Contributions of adaptation and purifying selection to SARS-CoV-2 evolution

Richard A. Neher

Biozentrum, University of Basel, Spitalstrasse 41, Basel 4053, Switzerland and Swiss Institute of Bioinformatics, Spitalstrasse 41, Basel 4053, Switzerland

Abstract

Continued evolution and adaptation of SARS-CoV-2 has led to more transmissible and immune-evasive variants with profound impacts on the course of the pandemic. Here I analyze the evolution of the virus over 2.5 years since its emergence and estimate the rates of evolution for synonymous and non-synonymous changes separately for evolution within clades—well-defined monophyletic groups with gradual evolution—and for the pandemic overall. The rate of synonymous mutation is found to be around 6 changes per year. Synonymous rates within variants vary little from variant to variant and are compatible with the overall rate of 7 changes per year (or 7.5×10⁻⁴ per year and codon). In contrast, the rate at which variants accumulate amino acid changes (non-synonymous mutations) was initially around 12-16 changes per year, but in 2021 and 2022 it dropped to 6-9 changes per year. The overall rate of non-synonymous evolution, that is across variants, is estimated to be about 26 amino acid changes per year (or 2.7×10⁻³ per year and codon). This strong acceleration of the overall rate compared to within clade evolution indicates that the evolutionary process that gave rise to the different variants is qualitatively different from that in typical transmission chains and likely dominated by adaptive evolution. I further quantify the spectrum of mutations and purifying selection in different SARS-CoV-2 proteins and show that the massive global sampling of SARS-CoV-2 is sufficient to estimate site-specific fitness costs across the entire genome. Many accessory proteins evolve under limited evolutionary constraints with little short-term purifying selection. About half of the mutations in other proteins are strongly deleterious.

Key words: fitness; landscape; evolutionary rates.

Since its emergence in late 2019 (Zhu et al. 2020), SARS-CoV-2 has displayed a discontinuous pattern of evolution with large jumps in sequence space giving rise to phylogenetically distinct variants (Hodcroft et al. 2021; Volz et al. 2021; Tegally et al. 2021; Faria et al. 2021; Naveca et al. 2021; Viana et al. 2022). Many of these variants spread considerably faster and quickly displaced the resident variants at the time either because of intrinsically increased transmissibility, evasion of prior immunity in the population, or a combination of both. Specific Variants of Concern or Interest were designated by the World Health Organization and labeled by Greek letters (Konings et al. 2021). The branches leading to these variants are characterized by many amino-acid-changing mutations that often cluster in the S1 domain of the spike protein (Kistler et al. 2022).

This pattern of rapid non-synonymous evolution in viral surface proteins that interact with the host cells is common among many RNA viruses and, for example, well studied in influenza A virus evolution (Bhatt, Holmes, and Pybus 2011; Strelkowa and Lässig 2012). But the adaptive evolution of influenza viruses tends to be gradual without large jumps in sequence space, while new variants of SARS-CoV-2 with tens of novel mutations emerged suddenly without intermediate genomes being observed, the most dramatic being the emergence of Omicron in late 2021 (Viana et al. 2022). One explanation for the sudden appearance of such highly mutated, transmissible, and immune evasive variants is rapid hidden evolution in chronic infections. Such prolonged infections are common in patients with impaired immune systems, either through HIV-1 infection (Cele et al. 2022) or medical intervention (Choi et al. 2020; Kemp et al. 2021). During such chronic infections, extensive intra-host diversity can develop through accelerated evolution (Chaguzza et al. 2022). Onward transmission from such chronic infections has also been documented (Gonzalez-Reiche et al. 2022). However, to date there is no direct evidence for the mode of emergence of any variant. The case for chronic infection being an important contributor is strongest for the variants Alpha and Omicron (Hill et al. 2022).

The dichotomous pattern of SARS-CoV-2 evolution with stepwise evolution within variants and atypical bursts of evolution leading to new variants has been investigated by Tay et al. (2022) and Hill et al. (2022), who showed that the rate of evolution along branches giving rise to new variants is up to fourfold higher than the background rate. Here, I build on these results and investigate the patterns of SARS-CoV-2 diversification within variants and compare these to the global dynamics of evolution and adaptation. This comparison reveals a consistent dichotomy between slow within-variant evolution and rapid adaptive evolution, giving
rise to new variants. This difference in evolutionary rates is only seen for non-synonymous changes—the rate of synonymous evolution within variants is compatible with that seen between variants. Furthermore, early variants display more rapid non-synonymous evolution than later variants, suggesting more ubiquitous adaptive evolution early on. I further quantify the level of functional constraint of different open reading frames and infer a map of mutational tolerance across the genome from patterns of rare diversity.

Results
Evolutionary rates and divergence times are typically estimated using phylogenetic approaches (Drummond et al. 2006). These methods, however, cannot handle the volume of SARS-CoV-2 data available and data have to be dramatically down-sampled. Furthermore, phylogenetic methods impose an hierarchical structure on the data and are thus very sensitive to problematic sequences or metadata: Any misdated, recombinant, contaminated, or otherwise chimeric sequence can fatally distort the analysis. Problematic sequences can be particularly common when a new variant takes over since sequencing protocols need adjusting (De Maio et al. 2020).

To circumvent many of the above-mentioned problems and still use the majority of the available data, I use a combination of automated filtering and simple robust approaches to analyze the evolutionary patterns (see Materials and Methods). I first use Nextclade (Aksamentov et al. 2021) to assign 12 million sequences available in GISAID (2022-07-25) (Shu and McCauley 2017) to one of the Nextstrain clades (Hadfield et al. 2018; Roemer et al. 2022), which are denoted by a year–letter combination (e.g. 19A, 20A, 20B,…), see Fig. 1. Sequences belonging to recombinant Pango lineages are excluded (Rambaut et al. 2020). These clades represent well-defined groups of sequences with little evidence of recombination within them and are analyzed independently. In this analysis, I only consider clades that have had significant circulation for at least 6 months.

For each clade, I define a ‘founder’ genotype and exclude any sequence that does not have the full set of clade-defining mutations relative to the reference sequence Wuhan/Hu-1. This founder sequence is manually curated for each clade considered. This filtering removes most incomplete sequences as well as sequences where amplicon dropouts are back-filled with the reference sequence, but will ignore a few sequences with true reversion mutations. In addition, sequences with a Nextclade quality control (QC) score above 30 (80 for 21H because of an unaccounted frameshift) are removed. The results are insensitive to the stringency of this filtering. For this reduced set of sequences, I determine the mutations they carry on top of the founder genotype of the clade to analyze diversification and divergence within the clade. The latter step is done for nucleotide changes as well as for amino acid changes.

Within each clade, the number of mutations is expected to increase linearly in time and the variation around this mean would, in an ideal case, obey Poisson statistics. For the majority of sequences, this is approximately true, but some problematic sequences have more mutations than expected. To exclude these outliers, I perform a simple linear regression of the number of ‘intra-clade’ mutations against time and remove sequences whose deviation from the linear fit exceeds twice the expected standard deviation by three mutations (see Fig. 2). These outliers are never more than 1 per cent of all sequences that pass the first round of filtering, and the results are insensitive to the filtering criteria.
Figure 2. Within-clade divergence increases linearly with time (example clade 20C). Top: Each panel shows the number of within-clade mutations (total (A), amino acid changing (B), and synonymous (C)) as a function of time. The additional green line in Panel A indicates the divergence cut-off; Panels B & C only show sequences that pass the divergence filter. Each panel also shows mean ± standard deviation and a weighted linear fit. Bottom: Panels D and E show the prevalence of the most common genotypes (D) and mutations (E) during the first 3 months. Genotypes in (D) refer to sequences with all and only the mutations indicated, while mutations in (E) count all sequences with a specific mutation, regardless of mutations elsewhere. In the case shown (20C), the founder genotype initially dominates and no specific mutation ever reaches high frequency. Note that specific genotypes invariably decrease in frequency due to subsequent mutations, while mutations irrespective of their genetic background persist. Panel F shows a Poisson model fit to the breakdown of the population into genotypes with different numbers of mutations over time. Analogous figures for other clades are included in the Appendix.

After removing these outliers, the data are binned by calendar week. Evolutionary rate and putative emergence date of the variant are then estimated by weighted linear regression where each time bin is weighted with the fourth root of the number of sequences in the bins. The exact functional form of this weighting does not have a big influence on the results, but a sublinear weighting helps to counter the large variation in sequencing effort across countries and the unavoidable imbalance due to the fact that few sequences are available early on when an emerging variant is still rare. Figure 2 shows the increasing intra-clade divergence for clade 20C, a large clade that emerged in early 2020 that was common in North America and Scandinavia. Both synonymous and non-synonymous within-clade average divergence increase linearly over time, allowing for a robust estimate of the rate.

Due to shared ancestry, divergences of sequences are not independent data points and a regression against time is generally not a suitable method to estimate evolutionary rates. In particular, confidence intervals are difficult to obtain. However, in the case of rapidly expanding variants we typically observe a large number of independent lineages emanating from one or several basal polytomies. Along each of these lineages, mutation accumulation is independent. Although not every sequence is an independent sample, the effective number of independent samples is large and the steadily increasing average divergence allows to estimate the rate robustly.

A simple model for diversity within a growing variant is a supercritical branching process with growth rate $\alpha$ and an embedded mutation process. Offspring of genomes with $i$ mutations will carry $i + j$ mutations, where $j$ is a Poisson distributed number with mean $\mu t$ (mutation rate $\mu$ and generation time $t$). The probability that offspring genomes are different from their parents is $u = 1 - e^{-\mu t}$, which for a generation time of $t=5$ days and a rate of $\mu = 15/\text{year}$ evaluates to $u \approx 0.2$ (note that this rate excludes strongly deleterious mutations, see Discussion).

When considering a rapidly growing well-sampled outbreak, typically a single founder genotype will give rise to a large number of daughter lineages that evolve independently. In this case, the diversification processes are robustly described by their means. Since the above branching process is linear, the mean number of cases $n$ will increase exponentially with rate $\alpha$, while the number
of genomes with \(i\) mutations relative to the founder \(m_i\) grows with rate \(\alpha - u\) per generation:

\[
\frac{dn}{dt} = \alpha n \\
\frac{dm_0}{dt} = (\alpha - u)m_0 \\
\frac{dm_1}{dt} = um_0 + (\alpha - u)m_1 \\
\frac{dm_2}{dt} = um_1 + (\alpha - u)m_2 \\
\ldots
\]

with solution \(m_i = e^{(\alpha-u)t} \frac{\alpha^i}{i!}\). Note that this model assumes continuous time and only allows increments by one mutation at a time. Multiple mutations in one serial interval can still happen through successive mutations within one host. At time \(t\) after the emergence of the variant, the number of mutations in the population is expected to be Poisson distributed with mean \(ut\). The overall number of cases is \(e^{\alpha t}\) or more generally \(e^{\int \alpha(t)dt}\) if growth rate or ascertainment varies over time.

For some clades, especially those that are well sampled soon after their emergence, this Poisson model is a good fit to diversity accumulation and yields estimates of rates and time of origin that are compatible with the divergence regression, see Fig. 2F for clade 20C. In this case, the founder genotype initially dominates, but is gradually replaced—first by single mutant genotypes, then double mutants, and so forth (comp. Fig. 2). Analogous graphs for all other clades considered are included in the Appendix.

In other cases, notable subclades did become dominant during the early stochastic dynamic. In this case, extrapolation of the linear fit to zero divergence is not necessarily a good estimator of the emergence time of a variant: if branches leading to big subclades carry anomalously many or few mutations, the divergence time will be over- or underestimated. For several clades 19B, 20H (Beta), 21D (Eta), 21G (Lambda), 21H (Mu), 21I (Delta), 21J (Delta), 21L (Omicron, BA.2), and 22B (Omicron, BA.5), founder-like variants are a minority even in early data.

These Poisson weights of mutation numbers are again only valid if mutations accumulate along many independent lineages. In particular, this assumption is violated if some lineages spread systematically faster than others either because of epidemiological factors or because they carry adaptive mutations. In variant 21K (Omicron, BA.1), a sublineage with mutation S:R346K might have enjoyed a transmission advantage. A possible consequence of this advantage is visible in Supplementary Fig. S21, where genotypes with 0, 1, or 2 mutations are decaying more rapidly than those with more mutations.

Despite these caveats, for almost all Nextstrain clades, the slope at which diversity increases is robust (similar patterns as Fig. 2A-C for clade 20C), allowing us to estimate clade-specific evolutionary rates for amino acid and synonymous changes. These rates are summarized in Fig. 3 and Table 1. In addition to these within-clade rates, I estimated the rate at which the clades themselves accumulated amino acid and synonymous changes by regressing the number of differences of the clade’s founder sequence (relative to the putative root in clade 19B (Caraballo-Ortiz et al. 2022)) against the estimated time of origin of the clade. These regressions are shown in thick gray lines in Fig. 3.

Rates of synonymous change are very consistent across clades (about 5–8 changes per genome per year) and also agree with the overall rate of synonymous changes of 7.3 changes per genome per year. The rates of non-synonymous changes are much more variable (Fig. 3D). Within clades, the rate of non-synonymous changes varies between 5 and 16 changes per year. Earlier clades are estimated to have larger rates around 10–15 changes per year, while rate estimates for later clade fall between 3 and 9 changes per year (see Fig. 4). In contrast, the inter-clade non-synonymous rate exceeds 25 changes per year. The spike protein does not contribute to this decreasing trend in the rate of within-clade non-synonymous evolution, which is most evident in ORF1ab and to a lesser degree in accessory proteins and \(N\) (see Supplementary Figure S1).

Nextstrain clades tend to be defined by long branches leading to a large polytomy. It could thus be that the estimated inter-clade rate exceeds the intra-clade rate purely because of this conditioning. This effect might be particularly important early on in the pandemic when diversity was low and when branches with as few as two mutations were used to define new clades. I therefore also included composite clades 19B+, 19B++, and 20A+ containing sequences from 19A and 19B (rooted on 19B), 19A, 19B, 20A, 20B, 20C, and 20D (rooted on 19B), and 20A, 20B, 20C, and 20D (rooted on 20A). The estimates for composite clades 19B+ and 20A+ are consistent with the estimates of the individual clades, while apparent rates of 19B++ are considerably higher. The latter is due to the rapid expansion and subsequent dominance of clade 20A, the first clade with spike mutation Del/4G, and its descendants which rapidly fixed four additional mutations (Korber et al. 2020). This is an early example of an accelerated global rate of evolution due to adaptive evolution, see Discussion.

At the other extreme, some Omicron clades have very low rate estimates which should be interpreted carefully. Founder genotypes of Omicron clades 21L (BA.2), 22A (BA.4), and 22B (BA.5) are rarely sampled. These clades showed considerable diversity, including reversions to the reference, soon after their initial discovery (Tegally et al. 2022). This diversity and the complex, possibly recombinant, origin of the Omicron clades make estimating their rates of evolution challenging.

### Purifying selection and mutation tolerance

As described earlier, the rate of synonymous mutations is comparable within and between clades without a strong indication that this rate might have changed over time. This is expected, as synonymous positions are rarely a locus of adaptation and tend to have a small effects on fitness in large parts of the genomes of RNA viruses (Zanini et al. 2017) (outside of specific regions with important RNA elements or splice sites). To quantify how much of the SARS-CoV-2 genome is constrained and how strongly purifying selection operates on different genomic regions, I made use of the ‘rare mutations’ annotation provided by Nextclade. Nextclade attaches each sequence to a reference tree and determines by which mutations it differs from the attachment point. For each Pango lineage (as determined by Nextclade) (Rambaut et al. 2020; Aksamentov et al. 2021), I count how often these ‘rare mutations’ (including reversions to the reference) are observed. This way, for each position in the genome, one obtains the fraction of lineages with minor variation (excluding singletons). I normalize this fraction against the relative rate of mutation away from the ancestral nucleotide (see Supplementary Fig. S2) and use this as a semi-quantitative proxy of mutational tolerance.

Simply splitting the genome into at first, second, and third positions of codons already reveals strong signatures of purifying selection, see Fig. 5. Between 15 and 20 per cent of first and second positions in codons show almost no variation, while half of these
Figure 3. Divergence and evolutionary rates of different Nextstrain clades. Panels A, B, and C show the estimated divergence of the founder genotype of each clade (big dot or square) and the subsequent divergence trend for all nucleotide changes, amino acid changes, and synonymous changes, respectively. In addition, each panel contains a regression of the divergence of clade founders vs. time (thick gray line). The standard deviation expected based on Poisson statistics is indicated as shaded area. Panel D summarizes the individual rate estimates (dots and squares) and compares them to the estimate of inter-clade rates (thick gray lines). In panel D, clades are in alphabetical order, which is similar to their (uncertain) order of emergence. The red outlier is ‘composite’ clade 19B++ (containing 19B, 19A, 20A, 20B, and 20C) with inflated rates due to adaptive mutations on the branch leading to clade 20A.

Discussion

The inferred evolutionary rate of RNA viruses often decreases with the timescale across which it is estimated (Wertheim and Kosakovsky Pond 2011; Ghafari et al. 2021). This effect can be particularly pronounced at the beginning of an outbreak following a host switch and has been attributed to incomplete purifying selection or methodological issues leading to inflated measures of diversity (Meyer et al. 2015; Ghafari et al. 2022). In addition to segregating deleterious mutations, early viral evolution after a host switch can also be driven by anomalously fast adaptation. A dramatic change in environment, e.g. a host switch, likely results in many opportunities for mutations that increase fitness. Such transient increases in the rate of adaptation are common in experimental evolution (Elena and Sanjuán 2007).
Table 1. Evolutionary rates estimates from root-to-tip regressions for the overall nucleotide changes, amino acid (aa) changes, and synonymous changes. The column ‘# of seq.’ shows the number of sequences that entered the analysis after filtering and restriction to the first 6 months after the emergence of the variant. The last three columns give the distances of the clade founder sequence from putative MRCA of SARS-CoV-2 (19B).

| Clade   | # of seq. | Overall rate [y⁻¹] | aa rate [y⁻¹] | Syn rate [y⁻¹] | Overall div. | aa div. | Syn div. |
|---------|-----------|---------------------|---------------|---------------|--------------|---------|---------|
| 19B     | 8,187     | 19.46               | 12.37         | 7.09          | 0            | 0       | 0       |
| 19B+    | 22,745    | 18.56               | 11.92         | 6.64          | 0            | 0       | 0       |
| 19B++   | 1,26,823  | 17.91               | 16.41         | 11.50         | 0            | 0       | 0       |
| 19A     | 14,382    | 19.73               | 12.79         | 6.95          | 2            | 1       | 1       |
| 20A     | 44,629    | 15.08               | 7.94          | 7.15          | 6            | 3       | 3       |
| 20B     | 42,077    | 14.29               | 7.69          | 6.59          | 9            | 5       | 4       |
| 20C     | 47,865    | 16.73               | 9.92          | 6.81          | 8            | 5       | 3       |
| 20A+    | 1,36,863  | 16.30               | 9.40          | 6.90          | 6            | 3       | 3       |
| 20E     | 82,240    | 14.04               | 5.65          | 8.39          | 13           | 5       | 8       |
| 20H (Beta) | 5,077   | 15.48               | 9.51          | 5.96          | 21           | 16      | 5       |
| 20I (Alpha) | 2,77,226 | 11.56              | 5.49          | 6.07          | 30           | 19      | 11      |
| 20J (Gamma) | 24,970   | 12.97              | 6.29          | 6.68          | 35           | 22      | 13      |
| 21D (Eta) | 3,653    | 11.36               | 6.29          | 5.07          | 27           | 13      | 14      |
| 21G (Lambda) | 4,567    | 12.04              | 6.51          | 5.53          | 28           | 19      | 9       |
| 21H (Mu) | 2,897     | 18.25               | 10.05         | 8.20          | 31           | 21      | 10      |
| 21I (Delta) | 86,585    | 16.55              | 8.93          | 7.62          | 30           | 24      | 6       |
| 21J (Delta) | 7,33,825 | 14.30              | 7.17          | 7.14          | 34           | 28      | 6       |
| 21K (Omicron) | 9,56,966 | 13.48              | 6.45          | 7.03          | 56           | 46      | 10      |
| 21L (Omicron) | 6,90,177 | 9.24               | 4.39          | 4.85          | 68           | 50      | 18      |
| 22A (Omicron) | 33,451   | 7.71               | 2.81          | 4.90          | 71           | 52      | 19      |
| 22B (Omicron) | 1,24,239 | 11.90              | 5.03          | 6.87          | 66           | 50      | 16      |

Figure 4. Divergence and evolutionary rates of different Nextstrain clades over time. Synonymous rates estimates are stable in time and fluctuate around the rate estimates between clades. Non-synonymous rate estimates are highest for clades 19A to 20C that emerged early in the pandemic.

Here, I showed that the deceleration during the first 6 months of the pandemic is only observed for non-synonymous mutations. Since I analyzed evolution within short-lived clades of SARS-CoV-2 over a span of up to 6 months, purifying selection has comparable efficiency and should affect the estimates equally for all clades. Similarly, sequencing artifacts are expected to increase divergence at all time points and affect synonymous and non-synonymous rates in similar ways. The same applies to potential changes in the fidelity of the polymerase. Nevertheless, the estimated non-synonymous evolutionary rate of clades circulating in late 2019 and early 2020 is about twice as high as that of clades in 2021 and 2022, while the synonymous rate does not change (see Fig. 3 and Table 1). One possible explanation is that the early evolutionary rate of SARS-CoV-2 was inflated by adaptive evolution.

Most of this apparent initial acceleration of within-variant evolution ceased in mid-2020 and even early variants like 20E accumulated non-synonymous changes at a rate of about 6 instead of 12 changes per year. By that time, the number of non-synonymous differences relative to the root of the tree was small (5 in the case of 20E) and it is implausible that this small change would have exhausted the pool of beneficial mutations. So why would the rate of adaptive evolution slow down? Maybe it is just chance, after all it is only clades 19A and 19B where the rate is substantially higher. Another possible explanation could be diminishing returns epistasis. Viruses with the S.D614G mutation showed faster replication
Figure 6. Landscape of selective constraint along the SARS-CoV-2 genome. Solid lines show sliding window smoothing of the estimated mutational tolerance at first, second, and third positions for a window size of 20 and 7 (faint lines) sites. The markers at the bottom of the panels show qualitatively how many Pango lineages are different at this position from the Wuhan-Hu-1 reference sequence.
Virus Evolution

The virus might, therefore, be operating closer to the maximal capacity at which cells can produce virions, reducing the scope for further optimization. Other mutations then have smaller benefits, rise in frequency more slowly, and the effect of adaptation does not manifest itself over the lifetime of the clades studied here. Such diminishing returns epistasis has been observed in the experimental evolution with yeast (Kryazhimskiy et al. 2014). Global epistasis of this nature can exist alongside epistasis of specific amino acid changes as, for example, suggested for the S:S:N501Y mutation (Martin et al. 2021; Rochman et al. 2022; Martin et al. 2022). Differences in diversification of SARS-CoV-2 with and without the S:D614G mutation were also observed in evolution experiments by Amicone et al. (2022), who concluded that these differences were not due to a change in the baseline mutation rate but had a selective origin.

An evolutionary rate of 6 synonymous changes per year at around 9,700 positions corresponds to a per-site evolutionary rate of $6.2 \times 10^{-6}$ per site per year or $1.7 \times 10^{-9}$ per day, slightly higher than the estimated baseline mutation rate of $1.3 \times 10^{-9}$ per day (Amicone et al. 2022). The total evolutionary rate within variants after mid-2020 (starting with 20E) is in the range of 9–16 changes per year, corresponding to a per-site rate of $3–5 \times 10^{-6}$ per site per year, consistent with recent estimates by Hill et al. (2022) and Tay et al. (2022).

When considering only the putative founder genotype and date of origin of each variant, all variants so far are compatible with a backbone evolutionary rate of 32 changes per year, corresponding to an per site rate of around $10^{-5}$ per site per year (see Fig. 3). This rate is a composite of the episode of cryptic accelerated evolution (possibly in chronically infected individuals) and regular transmission chains of acute infections. This estimate for the average rate is thus not inconsistent with results by Hill et al. (2022) and Tay et al. (2022), who estimated even higher rates specifically for the branches that gave rise to variants of concern. Different clades and variants probably emerged in different ways under different circumstances. Nevertheless, all clades and variants are compatible with a single ‘backbone’ molecular clock that runs substantially faster than the ‘within variant clock’. This accelerated backbone clock is likely driven by exponential amplification of beneficial mutations.

In addition to neutral and potentially adaptive mutations, I also quantified purifying selection on the SARS-CoV-2 genome. By analyzing the rate of mutations that spread only on short timescales within fine-grained Pango lineages, I estimated the level of constraint on different parts of the SARS-CoV-genome. The great majority of third positions in codons—at which most mutations are synonymous—do not show strong signatures of conservation. The major open reading frames are synonymous—do not show strong signatures of conservation. The majority of third positions in codons—at which most mutations result in increased diversity on short timescales compared to longer ones, which can lead to time-dependent effective evolutionary rates (Wertheim and Kosakovsky Pond 2011).

The effects of adaptive evolution are harder to control and account for. Since the number of sites that allow beneficial mutations is small, adaptive evolution tends to be very stochastic—it is not the typical events, but the rare and extreme events that determine the course of adaptive evolution. Unlike neutral evolution, the rate of adaptive evolution depends on the population size, stochastic nature of the transmission process, the environment, and previous adaptation (Neher 2013). These factors make extrapolation—whether it is reconstruction of past events or scenarios of future evolution—uncertain. Standard phylogenetic or phylodynamic methods will underestimate this uncertainty.

If the excess of non-synonymous changes on the pandemic scale compared to within-clade evolution is due to adaptive evolution, it would imply an adaptive rate of about 20 changes per year (plus insertion/deletions not considered here). This process is likely to continue even in the more and more diverse immunological landscape of the human population in which the viral population spreads and adapts. But it is unclear whether SARS-CoV-2 continues to evolve in a saltatory fashion with the repeated emergence of highly mutated variants or whether we will see a transition to a more gradual adaptive process. We currently observe multiple lineages with several amino acid changes and moderate transmission advantages emerging in a more stepwise fashion within 21L, 22A, and 22B (BA.2, BA.4, and BA.5), which could indicate a shift to a more influenza-like evolution.

Materials and Methods

Data

This analysis is based on SARS-CoV-2 sequencing data shared by researchers around the world via GISAID (Shu and McCauley, 2017) (EPISET_ID EPI_SET_220812zf). The bulk of the analysis can be reproduced using SARS-CoV-2 genomes shared via International Nucleotide Sequence Database Collaboration (INSDC). These data are available from Nextstrain at http://data.nextstrain.org/files/ncov/open/metadata.tsv.gz and http://data.nextstrain.org/files/ncov/open/nextclade.tsv.gz. However, for some clades and variants, coverage in INSDC is low.

Analysis

The code used to run this analysis is publicly available at https://github.com/neherlab/SC2_variant_rates. It is organized as
a Snakemake workflow (Köster and Rahmann 2012). I used Nextclade dataset 2022-07-26T12:00:00Z with the reference sequence MN908947 (Wuhan-Hu-1). The codon analysis ignores ORFs, which is fully contained in N but in a different reading frame. Tabular files with estimates of evolutionary rates, the mutation distribution, and fitness costs are included in the git repository.

The fitness costs of each positions are estimated from the Nextclade columns unlabeledSubstitutions and reversionSubstitutions for each sequence that has the QC status good. For each Pango lineage, I count how often each specific mutation was observed. I then calculate how many lineages show minor variations for each mutation. To avoid the loss of signal through sequencing errors, I drop lineages at which only a single mutation at this position is observed. To turn this count into a measure of mutational tolerance, I normalize the number of lineages with minor variation by the mutation rate away from the ancestral base at this position within the lineage.

Acknowledgements

I gratefully acknowledge the work by thousands of scientists and health-care professionals that collect the samples, generate, and share the sequence data on which this analysis is built. I am also grateful to Jesse Bloom, Louis du Plessis, and Florence Débarre for valuable feed-back on the manuscript and to Cornelius Roemer and Emma Hodcroft for stimulating discussions on many aspects of SARS-CoV-2 evolution.

Supplementary data

Supplementary data are available at Virus Evolution online.

References

Aksamentov, I. et al. (2021) 'Nextclade: clade assignment, mutation calling and quality control for viral genomes', Journal of Open Source Software, 6: 3773.

Amicone, M. et al. (2022) 'Mutation rate of SARS-CoV-2 and emergence of mutators during experimental evolution', Evolution, Medicine, and Public Health, 10: 142–55.

Bhatt, S. F. C., Holmes and O. G., Pybus (2011) 'The genomic rate of molecular adaptation of the human influenza A virus', Molecular Biology and Evolution, 28: 2443–51.

Caraballo-Ortiz, M. A. et al. (2022) 'TopHap: rapid inference of key phylogenetic structures from common haplotypes in large genome collections with limited diversity', Bioinformatics, 38: 2719–26.

Cele, S. et al. (2022) 'SARS-CoV-2 prolonged infection leading to advanced HIV disease evolves extensive immune escape', Cell Host and Microbe, 30: 154–62.e5.

Chagusa, C. et al. (2022), Accelerated SARS-CoV-2 intrahost evolution leading to distinct lineages during chronic infection. <https://www.medrxiv.org/content/10.1101/2022.06.29.22276868v1>.

Choi, B. et al. (2020) 'Persistence and evolution of SARS-CoV-2 in an immunocompromised host', New England Journal of Medicine, 383: 2291–3.

De Maio, N. et al. (2020), Issues with SARS-CoV-2 sequencing data-SARS-CoV-2 coronavirus/nCoV-2019 Genomic Epidemiology. <https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473> accessed 15 Oct 2022.

Drummond, A. J. et al. (2006) 'Relaxed phylogenetics and dating with confidence', PLoS Biology, 4: e88.

Elena, S. F. and R., Sanjuán (2007) 'Virus evolution: insights from an experimental approach', Annual Review of Ecology, Evolution, and Systematics, 38: 27–52.

Faria, N. R. et al. (2021), 'Genomics and epidemiology of the P1 SARS-CoV-2 lineage in Manaus, Brazil', Science 372: 815–21.

Ghafarí, M. et al. (2021) bioRxiv, 2021.02.09.430479 (Cold Spring Harbor Laboratory Section: New Results), <https://www.biorxiv.org/content/10.1101/2021.02.09.430479v1>.

Ghafarí, M. et al. (2022) 'Purifying selection determines the short-term time dependency of evolutionary rates in SARS-CoV-2 and pH1N1 influenza', Molecular Biology and Evolution, 39: msac009.

Gonzalez-Reiche, A. S. et al. (2022) medRxiv, 2022.05.25.22275533 ISSN-2227-5533 Type: article, <https://www.medrxiv.org/content/10.1101/2022.05.25.22275533v1>.

Hadfield, J. et al. (2018) 'Nextstrain: real-time tracking of pathogen evolution', Bioinformatics, 34: 4121–3.

Hill, V. et al. (2022) 'The origins and molecular evolution of SARS-CoV-2 lineage B.1.1.7 in the UK’, Virus Evolution, 8: veac080.

Hodcroft, E. B. et al. (2021) 'Spread of a SARS-CoV-2 variant through Europe in the summer of 2020', Nature, 595: 707–12.

Kemp, S. A. et al. (2021) 'SARS-CoV-2 evolution during treatment of chronic infection', Nature, 592: 277–82.

Kistler, K. E., J., Huddleston and T., Bedford (2022) 'Rapid and parallel adaptive mutations in spike S1 drive clade success in SARS-CoV-2', Cell Host and Microbe, 30: 545.

Konings, F. et al. (2021) 'SARS-CoV-2 Variants of Interest and Concern naming scheme conducive for global discourse', Nature Microbiology, 6, 821–3.

Korber, B. et al. (2020) 'Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus’, Cell, 182: 812.

Köster, J. and S., Rahmann (2012) ‘Snakemake—a scalable bioinformatics workflow engine’, Bioinformatics, 28: 2520.

Kryazhimskiy, S. et al. (2014) 'Global epistasis makes adaptation predictable despite sequence-level stochasticity', Science, 344: 1519.

Martin, D. P. et al. (2021) 'The emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages’, Cell, 184: 5189–200.e7.

Martin, D. P. et al. (2022) ‘Selection analysis identifies clusters of unusual mutational changes in Omicron lineage BA.1 that likely impact Spike function’, Molecular Biology and Evolution, 39: msar061.

Meyer, A. G. et al. (2015) 'Time dependence of evolutionary metrics during the 2009 pandemic influenza virus outbreak’, Virus Evolution, 1: vev006.

Naveca, F. G. et al. (2021) 'COVID-19 in Amazonas, Brazil, was driven by the persistence of endemic lineages and P.1 emergence’, Nature Medicine, 27: 1230–8.

Neher, R. A. (2013) ‘Genetic draft, selective interference, and population genetics of rapid adaptation’, Annual Review of Ecology, Evolution, and Systematics, 44: 195–215.

Rambaut, A. et al. (2020) 'A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology’, Nature Microbiology, 5: 1403–7.

Rochman, N. D. et al. (2021) ‘Ongoing global and regional adaptive evolution of SARS-CoV-2’, Proceedings of the National Academy of Sciences, 118: e2104241118.

Rochman, N. D. et al. (2022) 'Epistasis at the SARS-CoV-2 receptor-binding domain interface and the propitiously boring implications for vaccine escape’, mBio, 13: e00135.

Roemer, C. et al. (2022), SARS-CoV-2 clade naming strategy for 2022, <https://nextstrain.org/blog/2022-04-29-SARS-CoV-2-clade-naming-2022> accessed 15 Oct 2022.
Shu, Y. and McCauley (2017) ‘GISAID: Global initiative on sharing all influenza data – from vision to reality’, Eurosurveillance, 22: 30494.

Strelkowa, N. and M., Lässig (2012) ‘Clonal interference in the evolution of influenza’, Genetics, 192: 671–82.

Tay, J. H. et al. (2022) ‘The emergence of SARS-CoV-2 Variants of Concern is driven by acceleration of the substitution rate’, Molecular Biology and Evolution, 39: msac013.

Tegally, H. et al. (2021) ‘Detection of a SARS-CoV-2 variant of concern in South Africa’, Nature, 592: 438–43.

Viana, R. et al. (2022) ‘Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa’, Nature, 603: 679–82.

Volz, E. et al. (2021) ‘Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England’, Nature, 593: 266–9.

Wertheim, J. O. and S. L., Kosakovsky Pond (2011) ‘Purifying selection can obscure the ancient age of viral lineages’, Molecular Biology and Evolution, 28: 3355–65.

Zanini, F. et al. (2017) ‘In vivo mutation rates and the landscape of fitness costs of HIV-1’, Virus Evolution, 3: vex003.

Zhu, N. et al. (2020) ‘A novel coronavirus from patients with pneumonia in China, 2019’, New England Journal of Medicine, 382: 727–33.