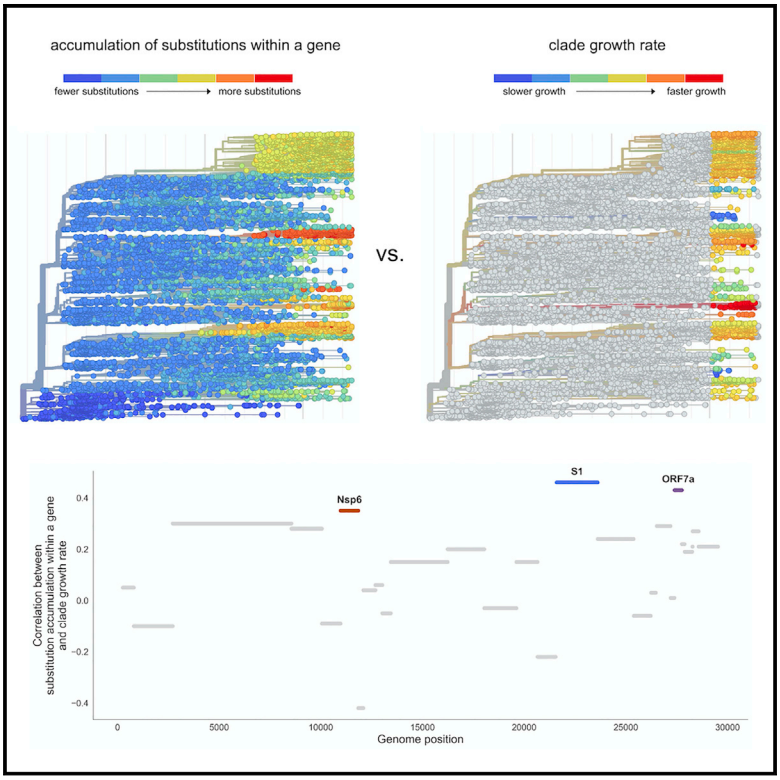
Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.

Article

Cell Host & Microbe

Rapid and parallel adaptive mutations in spike S1 drive clade success in SARS-CoV-2

Graphical abstract



Highlights

- SARS-CoV-2 is evolving adaptively in multiple regions of the genome
- The spike S1 subunit is the focus of adaptive evolution on a per-site basis
- S1 adaptation in the first 1.5 years was 2.5× greater than in the 2009 H1N1 pandemic

Authors

Kathryn E. Kistler, John Huddleston, Trevor Bedford

Correspondence

kkistler@fredhutch.org

In brief

Kistler et al. show that SARS-CoV-2 is undergoing adaptive evolution in multiple regions of the genome. Adaptive changes in spike S1 cluster temporally and accumulate rapidly, resulting in a high ratio of nonsynonymous to synonymous divergence. The extent of adaptation in S1 suggests that the virus might undergo antigenic drift.

Kistler et al., 2022, Cell Host & Microbe *30*, 545–555
April 13, 2022 © 2022 Elsevier Inc.
https://doi.org/10.1016/j.chom.2022.03.018

CellPress

Cell Host & Microbe

CellPress

Article

Rapid and parallel adaptive mutations in spike S1 drive clade success in SARS-CoV-2

Kathryn E. Kistler,[1,2,4,*] John Huddleston,[1] and Trevor Bedford[1,2,3]
[1]Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA
[2]Molecular and Cellular Biology Program, University of Washington, Seattle, WA, USA
[3]Howard Hughes Medical Institute, Seattle, WA, USA
[4]Lead contact
*Correspondence: kkistler@fredhutch.org
https://doi.org/10.1016/j.chom.2022.03.018

SUMMARY

The SARS-CoV-2 pandemic has resulted in numerous virus variants, some of which have altered receptor-binding or antigenic phenotypes. Here, we quantify the degree to which adaptive evolution is driving the accumulation of mutations across the genome. We correlate clade growth with mutation accumulation, compare rates of nonsynonymous to synonymous divergence, assess temporal clustering of mutations, and evaluate the evolutionary success of individual mutations. We find that spike S1 is the focus of adaptive evolution but also identify positively selected mutations in other proteins (notably Nsp6) that are sculpting the evolutionary trajectory of SARS-CoV-2. Adaptive changes in S1 accumulated rapidly, resulting in a remarkably high ratio of nonsynonymous to synonymous divergence that is 2.5× greater than that observed in influenza hemagglutinin HA1 at the beginning of the 2009 H1N1 pandemic. These findings uncover a high degree of adaptation in S1 and suggest that SARS-CoV-2 might undergo antigenic drift.

INTRODUCTION

After 20 months of global circulation, basal lineages of SARS-CoV-2 have been almost completely replaced by derived, variant lineages. These lineages are classified by the World Health Organization as variants of concern (VOCs) or variants of interest (VOIs) on the basis of genetic, phenotypic, and epidemiological differences (Konings et al., 2021). The effort to track the spread of these variants (and of the pandemic in general) through genomic epidemiology has resulted in a massive corpus of sequenced viral genomes. In the GISAID EpiCoV database alone, there are 2.5 million sequences and counting as of the end of July 2021 (Shu and McCauley, 2017). This thorough sampling offers an opportunity to investigate the evolutionary dynamics of a virus as it entered a naive population, spread rampantly, and subsequently began to transmit through previously exposed hosts. Here, we are particularly interested in whether SARS-CoV-2 viruses show phylogenetic evidence of adaptive evolution during the first year and a half of transmission in humans.

Seasonal influenza and seasonal coronaviruses both exhibit continual adaptive evolution during endemic circulation in the human population. In the case of influenza H3N2, transmission through an exposed host population results in adaptive evolution within hemagglutinin (HA). The HA1 subunit of HA both mediates binding to host cell receptors and is the primary target for neutralizing antibodies. Thus, in the context of an exposed host, selection for receptor binding avidity (Hensley et al., 2009) and for escape from humoral immunity (Bedford et al., 2014) drives the fixation of mutations in the HA1 subunit. The coronavirus protein subunit equivalent in function to HA1 is spike S1. Previously, we showed that at least two seasonal coronaviruses (229E and OC43) exhibit adaptive evolution concentrated in the S1 subunit of spike (Kistler and Bedford, 2021). By demonstrating that strong immune responses to a particular historical isolate of 229E do not neutralize 229E viruses that circulate years afterward, Eguia et al. confirmed that 229E evolves antigenically (Eguia et al., 2021).

Standard methods used for detecting adaptive evolution in seasonal influenza and seasonal coronaviruses rely on the fixation (or near fixation) of nonsynonymous changes and thus require years or decades of evolutionary time. These methods are ill fit to identify early adaptive evolution of a virus that has experienced a recent spillover event. For example, the common ancestor of globally circulating SARS-CoV-2 viruses is currently no earlier than January 2020, corresponding to the base of clade 20A or lineage B.1 (https://nextstrain.org/ncov/gisaid/global). Here, we present a method for identifying genomic regions undergoing adaptive evolution, which is well suited to early time points. This method correlates clade success with the accumulation of protein-coding changes in certain genes. Focusing on the period of VOC and VOI emergence, we apply this method to SARS-CoV-2 genomic data from December 2019 to May 2021.

With this method, we aim to present a rigorous quantification of the evolutionary process during this time and to show that

the observed success of variant viruses is a result of adaptive rather than neutral evolution. We conduct these analyses across the SARS-CoV-2 genome to identify foci of adaptive evolution. We complement these results with analyses of d_N/d_S (ratio of nonsynonymous to synonymous divergence) accumulation, evolutionary dynamics, and convergent evolution to provide evidence that genetic changes are contributing to viral fitness and identify genomic regions that are responsible.

RESULTS

Accumulation of nonsynonymous mutations in spike S1 correlates with clade success

RNA viruses are known for their remarkably high error rates and thus the rapid generation of mutations. Despite possessing some proof-reading capacity (a relatively rare function for an RNA virus), SARS-CoV-2 has been accumulating roughly 24–25 substitutions per year (https://nextstrain.org/groups/blab/ncov/adaptive-evolution/2021-05-15; Hadfield et al., 2018). The null hypothesis is that these substitutions reflect neutral evolution: the result of genetic drift acting on random mutations. To determine whether this is true, or whether adaptive evolution is also contributing to the accumulation of mutations, we started by comparing substitution rates in different regions of the genome.

We built a time-resolved phylogeny with a balanced geographic and temporal distribution of samples, including 9,544 viruses, collected between December 2019 and May 15, 2021 (Figure S1). For every internal branch on the phylogeny, we tallied the total number of mutations that occurred between the phylogeny root and that branch. We grouped deletion events with nonsynonymous single-nucleotide polymorphisms (SNPs) because they are protein changing and contribute to the evolution of some regions of the genome (Figure S2). Plotting mutation counts over time shows that spike S1 accumulates nonsynonymous changes at a rate of 8.4×10^{-3} substitutions/codon/year or about 5.5 substitutions/year (Figure 1A). This is a disproportionate percentage of the genome-wide estimate of 24 substitutions/year. As a control, we counted S1 synonymous mutations and found that they accumulate at 2.0×10^{-4} substitutions/codon/year, close to the naive expectation from the base composition that 22% of mutations should be synonymous. The per-codon rate of nonsynonymous mutation in S1 is roughly 17× higher than in RNA-dependent RNA polymerase (RdRp).

We hypothesize that adaptive evolution is driving the high rate of S1 nonsynonymous substitutions relative to S1 synonymous substitutions and RdRp nonsynonymous substitutions. And, although each S1 substitution will have a different effect on fitness, this observation suggests that this class of mutations is, on average, under positive selection. If this is the case, we would expect a correlation between S1 substitutions and a clade's evolutionary success: clades that happened to accumulate more S1 substitutions should have, on average, higher fitness (and hence faster growth in frequency) than clades that have accumulated fewer S1 substitutions. On the basis of this logic, we introduce a method for detecting adaptive evolution, which looks for genomic regions where mutation accumulation is associated with clade frequency growth. Because positive selection causes alleles or clades to increase in frequency in a logistic (rather than linear) fashion, we measure logistic growth rate and plot this against mutation accumulation.

Clade success and the number of nonsynonymous S1 mutations are positively correlated with a correlation coefficient r of 0.46 (Figure 1B). To test whether this correlation is greater than expected, we randomized the placement of mutations across branches of the phylogeny and computed a p value between the empirical r and the distribution of r values from 1,000 randomizations. The positive correlation between S1 mutations and logistic growth rate is statistically significant in comparison with the expected distribution ($p = 0.003$) but is absent for S1 synonymous mutations and is not significant for RdRp substitutions ($p = 0.256$) (Figure 1C).

We applied this method to every protein encoded by the SARS-CoV-2 genome (Table 1 and Figure S3A). The highest nonsynonymous mutation rate is observed in ORF8. However, ORF8 substitutions are not correlated with clade success (Table 1), and many lineages acquire premature stop codons in ORF8, indicating that the high rate of ORF8 substitutions is most likely due, at least in part, to a lack of functional constraints. Mutations within other regions of the genome, including spike S2 and nucleocapsid (N), also accumulate at reasonably high levels but do not correlate well with clade success (Table 1). Besides S1, only Nsp6 ($r = 0.35$, $p = 0.011$) and ORF7a ($r = 0.43$, $p < 0.001$) have a strong correlation with clade growth rates (Table 1 and Figure S3).

Although ORF7a substitutions appear highly correlated with clade success, this correlation is driven solely by the rapidly growing Delta variant, which possesses three mutations in ORF7a. Removing Delta clades from the analysis drops the r for ORF7a from 0.43 to 0.16, whereas the r values for S1 and Nsp6 dip from 0.46 to only 0.41 and from 0.35 to 0.32, respectively. This indicates that the correlation between S1 and Nsp6 substitutions and clade success is a general feature of SARS-CoV-2 lineages. Thus, the metric presented here provides evidence that SARS-CoV-2 is evolving adaptively and that the predominant locus of this evolution is spike S1.

The ratio of nonsynonymous to synonymous divergence is highest in S1

A classical method for assessing the average directionality of natural selection on some region of the genome is d_N/d_S, measuring the divergence of nonsynonymous sites relative to synonymous sites. A d_N/d_S value less than 1 indicates that the region is, on average, under purifying selection, whereas a d_N/d_S greater than 1 indicates positive selection on the region. Because even the most rapidly evolving proteins are still subject to structural and functional constraints, it is rare for an entire protein to have a d_N/d_S ratio greater than 1. For instance, the HA1 subunit of H3N2, which is the prototypical example of an adaptively evolving viral protein, has a d_N/d_S of 0.37 (Wolf et al., 2006).

For various regions of the SARS-CoV-2 genome, we computed the accumulation of nonsynonymous and synonymous divergence in 2-month windows between January 1, 2020, and May 15, 2021 (Figure 2A). This measures the d_N/d_S of branches leading to tips sampled within each 2-month window and captures the progressive enrichment of mutations by natural selection, i.e., mutations that persist and contribute to

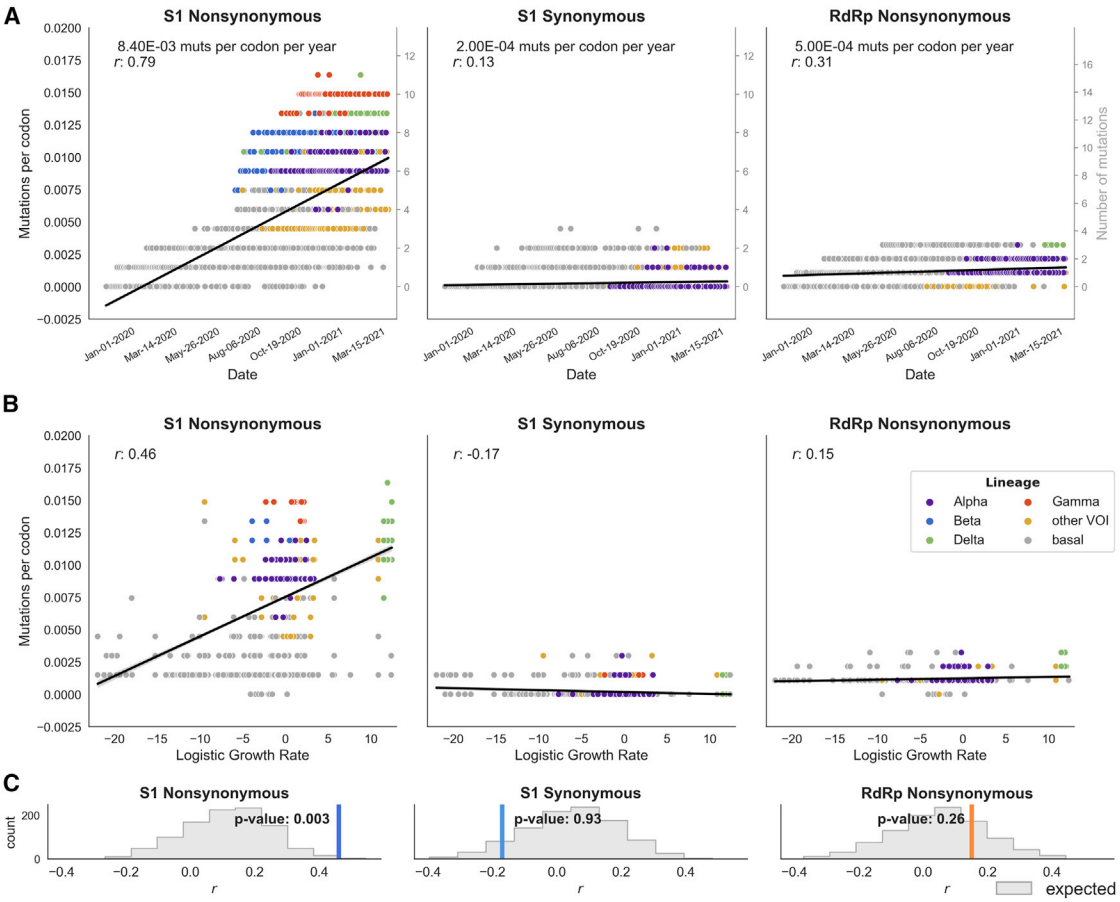


Figure 1. Accumulation of nonsynonymous S1 mutations is correlated with clade success

(A) For every clade in the phylogeny, mutations relative to the root of the phylogeny are tallied and plotted against the date of the base of that clade. Nonsynonymous S1, synonymous S1, and nonsynonymous RdRp mutations are plotted separately. Nonsynonymous mutations include nonsynonymous SNPs and deletions. The primary axis (left, black ticks) displays mutations per codon, and the secondary axis (right, gray ticks) shows the absolute number of mutations accumulated in each clade. Each point is colored according to the lineage it belongs to. Points are fit by linear regression.

(B) For every clade, mutation accumulation (as in A) is plotted against logistic growth rate, and the points are fit by linear regression.

(C) The empirical correlation coefficient *r* between mutation accumulation and logistic growth rate (colored bar) is compared with an expected distribution (gray) to yield a p value. Expected *r* values are determined from randomizing mutations across the phylogeny by using a multinomial draw with mutation likelihood proportional to relative branch length. The results of 1,000 iterations are shown.

the viral population will be captured in this measure, whereas mutations that die out will be excluded. The d_N/d_S ratio within RdRp, S2, and the structural proteins Envelope (E), Membrane (M), and Nucleocapsid (N) is consistently under 1 at all time points (Figure 2B). However, the d_N/d_S within S1 increases over time: there was an apparent inflection point in mid-2020, the d_N/d_S ratio exceeded 1 in late 2020 and 2021, and the most recent time point measured at 1.80. For comparison, we used the same methodology to compute d_N/d_S for influenza H3N2, influenza H1N1pdm, and seasonal coronavirus OC43 from 2009 to 2021 (Figure S4). We observe that after the 2009 emergence of influenza H1N1pdm in humans, the d_N/d_S in HA1 subunit peaked at 0.72 roughly a year after the beginning of that pandemic and declined in the following 4 years, whereas endemic viruses H3N2 and OC43 showed relatively stable d_N/d_S over this same time period.

The increase over time in SARS-CoV-2 S1 d_N/d_S could be due to a variety of reasons. Two non-mutually exclusive hypotheses include (1) the appearance of a new selective pressure on S1 substitutions and (2) the acquisition of mutations that change the mutational landscape to be more permissive toward S1 substitutions. Regardless of the cause, this change suggests a temporal structure to the adaptive evolution in the S1 subunit of SARS-CoV-2.

Nonsynonymous mutations in spike S1 cluster temporally

We can see a hint of this temporal structure by tracing individual mutational paths through the tree from root to tip. Figure S5 plots the accumulation of nonsynonymous S1 mutations along ten representative paths, leading to ten different emerging lineages (Rambaut et al., 2020). Along each of these paths, there appears to be an initial period of relative quiescence followed by a burst of S1 substitutions. To test whether this temporal clustering of mutations differs from what would be expected given the phylogenetic topology and the total number of observed S1

Table 1. Genome-wide correlation between nonsynonymous mutation accumulation and logistic growth rate

Protein	Nonsynonymous evolution rate (substitutions/codon/year)[a]	r[b]	p Value
Nsp1	0.1×10^{-3}	0.05	0.431
Nsp2	0.2×10^{-3}	−0.10	0.881
Nsp3	1.3×10^{-3}	0.30	0.083
Nsp4	0.8×10^{-3}	0.28	0.057
Nsp5	0.5×10^{-3}	−0.09	0.833
Nsp6	3.4×10^{-3}	0.35	0.011[c]
Nsp7	0.3×10^{-3}	−0.42	0.998
Nsp8	0.3×10^{-3}	0.04	0.443
Nsp9	0.7×10^{-3}	0.06	0.299
Nsp10	0.1×10^{-3}	−0.05	0.807
RdRp	0.5×10^{-3}	0.15	0.256
Nsp13	0.8×10^{-3}	0.20	0.153
Nsp14	0.3×10^{-3}	−0.03	0.698
Nsp15	0.3×10^{-3}	0.15	0.230
Nsp16	0.3×10^{-3}	−0.22	0.963
S1	8.4×10^{-3}	0.46	0.003[c]
S2	3.5×10^{-3}	0.24	0.105
ORF3a	1.8×10^{-3}	−0.06	0.865
E	1.6×10^{-3}	0.03	0.388
M	0.7×10^{-3}	0.29	0.045
ORF6	0.2×10^{-3}	0.01	0.528
ORF7a	1.5×10^{-3}	0.43	<0.001[c]
ORF7b	1.7×10^{-3}	0.22	0.050
ORF8	16.5×10^{-3}	0.19	0.185
N	6.10×10^{-3}	0.21	0.222
ORF9b	2.10×10^{-3}	0.27	0.050

[a]Evolutionary rate of nonsynonymous substitutions (and deletions) per codon per year.
[b]Correlation coefficient of mutation accumulation with logistic growth rate.
[c]Strong correlation. $p < 0.025$

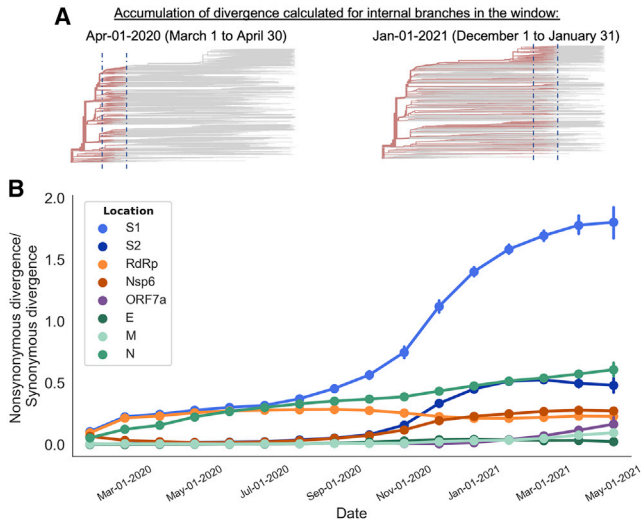


Figure 2. Ratio of nonsynonymous to synonymous divergence (d_N/d_S) is highest in S1

The d_N/d_S accumulation within various coding regions is calculated over time. The phylogeny of 9,544 viral sequences is divided into overlapping 2-month windows between January 1, 2020, and May 15, 2021.

(A) The phylogeny is colored to indicate the paths on which divergence accumulation is calculated for two of these time windows. The dashed blue lines indicate the time windows centered at April 1, 2020 (left), and January 1, 2021 (right). Every internal branch within these windows and the phylogenetic path that connects that branch back to the root is highlighted in red. The accumulation of divergence is calculated along these paths. Nonsynonymous divergence is calculated as the nonsynonymous Hamming distance between the sequence of an internal branch and the root sequence and normalized by the total possible number of nonsynonymous sites. The same is done for synonymous divergence.

(B) The d_N/d_S ratio is calculated for various coding regions within the genome. Each point shows the mean and 95% confidence interval of this ratio for all internal branches present in a 2-month window (centered at the date indicated on the x axis).

substitutions, we calculated wait times between mutations (diagrammed in Figure 3A). In brief, we created a null expectation by running 1,000 iterations of mutation randomization in which the phylogenetic placement of every observed mutation was shuffled. The distribution of wait times is dependent on tree topology and the total number of mutations, so the expectation is different for each category of mutations (Figure S6).

If mutations are clustered, there should be an excess of short wait times in the empirical data relative to the expectation. This is what we observe for S1 nonsynonymous mutations, where the distribution of wait times is left skewed with an overabundance of short wait times compared with the expected distribution (Figure 3B). The mean wait time between observed S1 substitutions is significantly lower than the expected mean wait time ($p < 0.001$), but there is no significant difference for S1 synonymous or RdRp wait times (Figure 3Ci). This difference is driven by short wait times because there is a significant difference between the proportion of observed versus expected wait times under 0.3 years for S1 nonsynonymous, but not S1 synonymous or RdRp, mutations (Figure 3Cii). These results indicate a temporal structure to the adaptive evolution of SARS-CoV-2 within the S1 subunit, which is characterized by mutation clustering.

Specific mutations associated with successful clades

We next sought to identify specific adaptive mutations throughout the genome. We note that convergent evolution is a good indicator of positive selection because each additional independent occurrence on the phylogeny of the mutation is increasingly unlikely under neutral evolution. As other groups have reported, many mutations shared by the VOCs have arisen via convergent evolution (van Dorp et al., 2020; Martin et al., 2021; Rochman et al., 2021). Here, we combine this observation of convergent evolution with logistic growth rate to find mutations that have arisen in the SARS-CoV-2 population multiple, independent times and expand into successful clades after each occurrence.

In this analysis, we focused on the evolutionary dynamics of SARS-CoV-2 during the period of time between the emergence of this virus in humans and mid-May 2021. We estimate that, during this period of time, VOC viruses were primarily competing with basal SARS-CoV-2 viruses. This allowed us to examine

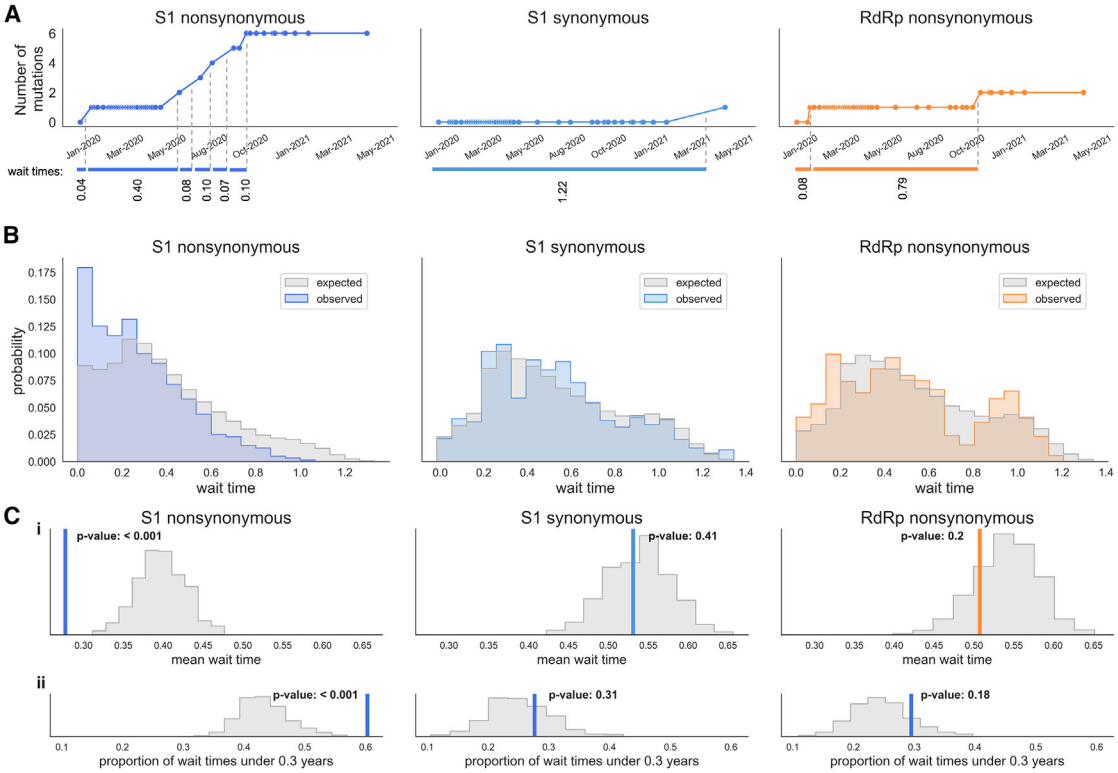


Figure 3. S1 substitutions are temporally clustered

(A) Timeline showing accumulation of S1 nonsynonymous, synonymous, and RdRp nonsynonymous mutations between the root and a representative tip (isolate USA/ME-HETL-J3202/2021); wait times between mutations are illustrated below. The exact date of a mutation is randomly selected along the branch that the mutation occurs on.

(B) Distribution of wait times between S1 nonsynonymous, S1 synonymous, and RdRp nonsynonymous mutations. Empirical wait times (in color) are plotted along with expected wait times (gray). Expected wait times are determined from randomizing mutations across the phylogeny by using a multinomial draw with mutation likelihood proportional to relative branch length. The results of 1,000 iterations are shown.

(C) (i) The mean empirical wait time from 1,000 iterations of the analysis (colored bar) is compared with the distribution of mean expected wait times (gray) to yield a p value. (ii) The proportion of observed wait times under 0.3 years (colored bar) is compared with the distribution of expected wait times under 0.3 years (gray).

the overall fitness effects of specific mutations in viral lineages that are successful during this period of time. After May 2021, VOCs make up the majority of the global virus population, and similar analyses on later time points would speak to the relative competitiveness of the variants.

For every deletion and substitution observed on the phylogeny, we tallied the number of independent occurrences and found the mean logistic growth rate of all clades where this mutation occurred. We limited this analysis to internal branches with 15 or more descending samples to limit the influence of stochasticity and sequencing errors that often occur on terminal branches. As expected, the bulk (84%) of the mutations occurred just once. Roughly 4% of mutations arose four or more times, and the majority of these mutations were located in S1 (Figure 4A). For seven of these convergently evolved mutations, the mean growth rate was higher than the tree-wide average growth rate. For three of these mutations (S:95I, S:452R, and ORF1a:3675-3677del), the mean growth rate exceeded the 90th percentile of mean growth rates expected from a mutation that occurs the same number of times on a randomized tree (Figure 4B).

This analysis reveals influential mutations during a snapshot of time in the ongoing adaptive evolution of SARS-CoV-2. In mid-May 2021, the Delta variant was rising in frequency. Both S1 mutations that we identified as important drivers of adaptive evolution (S:95I and S:452R) are present in the Delta variant as well as a handful of other emerging lineages (Figure S7). The specific mutations identified by this analysis will vary over time and depend on a multitude of factors (genetic, epidemiological, and otherwise) that determine clade success. However, ORF1a:3675-3677del consistently appears as a top hit (Figures 4C and S8). Remarkably, this deletion, which ablates amino acids 106–108 of Nsp6, arose eight independent times, and emerging lineages descend from each branch that this deletion occurs on (Figure S7).

Because recombination is common in coronaviruses (Müller et al., 2022; Turkahia et al., 2021), we investigated the possibility that these eight occurrences of the ORF1a:3675-3677 deletion were due to recombination rather than convergent evolution. We considered all pairs of lineages containing this mutation as potential recombinants and compared informative mutations in the potential donor and acceptor. The closest informative

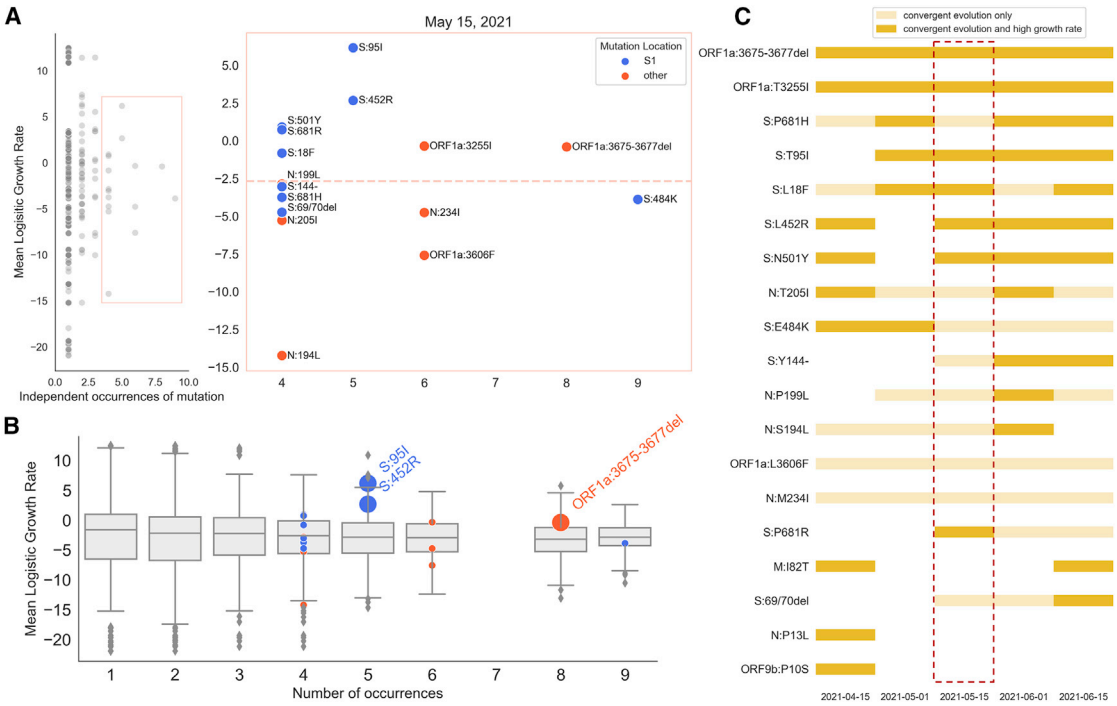


Figure 4. A 3-amino-acid deletion in Nsp6 displays convergent evolution and occurs in successful clades

(A) Every mutation observed on internal branches of the phylogeny is plotted according to the number of times this mutation occurs on the tree and the mean logistic growth rate of all clades it occurs in. Convergently evolved mutations that appear four or more times across the phylogeny are shown in the inset. The average growth rate of all clades is shown with a dotted line.

(B) Observed mean growth rates of convergently evolved mutations are compared with the mean growth rate expected for a mutation occurring on the phylogeny the same number of times. Convergently evolved mutations that have a mean growth rate falling at or above the 90th percentile of the expected values are labeled.

(C) The analysis shown in (A) was completed for five time points spanning 2 months. Each date represents the maximum date of sequences included in the analysis. Mutations that occur at least four times (convergent mutations) and result in a higher-than-average mean growth rate are shown in dark yellow. Mutations that display convergent evolution but do not result in high growth rates are in light yellow. The primary analysis was done at time point 2021-05-15 (outlined in red).

mutations flanking ORF1a:3675-3677del are not shared by any pairs of lineages, offering a lack of evidence for recombination and strong support for convergent evolution.

A 3-amino-acid deletion in Nsp6 is associated with accumulation of S1 substitutions

The ORF1a:3675-3677 deletion in Nsp6 exhibits striking convergent evolution and consistently precedes successful viral lineages. Because we have shown that accumulation of the S1 mutation is also associated with clade success, we next asked whether there is a relationship between clades with ORF1a:3675-3677del and a high number of S1 substitutions.

We created an expectation for the mean number of S1 mutations that should be observed in clades with ORF1a:3675-3677del by generating 100 randomized trees where the mutation occurred on eight branches selected by a multinomial draw. To make the expectation as fair as possible, we constrained the randomized branches to be on or after the date when the first Nsp6 deletion was observed. Under this expectation, there is no difference between the mean number of S1 or RdRp substitutions in clades that have the ORF1a:3675-3677 deletion and clades that do not (Figure 5A, left). However, in the empirical phylogeny, there are significantly more S1 substitutions in clades with the Nsp6 deletion than in clades without it (Figure 5A, right).

That clades with ORF1a:3675-3677del have higher numbers of S1 substitutions does not speak to the directionality of this relationship. In other words, it is possible that ORF1a:3675-3677del occurs in lineages that already have a lot of S1 substitutions or that a lot of S1 mutations accumulate in clades that already have ORF1a:3675-3677del. To determine the directionality of this difference, we considered every phylogenetic path that contains the Nsp6 deletion and found the difference between the final number of S1 substitutions on that path and the number of S1 substitutions that had accumulated before the deletion. On average, around 2.5 S1 nonsynonymous mutations accumulated after ORF1a:3675-3677del (Figure 5B). This is the second largest increase in S1 mutation accumulation following any convergently evolved mutation, behind S:681R. These results do not indicate that the deletion directly causes S1 substitutions, but they do add to the observations of convergent evolution and high clade growth rates in suggesting that ORF1a:3675-3677del is an adaptive mutation and an influential factor in the evolution of SARS-CoV-2.

DISCUSSION

Detecting adaptive evolution is both highly interesting from a basic scientific perspective as we seek to understand how and

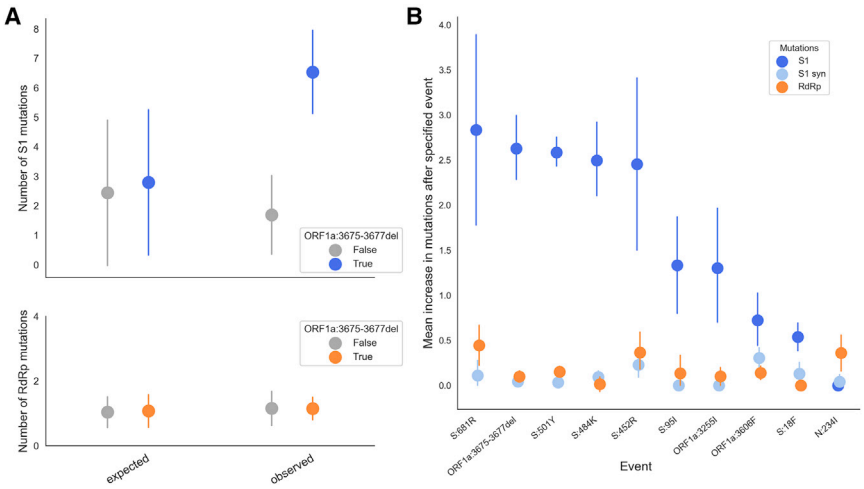


Figure 5. Clades with the 3-amino-acid deletion in Nsp6 have a high number of S1 mutations

(A) The mean number of S1 mutations (top) or RdRp mutations (bottom) that occur in clades that have (blue or orange) or do not have (gray) the 3-amino-acid deletion in Nsp6. The expected difference is shown on the left, and empirical data are shown on the right. Expectation is based on 100 randomizations of the tree. Error bars show the standard deviation.

(B) The difference in the number of nonsynonymous S1 (dark blue), S1 synonymous (light blue), and nonsynonymous RdRp (orange) mutations that occur before or after a convergently evolved mutation is shown. Error bars show 95% confidence intervals.

when this type of evolution occurs and highly relevant from a public health perspective as we strive to curb the transmission of infectious diseases. As widespread SARS-CoV-2 circulation continues, our best defense is through vaccination. The SARS-CoV-2 vaccines showed high efficacy in clinical trials, but we must be proactive to ensure their continued effectiveness. Vaccines against viruses that undergo antigenic drift, such as influenza, must be continually updated to match circulating variants. Therefore, the propensity of SARS-CoV-2 to evolve adaptively in spike S1 (the location of most neutralizing antibody epitopes) has important bearing on whether the SARS-CoV-2 vaccine will also need to be regularly updated.

SARS-CoV-2 exhibits convergent evolution (van Dorp et al., 2020; Martin et al., 2021; Rochman et al., 2021), and some of the notable mutations that have occurred multiple times independently (e.g., S:501Y and S:484K) appear in multiple VOCs, suggesting positive selection on these mutations. In the context of deep mutational scanning (DMS) experiments, mutations at 501 increase ACE2 binding affinity (Starr et al., 2020), and mutation to site 484 escapes antibody binding (Greaney et al., 2021). Recurrent mutations at S:681 enhance S1/S2 subunit cleavage (Liu et al., 2021; Lubinski et al., 2021), a protein modification that is essential for spike-mediated cell entry (Hoffmann et al., 2020) and thus is thought to contribute to increased viral replication (Liu et al., 2021). Many other convergently evolved mutations are also shared by VOCs and possess demonstrably different phenotypes, often altering antigenicity (Li et al., 2020; McCarthy et al., 2021; Wang et al., 2021).

Despite the demonstrably advantageous effects of observed mutations, it is too soon evolutionarily to pick up strong signals of adaptive evolution by classical methods that rely on the fixation of nonsynonymous mutations. Instead, we capitalize on the high temporal and geographic density of SARS-CoV-2 sequencing data to create a method for identifying adaptive evolution and genomic regions where this evolution is localized. This method identifies genes where amino acid substitutions significantly correlate with the clade growth rate. This can be intuitively interpreted as genes with high rates of amino acid substitutions (suggestive of positive selection) that result in more successful viruses (suggestive of a positive fitness effect) are undergoing adaptive evolution and is effectively a continuous analog to partitioning differences between polymorphism and divergence in the classical McDonald-Kreitman test (McDonald and Kreitman, 1991). We find that the spike S1 subunit shows strong signals of adaptive evolution by this method (Figure 1).

Our inference of adaptive evolution is based on a correlation between the accumulation of S1 substitutions and clade success that falls well outside the null expectation (Figure 1C). It is important to emphasize that these results speak to the average evolutionary effect of S1 substitutions. This does not mean that every S1 substitution is selectively advantageous, and it is likely that some mutations have larger effects on fitness than others. In fact, it is possible that successful viruses contain some S1 substitutions that do not contribute to their evolutionary success. One possibility is that these mutations could have arisen during long-term infections where they were advantageous within a single host. For instance, S1 mutations 484K and 501Y have been observed to arise from continued evolution within a single host (Choi et al., 2020). It is therefore possible that the parallel evolution of these particular mutations is due to a selective advantage at a within-host rather than between-host level. Within-host selection pressures could help to explain why some mutations such as 484K occur again and again across the phylogeny (Figure 4). However, the context in which S1 mutations arose does not affect our finding that viruses with more nonsynonymous mutations in S1 are more successful, on average, within the global population of SARS-CoV-2 viruses.

Phylogenetic inferences of evolution can be biased by the samples included in the analysis. To reduce sampling biases, our study is based on a phylogeny of 9544 SARS-CoV-2 genomes sampled evenly over space and time. The strong correlation between S1 mutation accumulation and clade growth rate persists if the number of genomes included in the phylogeny is doubled (Figure S9), indicating that our results are not biased by the number of samples included in the analysis. We also find that global adaptive evolution in S1 is not driven solely by certain geographic regions. Using phylogenies that include only samples from a particular geographic region, we observe that clade success strongly correlates with S1 substitutions in Asia, Europe, North America, Oceania, and South America (Figure S10). The only region where this correlation is not observed

at this time point is Africa, where the declining frequency of a particular clade of Beta drives an overall lack of correlation.

In addition to sampling biases, there are several limitations to our approach presented here. First, our analysis intentionally considers the average effect of mutations in different regions of the genome on viral fitness with the goal of taking a population-genetics approach to quantifying the adaptive evolution of SARS-CoV-2. This means that, although we observe a significant correlation, given a correlation coefficient of $r = 0.46$, our results cannot predict the fitness of specific variants solely on the basis of S1 mutation counts. Similarly, as mentioned above, this means that it is likely that some successful viral clades contain S1 substitutions that, rather than being advantageous, are hitchhiking along with positively selected mutations. Additionally, our analysis focuses on the period of VOC and VOI emergence from December 2019 to May 2021. So, although we can speculate on how our findings of high adaptive potential in S1 will translate to future evolution of the virus, we cannot directly predict how the pace of adaptive evolution will change over time.

We observe temporal structure in the adaptive evolution of SARS-CoV-2. We find that the correlation between clade success and S1 substitutions changes over time but showed strong signals of adaptive evolution from January to September 2021 (Figure S11). Enrichment of the d_N/d_S ratio in S1 also increases over time (Figure 2). Additionally, substitutions within S1 cluster temporally (Figure 3) rather than accrue at a steady rate. This temporal structure potentially indicates a changing evolutionary landscape either through the emergence of new selective pressure or through the occurrence of permissive mutations that made adaptive mutations more accessible.

Although the overall d_N/d_S ratio in S1 is 0.70, the d_N/d_S is 1.66 along persistent lineages in 2021 (Figure 2). This high ratio is remarkable in comparison with that of the antigenically evolving HA1 subunits of influenza H3N2 and H1N1pdm or the S1 subunit of seasonal coronavirus OC43 (Figure S4). We estimate the mean d_N/d_S ratio for HA1 in influenza H3N2 to be 0.36 (Figure S4), which is similar to the 0.37 estimated previously (Wolf et al., 2006). However, influenza H3N2 has been endemic in the human population for over 50 years, and its current evolution is largely driven by antigenic changes (Smith et al., 2004).

Viral evolution directly after spillover into a new host species can differ from evolution at endemicity. For instance, for 2 years following its emergence in 2009, the evolution of influenza H1N1pdm was thought to largely have been adaptation to a new host, whereas adaptive evolution after 2011 has been dominated by antigenic changes (Su et al., 2015). Accordingly, we observe that the d_N/d_S ratio in the H1N1pdm HA1 subunit peaked at 0.72 in 2010 shortly after pandemic emergence and then declined to a more stable value of ~0.3 beginning in 2014. An initially high rate of protein-coding changes is consistent with the idea that, soon after a spillover event, many evolutionarily accessible mutations are advantageous in the new host environment. It is unclear whether the observed d_N/d_S ratio in SARS-CoV-2 S1 will persist or whether it is a feature of this virus's recent emergence and will drop in the years to come.

Together, the results presented in Figures 1, 2, and 3 offer phylogenetic evidence that SARS-CoV-2 is evolving adaptively and that the primary locus of this adaptation is in S1. Our results are consistent with experimental demonstration of phenotypic changes conferred by VOC spike mutations (Greaney et al., 2021; Liu et al., 2021; McCarthy et al., 2021; Wang et al., 2021). Adaptive evolution in the S1 subunit during the period we focus on (December 2019 to May 2021) is most likely driven by selection to adapt to a new host by increasing infectivity of human cells. However, the amount of immunity to SARS-CoV-2 is rising globally, increasing the selection for antibody escape. Given the virus's demonstrated propensity for adaptive change in S1, antigenic drift will most likely begin to sculpt the evolution of SARS-CoV-2. The potential antigenic impact of adaptive S1 mutations, which are accruing at pace over four times that of influenza H3N2 (Figures 2 and S4), suggests that it might become necessary to update the SARS-CoV-2 vaccine strain. Indeed, the emergence of the Omicron VOC demonstrated a SARS-CoV-2 virus with an extraordinarily high number of S1 substitutions (Viana et al., 2022) that spread rapidly across the world and showed significantly lower neutralization titers than preceding variants (Cele et al., 2022). With Omicron, we now know that significant antigenic variants can emerge with highly modified S1 domains. However, the observed pace of adaptive evolution in S1 perhaps should have suggested the potential for emergence of such a variant.

Our results suggest that, in addition to S1, substitutions within Nsp6 and ORF7a could significantly contribute to the success of viral clades (Table 1). We expand on these gene-wide results by identifying specific adaptive mutations by using the confluence of convergent evolution and clade success. This analysis turned up many S1 mutations that have been extensively studied, along with mutations to nucleocapsid (N), another target of antibody-recognition (Kang et al., 2021), and a couple mutations in Nsp6, Nsp4, and M (Figure 4). The non-S1 mutations ORF1a:3255I (in Nsp4), M:82T, and N:205I in particular show compelling evidence of positive selection. These sites enrich our understanding from gene-wide analyses presented in Figures 1, 2, and 3 and Table 1: although S1 is the primary genomic locus of adaptive evolution, a handful of positively selected mutations in other genes are also influencing the evolution of SARS-CoV-2 in the human population.

Our analysis of specific adaptive mutations suggests the possibility of differences between within-host selection for viral replication and between-host selection for transmission. Viruses belonging to Delta have shown greater between-host transmission rates than other VOC or VOI viruses (Campbell et al., 2021) but are lacking mutations that have occurred repeatedly and that were associated with increased clade growth (notably ORF1a:3675-3677del, S:484K, and S:501Y). It is possible that some mutations display a large degree of parallelism as a result of specific within-host pressures that occur in secondary infections of partially immune individuals despite having only modest effects on between-host transmission.

It is important to note that the precise mutations that appear most influential depend on when the analysis is done (Figures 4C and S9). The fitness effect of a mutation is not an absolute quality—it depends on a multitude of influences, including the genetic background of the viral lineage, other co-circulating lineages, existing host immunity, and epidemiological factors (such as geographically heterogeneous mitigation efforts).

Additionally, lineages can grow in frequency as a result of stochastic effects. Therefore, it is expected that mutations associated with successful clades will change over time and that these changes reflect both a changing fitness landscape and the stochastic nature of evolution. Mutations that transcend this or, in other words, are associated with successful lineages at multiple time points are more likely to have important, adaptive functions. One such mutation is ORF1a:3675-3677del (Figures 4C and S9).

The ORF1a:3675-3677 deletion removes three amino acids (SGF) from a predicted transmembrane loop (Benvenuto et al., 2020) of the Nsp6 protein. Across the coronavirus family, the Nsp6 protein, in coordination with Nsp3 and Nsp4, forms double-membrane vesicles that are sites for viral RNA synthesis (Snijder et al., 2020). In SARS-CoV-2, Nsp6 suppresses the interferon-I response (Xia et al., 2020). It is unclear whether ORF1a:3675-3677del affects either of these functions. This deletion is not observed in other sarbecoviruses, residues 3675 and 3676 are 100% conserved, and only synonymous and conservative changes are seen at 3677 in this subgenus (Jungreis et al., 2021). However, in SARS-CoV-2, this deletion exhibits close to the highest level of convergence, presence in VOCs, mean logistic growth rate, and increase in S1 mutations in descending lineages. Future experimental study of this deletion would increase our understanding of what functions, apart from enhanced cell entry and potential antibody escape, were highly advantageous during the early adaptive evolution of SARS-CoV-2.

So far, ORF1a:3675-3677del has not been observed in Delta viruses, and our results suggest that the appearance of a Delta sublineage possessing ORF1a:3675-3677del could outcompete basal Delta viruses. However, the Omicron variant, which appeared in late November 2021 and spread rapidly, possesses a very similar deletion, where Omicron viruses from the primary BA.1 PANGO lineage exhibit ORF1a:3674-3676del. This provides yet another example of an association between deletion to this region of Nsp6 and consequential VOC viruses.

Limitations of this study

This study examined adaptive evolution during the period of December 2019 to May 2021. The results presented in this manuscript are therefore limited to this temporal range and do not directly predict how the pace of adaptive evolution will change in the future. In addition, the results presented in Figures 1 and 2 speak to the average evolutionary effect of S1 substitutions. As discussed above, this does not mean that every S1 substitution is selectively advantageous, and it is likely that some mutations have larger effects on fitness than others.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHODS DETAILS
  - Phylogenetic reconstruction of a subsampling of global SARS-CoV-2 genome sequences
  - Quantification of mutation accumulation
  - Estimation of the logistic growth rate of clades
  - Calculation of correlation between mutations and clade growth rate
  - Calculation of nonsynonymous to synonymous divergence ratio
  - Randomization of mutations across the phylogeny for wait time calculations
  - Calculation of wait times
  - Quantification of convergent evolution and logistic growth rates across the phylogeny
  - Randomization of recurrent mutations across the phylogeny
  - Consideration of recombination as an alternative to convergent evolution of Nsp6 deletion
  - Calculation of the mean number of S1 mutations per clade
  - Calculation of S1 mutations that precede and follow specific mutation events
  - Consideration of sampling biases
  - Analysis of duration of correlation between clade success and S1 substitutions
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Statistical analysis of correlation between growth rate and number of mutations
  - Statistical analysis of wait times
  - Statistical analysis of growth rates associated with convergent mutations
- ADDITIONAL RESOURCES

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.chom.2022.03.018.

ACKNOWLEDGMENTS

We thank members of the Bedford lab, Tami Lieberman, and two anonymous reviewers for helpful feedback on the manuscript. We acknowledge the authors for originating and submitting laboratories of the sequences from GISAID's EpiCoV database, on which this research is based. A full acknowledgments table is available as Table S1. T.B. is a Pew Biomedical Scholar. This work was supported by NIH R35 GM119774. K.E.K. was supported by the National Science Foundation Graduate Research Fellowship Program under grant no. DGE-1762114.

AUTHOR CONTRIBUTIONS

Conceptualization, K.E.K. and T.B.; methodology, K.E.K., J.H., and T.B.; software, K.E.K. and J.H.; formal analysis, K.E.K.; investigation, K.E.K.; writing – original draft, K.E.K.; writing – review & editing, K.E.K., J.H., and T.B.; visualization, K.E.K.; funding acquisition, K.E.K. and T.B.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: September 24, 2021
Revised: January 18, 2022
Accepted: March 14, 2022
Published: April 13, 2022

REFERENCES

Aksamentov, I., Roemer, C., Hodcroft, E., and Neher, R. (2021). Nextclade: Clade assignment, mutation calling and quality control for viral genomes. J. Open Source Softw. *6*, 3773.

Bedford, T., Suchard, M.A., Lemey, P., Dudas, G., Gregory, V., Hay, A.J., McCauley, J.W., Russell, C.A., Smith, D.J., and Rambaut, A. (2014). Integrating influenza antigenic dynamics with molecular evolution. eLife *3*, e01914.

Benvenuto, D., Angeletti, S., Giovanetti, M., Bianchi, M., Pascarella, S., Cauda, R., Ciccozzi, M., and Cassone, A. (2020). Evolutionary analysis of SARS-CoV-2: How mutation of non-structural protein 6 (NSP6) could affect viral autophagy. J. Infect. *81*, e24–e27.

Campbell, F., Archer, B., Laurenson-Schafer, H., Jinnai, Y., Konings, F., Batra, N., Pavlin, B., Vandemaele, K., Van Kerkhove, M.D., Jombart, T., et al. (2021). Increased transmissibility and global spread of SARS-CoV-2 variants of concern as at June 2021. Euro Surveill. *26*, 2100509.

Cele, S., Jackson, L., Khoury, D.S., Khan, K., Moyo-Gwete, T., Tegally, H., San, J.E., Cromer, D., Scheepers, C., Amoako, D.G., et al.; NGS-SA; COMMIT-KZN Team (2022). Omicron extensively but incompletely escapes Pfizer BNT162b2 neutralization. Nature *602*, 654–656.

Choi, B., Choudhary, M.C., Regan, J., Sparks, J.A., Padera, R.F., Qiu, X., Solomon, I.H., Kuo, H.-H., Boucau, J., Bowman, K., et al. (2020). Persistence and evolution of SARS-CoV-2 in an immunocompromised host. N. Engl. J. Med. *383*, 2291–2293.

Eguia, R.T., Crawford, K.H.D., Stevens-Ayers, T., Kelnhofer-Millevolte, L., Greninger, A.L., Englund, J.A., Boeckh, M.J., and Bloom, J.D. (2021). A human coronavirus evolves antigenically to escape antibody immunity. PLoS Pathog. *17*, e1009453.

Greaney, A.J., Loes, A.N., Crawford, K.H.D., Starr, T.N., Malone, K.D., Chu, H.Y., and Bloom, J.D. (2021). 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.

Hadfield, J., Megill, C., Bell, S.M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T., and Neher, R.A. (2018). Nextstrain: Real-time tracking of pathogen evolution. Bioinformatics *34*, 4121–4123.

Hensley, S.E., Das, S.R., Bailey, A.L., Schmidt, L.M., Hickman, H.D., Jayaraman, A., Viswanathan, K., Raman, R., Sasisekharan, R., Bennink, J.R., and Yewdell, J.W. (2009). Hemagglutinin receptor binding avidity drives influenza A virus antigenic drift. Science *326*, 734–736.

Hoffmann, M., Kleine-Weber, H., and Pöhlmann, S. (2020). A multibasic cleavage site in the spike protein of SARS-CoV-2 is essential for infection of human lung cells. Mol. Cell *78*, 779–784.e5.

Huddleston, J., Hadfield, J., Sibley, T.R., Lee, J., Fay, K., Ilcisin, M., Harkins, E., Bedford, T., Neher, R.A., and Hodcroft, E.B. (2021). Augur: A bioinformatics toolkit for phylogenetic analyses of human pathogens. J Open Source Softw *6*, 2906.

Jungreis, I., Sealfon, R., and Kellis, M. (2021). SARS-CoV-2 gene content and COVID-19 mutation impact by comparing 44 Sarbecovirus genomes. Nat. Commun. *12*, 2642.

Kang, S., Yang, M., He, S., Wang, Y., Chen, X., Chen, Y.-Q., Hong, Z., Liu, J., Jiang, G., Chen, Q., et al. (2021). A SARS-CoV-2 antibody curbs viral nucleocapsid protein-induced complement hyperactivation. Nat. Commun. *12*, 2697.

Kistler, K.E., and Bedford, T. (2021). Evidence for adaptive evolution in the receptor-binding domain of seasonal coronaviruses OC43 and 229e. eLife *10*, e64509.

Konings, F., Perkins, M.D., Kuhn, J.H., Pallen, M.J., Alm, E.J., Archer, B.N., Barakat, A., Bedford, T., Bhiman, J.N., Caly, L., et al. (2021). SARS-CoV-2 variants of interest and concern naming scheme conducive for global discourse. Nat. Microbiol. *6*, 821–823.

Li, Q., Wu, J., Nie, J., Zhang, L., Hao, H., Liu, S., Zhao, C., Zhang, Q., Liu, H., Nie, L., et al. (2020). The impact of mutations in SARS-CoV-2 spike on viral infectivity and antigenicity. Cell *182*, 1284–1294.e9.

Liu, Y., Liu, J., Johnson, B.A., Xia, H., Ku, Z., Schindewolf, C., Widen, S.G., An, Z., Weaver, S.C., Menachery, V.D., et al. (2021). Delta spike P681R mutation enhances SARS-CoV-2 fitness over Alpha variant. Preprint at bioRxiv. https://doi.org/10.1101/2021.08.12.456173.

Lubinski, B., Frazier, L.E.T., Phan, M.V.T., Bugembe, D.L., Tang, T., Daniel, S., Cotten, M., Jaimes, J.A., and Whittaker, G.R. (2021). Spike protein cleavage-activation mediated by the SARS-CoV-2 P681R mutation: A case-study from its first appearance in variant of interest (VOI) A.23.1 identified in Uganda. Preprint at bioRxiv. https://doi.org/10.1101/2021.06.30.450632.

Martin, D.P., Weaver, S., Tegally, H., San, J.E., Shank, S.D., Wilkinson, E., Lucaci, A.G., Giandhari, J., Naidoo, S., Pillay, Y., et al.; NGS-SA; COVID-19 Genomics UK (COG-UK) (2021). The emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages. Cell *184*, 5189–5200.e7.

McCarthy, K.R., Rennick, L.J., Nambulli, S., Robinson-McCarthy, L.R., Bain, W.G., Haidar, G., and Duprex, W.P. (2021). Recurrent deletions in the SARS-CoV-2 spike glycoprotein drive antibody escape. Science *371*, 1139–1142.

McDonald, J.H., and Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. Nature *351*, 652–654.

Müller, N.F., Kistler, K.E., and Bedford, T. (2022). Recombination patterns in coronaviruses. Preprint at bioRxiv. https://doi.org/10.1101/2021.04.28.441806.

Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. *32*, 268–274.

Rambaut, A., Holmes, E.C., O'Toole, Á., Hill, V., McCrone, J.T., Ruis, C., du Plessis, L., and Pybus, O.G. (2020). A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat. Microbiol. *5*, 1403–1407.

Rochman, N.D., Wolf, Y.I., Faure, G., Mutz, P., Zhang, F., and Koonin, E.V. (2021). Ongoing global and regional adaptive evolution of SARS-CoV-2. Proc. Natl. Acad. Sci. USA *118*, e2104241118.

Sagulenko, P., Puller, V., and Neher, R.A. (2018). TreeTime: Maximum-likelihood phylodynamic analysis. Virus Evol. *4*, vex042.

Shu, Y., and McCauley, J. (2017). GISAID: Global initiative on sharing all influenza data - from vision to reality. Euro Surveill. *22*, 30494.

Smith, D.J., Lapedes, A.S., de Jong, J.C., Bestebroer, T.M., Rimmelzwaan, G.F., Osterhaus, A.D.M.E., and Fouchier, R.A.M. (2004). Mapping the antigenic and genetic evolution of influenza virus. Science *305*, 371–376.

Snijder, E.J., Limpens, R.W.A.L., de Wilde, A.H., de Jong, A.W.M., Zevenhoven-Dobbe, J.C., Maier, H.J., Faas, F.F.G.A., Koster, A.J., and Bárcena, M. (2020). A unifying structural and functional model of the coronavirus replication organelle: Tracking down RNA synthesis. PLoS Biol. *18*, e3000715.

Starr, T.N., Greaney, A.J., Hilton, S.K., Ellis, D., Crawford, K.H.D., Dingens, A.S., Navarro, M.J., Bowen, J.E., Tortorici, M.A., Walls, A.C., et al. (2020). Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. Cell *182*, 1295–1310.e20.

Su, Y.C.F., Bahl, J., Joseph, U., Butt, K.M., Peck, H.A., Koay, E.S.C., Oon, L.L.E., Barr, I.G., Vijaykrishna, D., and Smith, G.J.D. (2015). Phylodynamics of H1N1/2009 influenza reveals the transition from host adaptation to immune-driven selection. Nat. Commun. *6*, 7952.

Turkahia, Y., Thornlow, B., Hinrichs, A., McBroome, J., Ayala, N., Ye, C., De Maio, N., Haussler, D., Lanfear, R., and Corbett-Detig, R. (2021). Pandemic-scale phylogenomics reveals elevated recombination rates in the SARS-CoV-2 spike region.

van Dorp, L., Acman, M., Richard, D., Shaw, L.P., Ford, C.E., Ormond, L., Owen, C.J., Pang, J., Tan, C.C.S., Boshier, F.A.T., et al. (2020). Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infect. Genet. Evol. *83*, 104351.

Viana, R., Moyo, S., Amoako, D.G., Tegally, H., Scheepers, C., Althaus, C.L., Anyaneji, U.J., Bester, P.A., Boni, M.F., Chand, M., et al. (2022). Rapid

epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa. Nature.

Wang, Z., Schmidt, F., Weisblum, Y., Muecksch, F., Barnes, C.O., Finkin, S., Schaefer-Babajew, D., Cipolla, M., Gaebler, C., Lieberman, J.A., et al. (2021). mRNA vaccine-elicited antibodies to SARS-CoV-2 and circulating variants. Nature *592*, 616–622.

Wolf, Y.I., Viboud, C., Holmes, E.C., Koonin, E.V., and Lipman, D.J. (2006). Long intervals of stasis punctuated by bursts of positive selection in the seasonal evolution of influenza A virus. Biol. Direct *1*, 34.

Xia, H., Cao, Z., Xie, X., Zhang, X., Chen, J.Y.-C., Wang, H., Menachery, V.D., Rajsbaum, R., and Shi, P.-Y. (2020). Evasion of type I interferon by SARS-CoV-2. Cell Rep. *33*, 108234.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
GISAID EpiCoV SARS-CoV-2 sequence isolates	GISAID	full list of contributing labs and accessions: Mendeley Data: 10.17632/mxjt9nhgmj.1
Software and algorithms		
Augur, version 12.0.0	Huddleston et al., 2021	https://github.com/nextstrain/augur
TreeTime, version 0.8.2	Sagulenko et al., 2018	https://github.com/neherlab/treetime
IQ-TREE, version 2.1.2	Nguyen et al., 2015	https://github.com/Cibiv/IQ-TREE
Nextclade	Aksamentov et al., 2021	https://clades.nextstrain.org
Custom code	this paper	https://github.com/blab/sarscov2-adaptive-evolution/

RESOURCE AVAILABILITY

Lead contact

Inquiries for further information should be directed to the lead contact, Kathryn Kistler (kkistler@fredhutch.org).

Materials availability

This study did not generate any new reagents.

Data and code availability

- This paper analyzes existing, publicly accessible data. The accession numbers are listed in Table S1, which is available from Mendeley Data: 10.17632/mxjt9nhgmj.1.
- Source code for all analyses presented in this manuscript is available at https://github.com/blab/sarscov2-adaptive-evolution. This repository has been archived at Zenodo: 10.5281/zenodo.6126495 and is publicly available as of the date of publication.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHODS DETAILS

Phylogenetic reconstruction of a subsampling of global SARS-CoV-2 genome sequences

All primary analyses in this manuscript were performed using data downloaded from the GISAID EpiCoV database (https://www.gisaid.org/; Shu and McCauley, 2017) on July 29, 2021 and curated by the Nextstrain nCoV ingest pipeline (https://github.com/nextstrain/ncov-ingest). This dataset contained 2,459,376 viral genomes and associated metadata. These genomes were aligned with Nextalign (https://docs.nextstrain.org/projects/nextclade/en/latest/user/nextalign-cli.html) and masked to minimize error in phylogenetic inference associated with problematic amplicon sites. Masked alignments were filtered to exclude strains that were known outliers, sequenced due to "S dropout," mis-annotated with an admin division of "USA," shorter than 27,000 bp of A, C, T, or G bases, missing complete date information, annotated with a date prior to October 2019, flagged with more than 20 mutations above the expected number based on the mutational clock rate, or flagged by Nextclade (https://docs.nextstrain.org/projects/nextclade/en/latest/user/algorithm/07-quality-control.html; Aksamentov et al., 2021) with one or more clusters of 6 or more private differences in a 100-nucleotide window. After filtering 2,213,085 genomes remained.

After filtering, SARS-CoV-2 genomes were evenly sampled across geographic scales and time. Specifically, a maximum of 1,600 strains were sampled from each continental region including Africa, Asia, Europe, North America, Oceania, and South America for an approximate total of 9,600 genomes per phylogeny. For each region except North America and Oceania, strains were sampled from each distinct combination of country, year, and month. For North America and Oceania, genomes were sampled from each distinct combination of division (i.e., state-level geography), year, and month.

Time-resolved phylogenies were inferred using Augur 12.0.0 (Huddleston et al., 2021), IQ-TREE 2.1.2 (Nguyen et al., 2015), and TreeTime 0.8.2 (Sagulenko et al., 2018). Ancestral sequences were inferred with TreeTime using the joint inference mode. The primary analysis was conducted on 9,544 genomes collected on or before May 15, 2021, and the phylogeny reconstructed from these data can be found at https://nextstrain.org/groups/blab/ncov/adaptive-evolution/2021-05-15. Phylogenies used for secondary

analyses of convergent evolution (Figures 4C and S9) can be viewed using the date drop-down menu in the left-hand sidebar. The secondary analyses included isolates sequenced up until April 15, 2021 (9,467 genomes), May 1, 2021 (9,449 genomes), June 1, 2021 (9,343 genomes), and June 15, 2021 (9,401 genomes). All isolates used in these analyses are listed in Table S1.

Quantification of mutation accumulation

For every internal branch on the phylogeny, the number of mutations that accumulated between the root of the tree and that branch was counted. For this and all subsequent analyses, deletions are grouped with nonsynonymous substitutions. Deletions that span multiple, adjacent amino acids are counted as one mutation. Mutations to a premature stop codon are also counted as one mutation event. Mutations were separated by which gene they occur in (according to the Wuhan-Hu-1 reference sequence, found at "analysis/reference_seq_edited.gb") and whether they are synonymous or nonsynonymous. Genomic locations of the 15 NSPs were found in the NC_045512.2 annotation of the ORF1ab polyprotein (http://www.ncbi.nlm.nih.gov/gene/43740578). Code for mutation accumulation counting and plotting of Figure 1A is found in "fig1-muts_by_time_and_growthrate.ipynb."

Estimation of the logistic growth rate of clades

Logistic growth of individual clades was estimated from the time-resolved phylogeny and the estimated frequencies for each strain in the tree. Frequencies were estimated with Augur 12.0.0 (Huddleston et al., 2021) using the KDE estimation method that creates a Gaussian distribution for each strain with a mean equal to the strain's collection date and a variance of 0.05 years. At weekly intervals, the frequencies of each strain at a given date were calculated by summing the corresponding values in their Gaussian distributions and normalizing the values to sum to 1. The frequency of each clade at a given time was the sum of its corresponding strain frequencies at that time.

Logistic growth was calculated for each clade in the phylogeny that was currently circulating at a frequency >0.0001% and < 95% and that had at least 50 descendant strains. Each clade's frequencies for the last 6 weeks were logit transformed and used as the dependent variable for a linear regression where the independent variable was the corresponding date value for each transformed frequency. The logistic growth of the clade was then annotated as the slope of the linear regression of the logit-transformed frequencies.

Calculation of correlation between mutations and clade growth rate

For internal branch on the phylogeny, the logistic growth rate was plotted against the number of nonsynonymous S1 mutations per codon. The correlation coefficient r for the linear regression was computed. This empirical r value is then compared to a distribution of 1,000 expected r values to yield a p value. The expected correlation coefficients were determined by stripping all observed nonsynonymous S1 mutations from the phylogeny and then randomly assigning each mutation to an internal branch of the phylogeny by a multinomial draw where the likelihood of a mutation occurring on a branch is proportional to the branch length. The was repeated for every gene in the genome. The code to compute the empirical r and p value for each gene and to reproduce Table 1 is at "table1-comprehensive_mut_accumulation_analysis.ipynb."

Calculation of nonsynonymous to synonymous divergence ratio

A time-course of d_N/d_S ratios was calculated in non-overlapping time windows by splitting all internal branches included in the phylogeny according to their date. Within each gene, the nonsynonymous and synonymous Hamming distances were found between the reference sequence and every internal branch. The Hamming distances were normalized by the total number of possible nonsynonymous or synonymous sites within that gene to give a measure of divergence. The nonsynonymous divergence was divided by synonymous divergence. Then, for each time window, the mean of this ratio was found for all internal branches within the window. For SARS-CoV-2, the time windows were 2 months and overlap by 1 month. The code to run this analysis and reproduce Figure 2 is at "fig2-divergence.ipynb."

Phylogenies for seasonal influenza A/H3N2 and A/H1N1pdm were built using the Nextstrain pipeline from https://github.com/nextstrain/seasonal-flu. They include 2,274 and 2,169 genomes, respectively, that were sampled between 2009 and 2021 to capture the earliest sequences from the H1N1pdm pandemic (March 2009). The OC43 phylogeny was built from all available OC43 lineage A genomes sampled in 2009 or later (214 genomes) using the workflow in https://github.com/blab/seasonal-cov-adaptive-evolution/tree/master/oc43/separate_lineages. Divergence accumulation ratios were computed from the root of each tree using 1-year time windows overlapping by 0.5 years. These phylogenies can be found in "seasonal-flu_trees/." The code in "fig2supp-divergence_seasonalflu.ipynb" reproduces Figure 2.

Randomization of mutations across the phylogeny for wait time calculations

For each type of mutation (S1 nonsynonymous, S1 synonymous, and RdRp nonsynonymous), the total number of mutations observed on the phylogeny was randomly scattered across phylogeny. Only internal branches with 3 or more descending tips were used. Random branches were selected by a multinomial draw, where the likelihood of a branch having a mutation is proportional to its branch length in years. Multiple mutations were allowed to occur on the same branch, just as with the empirical phylogeny. Randomizations were run 1,000 times for each mutation type used in Figures 3B and 3C and ten times for the distributions shown in Figure S6. Code for this analysis is in "fig3-wait_times.ipynb."

Calculation of wait times

Wait times were counted for the following classes of mutations: S1 nonsynonymous, S1 synonymous, and RdRp nonsynonymous. For each class of mutation, a wait time was calculated between each branch that has a mutation of this type and its first child branch on each descending path that has a mutation of this type. A wait time was also calculated between the tree root and the first branch on any independent path that has a mutation of this type. Conceptually, the result of this is that wait times are computed between every sequential mutation that occurs along every path on the tree (as diagrammed in Figure 3A), without double counting any pairs of branches. Only mutations on internal branches (defined as having three or more descending tips) are considered.

A wait time is simply the time between mutations and is calculated by subtracting the date (in decimal years) of the earlier mutation from the date of the later mutation. Because the exact date a mutation occurred cannot be known, each mutation is assigned a random date along the branch it occurred on. If multiple mutations of the same type occurred on one branch, each mutation is assigned a different random date and the wait times between mutations on that branch are calculated.

Empirical and expected wait times were calculated for each type of mutation 1,000 times and the results of all 1,000 iterations can be found in "wait_time_stats/." Code to calculate wait times and reproduce Figures 3B, 3C, and S6 is found in "fig3-wait_times.ipynb."

Quantification of convergent evolution and logistic growth rates across the phylogeny

Every substitution that occurred on an internal branch with at least 15 descending tips was tallied. For every substitution that was observed at least 4 times on internal branches, the average growth rate of clades containing this mutation was calculated by taking the mean logistic growth rate of clades where this mutation occurred. Code to count occurrences, calculate mean logistic growth, and determine which emerging lineages descend from recurrent mutations is found in "fig4-convergent_evolution.ipynb." This code will reproduce Figures 4A, S8, and S9.

Randomization of recurrent mutations across the phylogeny

One hundred randomized trees were created by shuffling the phylogenetic positions of each substitution that was observed on an internal branch with at least 15 descending tips (those calculated above and shown in Figure 4A). Randomized branches were also limited to internal branches with at least 15 descending tips. The position of each randomized substitution was constrained to branches that "make phylogenetic sense": meaning, a given substitution cannot occur twice on the same path. This results in a tree with exactly the same distribution of mutation occurrences as the empirical phylogeny, but where those mutations occur on different branches. Code to implement these randomizations and reproduce Figure 4B is in "fig4-convergent_evolution.ipynb."

Consideration of recombination as an alternative to convergent evolution of Nsp6 deletion

For each occurrence of the ORF1a:3675-3677 deletion, all nucleotide mutations that occurred between the root and the branch where the deletion occurred were recorded. Then, recombination between every pair of the 8 inferred occurrences of ORF1a:3675-3677del was considered. For each pair, informative mutations that did not occur in a common ancestor of the potential recombinant lineages were identified. The informative mutations closest to the Nsp6 deletion on the upstream side were compared between potential donor and acceptor (and the same was done for the downstream side). If the closest mutations were shared between any donor/acceptor pair, this would be evidence that this mutation and the Nsp6 deletion were transferred from the donor to the acceptor by recombination. If the closest mutations are not shared between the donor and acceptor, the only way the acceptor could have acquired the ORF1a:3675-3677del through recombination is if both recombination break points occurred within a genomic window defined by the closest informative mutations on either side of the Nsp6 deletion. Code for this analysis as well as a table summarizing the results is in "nsp6del_recombination.ipynb."

Calculation of the mean number of S1 mutations per clade

The phylogeny was divided into clades that have the ORF1a:3675-3677 deletion and those that do not, and the mean number of S1 and RdRp substitutions was computed for each category. The tree was limited to only branches occurring on or after the date of the first ORF1a:3675-3677del occurrence. The expectation was created by randomizing the locations of the eight occurrences of ORF1a:3675-3677del as was done above in "randomization of recurrent mutations across the phylogeny." Code for this analysis is in "fig5-nsp6del_s1mutations_correlation.ipynb."

Calculation of S1 mutations that precede and follow specific mutation events

For each convergently-evolved mutation, every path through the phylogeny containing this mutation was considered. The total number of S1 mutations accumulated between the root and the occurrence of the convergently-evolved mutation is considered to be the number of S1 mutations before the event. The number of mutations after is the final number of S1 mutations present on the path. The before total is subtracted from the after total to give the increase in S1 mutations after the event. The mean of this increase is calculated for every path containing the convergently-evolved mutation. Code to implement this analysis is in "fig5b-s1_muts_before_vs_after.ipynb."

Consideration of sampling biases

The impact of phylogeny size on the results presented in this manuscript was tested by running the same analyses on a phylogeny containing twice as many samples. This phylogeny was built according to the same methods described in the "phylogenetic reconstruction of a subsampling of global SARS-CoV-2 genome sequences" section above, except that a maximum of 3,300 strains were sampled from each continental region. This resulted in a tree with a total of 19,694 genomes, sampled on or before May 15, 2021 and distributed roughly evenly over geography and time (https://nextstrain.org/groups/blab/ncov/adaptive-evolution/2021-05-15/20k). The correlation between mutation accumulation in eight genes (or subunits) and clade success (Figure S9A) is done in "fig1supp-20ktree.ipynb." The d_N/d_S ratio was calculated on the 19,694-tip tree (Figure S9B) as described in the "calculation of nonsynonymous to synonymous divergence ratio" section, and this analysis is in "fig2supp-divergence-20k.ipynb." Convergently-evolved mutations (Figure S9C) were identified on this larger tree in "fig4supp-convergent_evolution-20k."

The impact of specific geographic regions on the results presented in Figure 1B was analyzed by computing the correlation coefficient *r* between S1 substitution accumulation and logistic growth rate for each geographic region separately. This was done by constructing a separation 10,000-sample tree for each of the six continental regions: Africa, Asia, Europe, North America, Oceania, and South America. Samples are from May 15, 2021 or earlier and are roughly evenly distributed over time. Each regional tree was built according to "phylogenetic reconstruction of a subsampling of global SARS-CoV-2 genome sequences," except that all sequences were restricted to that geographic region. The Asia-specific tree can be interactively-viewed at https://nextstrain.org/groups/blab/ncov/adaptive-evolution/2021-05-15/asia, and other regional trees can be accessed by substituting the region's name at the end of the URL. Code to conduct the analysis presented in Figure S10 is in "fig1-followup-regional.ipynb."

Analysis of duration of correlation between clade success and S1 substitutions

The correlation coefficient *r* between S1 substitution accumulation and logistic growth rate was computed over time using 13 phylogenies spanning a year of time surrounding the primary analysis. Trees for this analysis were built according to the methods in "phylogenetic reconstruction of a subsampling of global SARS-CoV-2 genome sequences" except that end date was changed to the 15th of the month, for each month between November 15, 2020 and November 15, 2021. The November 15, 2020 tree can be viewed at https://nextstrain.org/groups/blab/ncov/adaptive-evolution/2020-11-15, and all other dates can be accessed by changing the date at the end of the URL. Code to conduct the analysis presented in Figure S11 is in "fig1-followup-timeseries.ipynb."

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical analysis of correlation between growth rate and number of mutations

The p value for the correlation between growth rate and number of mutations was determined by comparing the empirical correlation coefficient *r* to the distribution of expected *r* values. If the observed correlation is just due to chance, the empirical *r* should not differ from the expectation. The null expectation is created by stripping all observed mutations from the phylogeny and then randomly assigning them to branches of the phylogeny as explained in "calculation of correlation between mutations and clade growth rate." Using the randomized phylogeny, the logistic growth rate is compared to the number of mutations on every internal branch using linear regression. The p value of the empirical correlation is then computed as the number of expected *r* values that are greater than the observed *r*, divided by 1,000. This method of statistical analysis is shown visually in Figure 1C and is used to provide the p values listed in Table 1 and those shown in Figures S3B, S10, and S12.

Statistical analysis of wait times

For S1 nonsynonymous, S1 synonymous and RdRp nonsynonymous mutations, the statistical significance of the mean wait time and proportion of wait times under 0.3 years was evaluated by comparing the empirical values to those generated from 1,000 phylogenies where the location of mutations on the phylogeny were randomized as described above in "randomization of recurrent mutations across the phylogeny."

Statistical analysis of growth rates associated with convergent mutations

The mean logistic growth rate of clades possessing convergent mutations that arose four or more times was compared to the expected value, given that the mutation occurred that many times as described in "randomization of recurrent mutations across the phylogeny." Mean growth rates that fall in the 90th percentile or higher of the expected distribution are labeled in Figure 4B.

ADDITIONAL RESOURCES

Interactive versions of all phylogenies used in this study can be found at https://nextstrain.org/groups/blab/ncov/adaptive-evolution/2021-05-15. The drop-down menus on the side can be used to select a different build date, appending "/20k" to the URL will bring up the phylogeny containing ~20,000 samples, and appending "/africa" (or any of the other geographic regions considered in Figure S10) will show the region-specific phylogeny.