Fitness effects of mutations to SARS-CoV-2 proteins

Jesse D. Bloom[1,2,3*] **and Richard A. Neher**[4,5*]
[1]Basic Sciences and Computational Biology, Fred Hutchinson Cancer Center
[2]Department of Genome Sciences, University of Washington
[3]Howard Hughes Medical Institute
[4]Biozentrum, University of Basel
[5]Swiss Institute of Bioinformatics

ABSTRACT Knowledge of the fitness effects of mutations to SARS-CoV-2 can inform assessment of new variants, design of therapeutics resistant to escape, and understanding of the functions of viral proteins. However, experimentally measuring effects of mutations is challenging: we lack tractable lab assays for many SARS-CoV-2 proteins, and comprehensive deep mutational scanning has been applied to only two SARS-CoV-2 proteins. Here we develop an approach that leverages millions of publicly available SARS-CoV-2 sequences to estimate effects of mutations. We first calculate how many independent occurrences of each mutation are expected to be observed along the SARS-CoV-2 phylogeny in the absence of selection. We then compare these expected observations to the actual observations to estimate the effect of each mutation. These estimates correlate well with deep mutational scanning measurements. For most genes, synonymous mutations are nearly neutral, stop-codon mutations are deleterious, and amino-acid mutations have a range of effects. However, some viral accessory proteins are under little to no selection. We provide interactive visualizations of effects of mutations to all SARS-CoV-2 proteins (https://jbloomlab.github.io/SARS2-mut-fitness/). The framework we describe is applicable to any virus for which the number of available sequences is sufficiently large that many independent occurrences of each neutral mutation are observed.

The rapid evolution of SARS-CoV-2 has led to the emergence of viral variants with enhanced transmissibility, escape from therapeutics, or reduced recognition by immunity [1, 2]. To anticipate and mitigate this evolution, the scientific community has launched efforts to assess the risk of new viral variants [3] and create therapeutics that target constrained regions of the virus where resistance is less likely to evolve [4, 5, 6]. Both efforts require determining how specific mutations affect viral fitness.

Unfortunately, experimentally measuring the effects of mutations is challenging for most SARS-CoV-2 proteins. For spike, tractable lab assays have identified key functional and antigenic mutations [1, 7], and enabled deep mutational scanning measurements of how most mutations affect receptor binding, cellular infection, and antibody recognition [8, 9, 10, 11]. These experimental data are valuable for assessing new spike variants [3, 12, 13] and designing antibody therapeutics with greater resistance to escape [14, 15, 16]. But most SARS-CoV-2 proteins lack tractable lab assays, despite contributing to viral fitness [17, 18, 19] and being targets of efforts to develop anti-viral drugs [20]. The only non-spike SARS-CoV-2 protein with large-scale experimental measurements of mutation effects is Mpro [21, 22].

An alternative to experiments is to estimate effects of mutations by analyzing natural viral sequences. The amount of data available for such analyses has increased dramatically over the last few years with the sequencing of SARS-CoV-2 from millions of human infections. So far analyses of these sequences have focused on analyzing expanding viral clades to identify mutations that mediate immune escape or increase transmissibility [23, 24, 25]. The basic idea is that mutations that repeatedly appear near the base of clades that increase in relative frequency are likely beneficial to the virus. However, only a small minority of all possible mutations are beneficial, with most being nearly neutral or deleterious. For purposes such as identifying constrained drug targets or understanding the function of viral proteins, it is important to estimate the effects of neutral or deleterious mutations as well as beneficial ones. Other studies have analyzed broader alignments of coronaviruses substantially diverged from SARS-CoV-2 [26, 27], but the resulting estimates are limited by sparse sampling and possible changes in the impacts of some mutations across divergent viruses.

Here we develop a new approach that uses natural sequences to estimate the effects of mutations. Our basic insight is that there are now so many SARS-CoV-2 sequences that all non-deleterious single-nucleotide mutations are expected to independently occur many times along the observed phylogenetic tree. We therefore first calculate the number of expected observations of independent occurrences of each mutation based on the neutral mutation rate of SARS-CoV-2. We then compare these expected observations to the actual observations in the SARS-CoV-2 tree to estimate the effect of each mutation. The resulting estimates correlate well with existing deep mutational scanning data. Most viral proteins have regions under strong selective constraints. However the accessory proteins mostly show only weak selection against amino-acid and even stop-codon mutations. Overall, our work demonstrates a new approach to determine the effects of mutations, and provides detailed maps of functional constraint across the SARS-CoV-2 proteome.

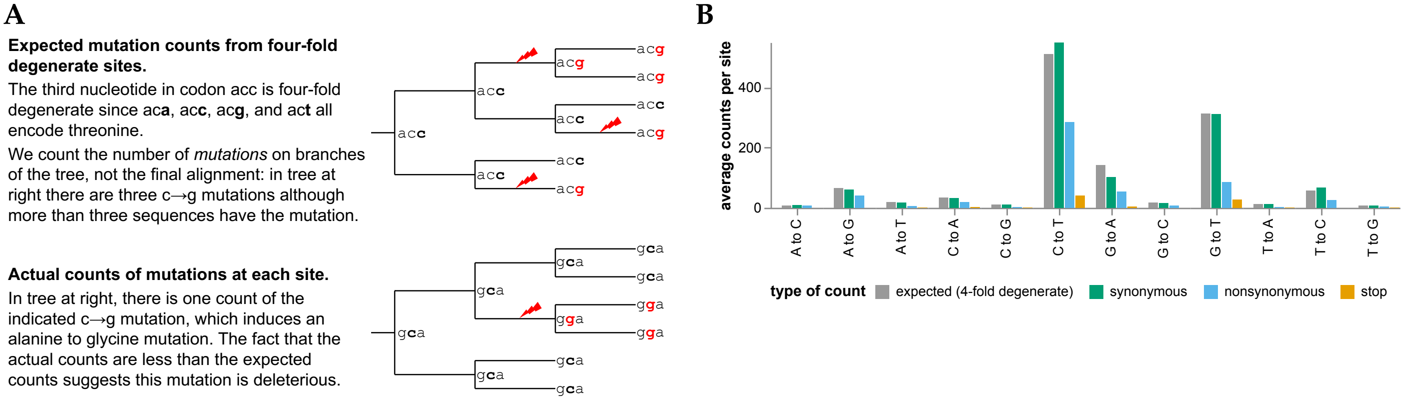
Results

Mutation effects from actual versus expected counts

To determine how many times each mutation is expected to be observed, we used the pre-built UShER tree [28, 29, 30] of ∼6.5-million public SARS-CoV-2 sequences to count nucleotide muta-

Manuscript compiled: Monday 30th January, 2023

*jbloom@fredhutch.org or richard.neher@unibas.ch



Expected mutation counts from four-fold degenerate sites.

The third nucleotide in codon acc is four-fold degenerate since ac**a**, ac**c**, ac**g**, and ac**t** all encode threonine.

We count the number of *mutations* on branches of the tree, not the final alignment: in tree at right there are three c→g mutations although more than three sequences have the mutation.

Actual counts of mutations at each site.

In tree at right, there is one count of the indicated c→g mutation, which induces an alanine to glycine mutation. The fact that the actual counts are less than the expected counts suggests this mutation is deleterious.

Figure 1 Expected versus actual counts of mutations. **(A)** The number of expected counts of each type of nucleotide mutation is computed from four-fold degenerate sites, and then compared the actual counts of each mutation. **(B)** Expected versus actual counts for each nucleotide mutation type aggregated across all viral clades and averaged across all sites where the mutation is four-fold degenerate, synonymous (including four-fold degenerate), nonsynonymous, or introduces a stop codon. See https://jbloomlab.github.io/SARS2-mut-fitness/avg_counts.html for an interactive version of panel B that enables mouseovers to read off specific values.

tions at four-fold degenerate sites [Figure 1A; 31]. Because mutations at such sites never alter the amino-acid sequence, these counts reflect the mutation process in the absence of protein-level selection (see Discussion for caveats about nucleotide-level selection). The expected counts of a mutation from nucleotide x to y is simply the average count of this type of mutation across all four-fold degenerate sites with parental identity x. Importantly, we count independent *occurrences* of each mutation along the branches of the tree, not the sequences with the mutation in the final alignment (Figure 1A. We also compute expected counts separately for each SARS-CoV-2 clade to account for shifts in mutation spectrum [31, 32], and apply quality-control steps to remove spurious mutations (see Methods).

The expected counts per mutation (summed across all viral clades) vary with mutation type, ranging from ∼500 for `C`→`T` to only ∼8 for `T`→`G` mutations (Figure 1B). This variation is because the SARS-CoV-2 mutation spectrum is highly biased towards specific mutation types [31, 32, 33, 34].

We compared the expected counts to the actual observed counts of mutations averaged across sites (Figure 1). For synonymous mutations, the expected and actual counts are similar. But for nonsynonymous and especially stop-codon mutations, the actual counts are substantially lower than the expected counts, reflecting purifying selection for protein function.

The ratio of actual to expected counts for each mutation is related to its effect on viral fitness. The intuition is straightforward: mutations arise at all sites, but viruses with deleterious mutations are less likely to transmit and be observed in sequencing of human SARS-CoV-2. Therefore, the ratio of actual to expected counts will be one for neutral mutations, and less than one for deleterious mutations. In the Methods and Appendix, we show that under plausible assumptions about SARS-CoV-2 evolution and sampling intensity (fraction of viruses sequenced), the fitness cost of a deleterious mutation scales roughly inversely with the ratio of actual to expected counts for mutations with costs greater than a few percent. A key result is the dependence on sampling intensity: if all human SARS-CoV-2 were sequenced even deleterious mutations would have a high chance of being sampled and we would need to study the subsequent spread of the mutations to assess their fitness. But since the actual sampling intensity is ∼0.1–1% the number of times a mutation is observed reflects more subtle reductions in transmission efficiency. We quantify the effect of each mutation as the logarithm of the ratio of actual to expected counts after summing counts for all nucleotides that encode the relevant amino-acid. The statistical noise is greater for mutations with fewer expected counts: the figures in this paper show mutations with ≥ 10 expected counts unless otherwise noted, with legends linking to interactive plots that enable adjustment of this threshold.

Mutation-effect estimates are robust to subsampling natural sequences, with some evidence of epistasis in spike

We computed the correlations among mutation-effect estimates made using subsets of SARS-CoV-2 sequences from different viral clades or geographic locations. These estimates were well correlated, with some modest variation in estimates across sequence subsets (Figure 2A,B).

The modest variation in estimates from different sequence subsets could have two causes: statistical noise due to finite mutation counts, or real shifts in mutation effects during SARS-CoV-2 evolution [35, 36]. To test for statistical noise, we computed correlations with different thresholds for how many expected counts are required before making an estimate for a mutation (Figure 2C). Correlations increased with this count threshold, consistent with reduced statistical noise for larger mutation counts. But the correlation for spike mutations was consistently lower for cross-clade but not cross-geography comparisons (Figure 2C). The lower cross-clade correlation for spike appears due to epistatic shifts in mutation effects [35, 36, 37, 38, 39] or changes in the selective landscape [40] between SARS-CoV-2 clades, since the correlation is lower between clades with higher spike divergence (Figure 2D).

Despite evidence for some shifts in mutation effects in spike, for the rest of this paper we aggregate counts across viral clades to make a single estimate for each amino-acid mutation. The reason is that the accuracy of the estimates increases with the number of counts (Figure 2C), and several mutation types only have enough counts for reasonable estimates when aggregating across clades (Figure 1B). For the purposes of this paper, we deemed it preferable to have more accurate and comprehensive

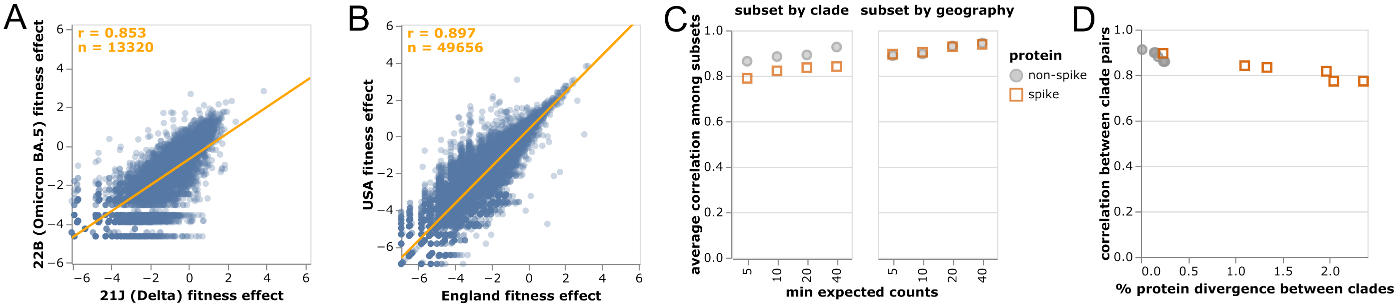


Figure 2 Correlations of mutation fitness effect estimates made using subsets of natural sequences. Correlations between estimates made **(A)** just using sequences from the Delta or Omicron BA.5 clades or **(B)** just from the USA or England. Each point is an amino-acid mutation, the orange line is a least-squares regression, and orange text at upper left shows the number of mutations and Pearson correlation coefficient. Only mutations with at least 10 expected counts are shown, which is why panels have different numbers of mutations shown (sequence subsets vary in size). **(C)** Correlations between clade or geography subsets become higher with an increasingly large threshold for minimum expected counts. Spike mutations have a worse correlation when subsetting by viral clade (plot shows average correlation over all pairwise combinations of Delta, BA.1, BA.2, and BA.5), but not when subsetting by geography (USA or England). **(D)** Correlations in estimated mutation effects decline for clades with higher protein divergence, with the effect most noticeable for spike since spike is more diverged among SARS-CoV-2 clades than other viral proteins. See https://jbloomlab.github.io/SARS2-mut-fitness/clade_corr_chart.html and https://jbloomlab.github.io/SARS2-mut-fitness/subset_corr_chart.html for versions of A and B that include all viral clades with at least 500,000 total expected counts (summed across all mutations) and have other interactive options.

pan-SARS-CoV-2 estimates than noisier clade-specific estimates for fewer mutations. However, the interactive version of Figure 2A linked in the legend enables exploration of mutations with disparate estimates among clades.

Structural and non-structural proteins are under strong purifying selection, but most accessory proteins are not

The distributions of mutation effects concur with biological intuition about how different classes of mutations impact protein function. Most synonymous mutations are nearly neutral, most stop codons are highly deleterious, and amino-acid mutations range from slightly beneficial to highly deleterious (Figure 3A).

To investigate differences among viral proteins, we computed the distributions of effects separately for each gene (Figure 3B). SARS-CoV-2 proteins are grouped into three categories: non-structural (or nsp) proteins, structural proteins (spike, M, N, and E), and accessory proteins (names prefixed with "ORF") [41]. The nonstructural and structural proteins are essential, and these proteins show strong selection against stop codons and clear although variable purifying selection against amino-acid mutations (Figure 3B; e.g., nsp13 is under stronger protein-level constraint than nsp1).

However, most accessory proteins are under little constraint (Figure 3B). Stop-codon and amino-acid mutations to ORF7a and ORF8 are not more deleterious than synonymous mutations (although recall that our estimates are only sensitive to fitness costs greater than a few percent). The lack of deleterious mutations to ORF8 is consistent with the fact that viruses with deletions in this gene have spread in humans [42] and that major variants had stop codons early in ORF8. The only accessory protein under strong purifying selection against stop codons is ORF3a (Figure 3B), for which stop codons in the first 240 residues are clearly deleterious (Figure S1). These observations concur with experiments showing SARS-CoV-2 is attenuated by deletion of ORF3a but there is little effect of deleting ORF6, ORF7a, or ORF8 [19, 43, 44]. However, ORF3a's function must be relatively insensitive to its protein sequence, since other than selection against stop codons there is only amino-acid level constraint at a few sites like 135 and 138 (Figure S1). Observations such as these could help guide experimental studies to better understand protein function.

Mutation-effect estimates correlate with experiments

We examined how the mutation effects estimated using our approach compare with prior high-throughput deep mutational scanning measurements. For spike, two distinct experimental methodologies have been used to characterize large numbers of mutations: yeast display of the receptor-binding domain (RBD) [8, 45] and spike pseudotyped lentiviruses [9]. For Mpro (also known as nsp5 or 3CLpro), two different labs have performed deep mutational scanning using the same basic methodology of assaying protease cleavage in yeast [21, 22].

For spike, our estimates from natural sequences correlate with each set of experiments almost as well as the two experimental methodologies correlate with each other (Figure 4A). If we increase the minimum expected counts from 10 to 20 and subset only on mutations shared among all three data sets, then the correlations between the estimates and experiments ($r = 0.66$) become even closer to the cross-experiment correlations ($r = 0.72$; see interactive version of Figure 4A linked in legend). Some of the mutations with the greatest divergence between our sequence-based estimates and the deep mutational scanning likely represent experimental artifacts. For instance, P527L, which is favorable in the RBD deep mutational scan but deleterious in the sequence-based estimates and full-spike scan, is at the C-terminus of the yeast-displayed RBD [8] where it may adopt a non-native conformation.

The sequence-based estimates for Mpro also correlate with the deep mutational scans for that protein, although in this case the two experiments correlate substantially better with each other than with our estimates (Figure 4B). Because both Mpro experiments use a similar yeast-based methodology [21, 22] it is possible that the higher correlation of the experiments with each other than the sequence-based estimates reflects shared artifacts of the yeast experiments. In particular, some Mpro mutations estimated to have deleterious effects in natural sequences are well

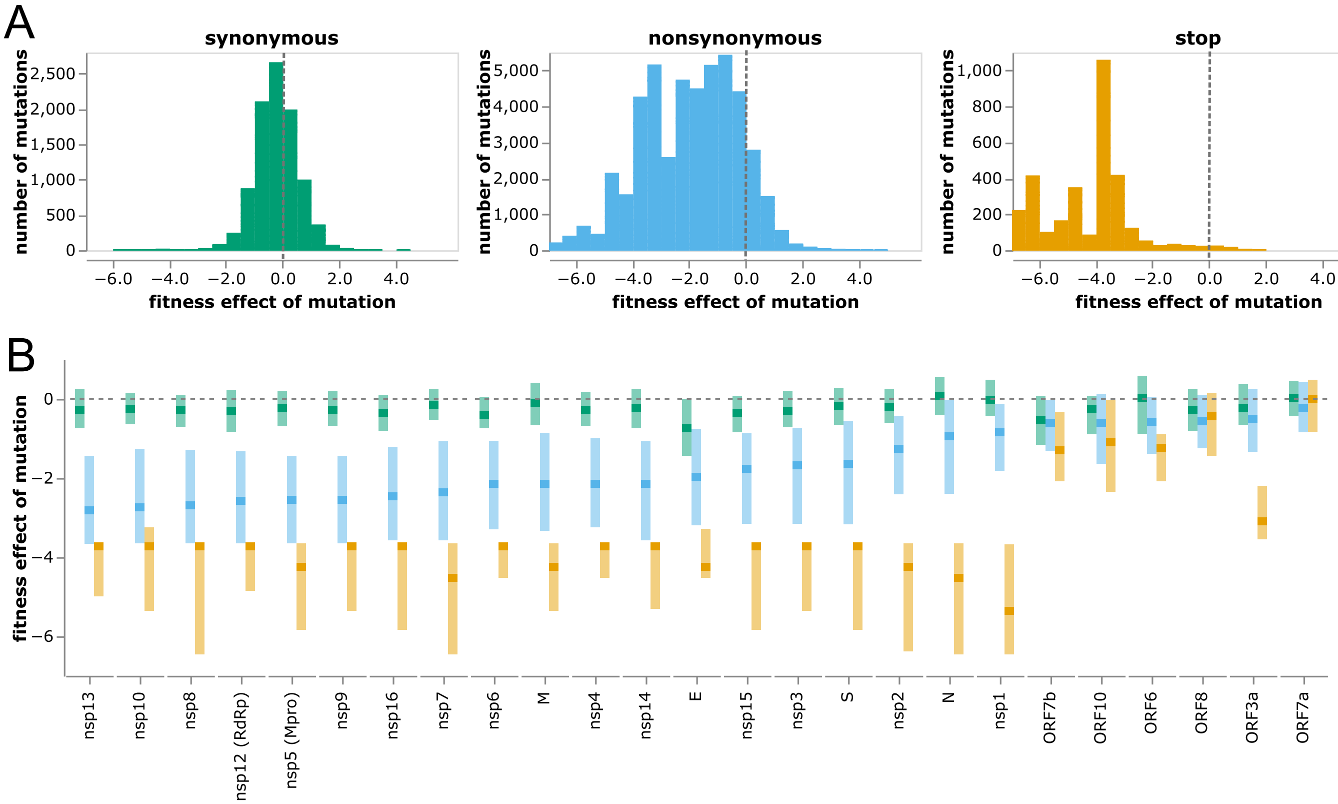


Figure 3 Distribution of effects of different classes of mutations. **(A)** Histograms of effects of synonymous, nonsynonymous, and stop-codon mutations across all viral genes. Neutral mutations have effects of zero (dashed gray vertical lines), and deleterious mutations have negative effects. **(B)** Effects of each class of mutation for each viral gene. Dark squares indicate the median effect, and the lighter rectangles span the interquartile range. Mutation types are color-coded as in panel A. See https://jbloomlab.github.io/SARS2-mut-fitness/effects_histogram.html and https://jbloomlab.github.io/SARS2-mut-fitness/effects_dist.html for plots that allow adjustment of the expected-count cutoff and other interactive options (such as separate histograms for each gene).

tolerated in the yeast experiments. This difference could arise if the yeast experiments only capture some of the constraints on Mpro in the context of actual virus. For instance, a stop codon at Q306 is well tolerated in both deep mutational scans but extremely disfavorable in our sequence-based estimates, and would be highly deleterious to actual virus as it would truncate the polyprotein. Similarly, K61N is well tolerated in the deep mutational scans but extremely disfavorable in our estimates, possibly because in the full viral polyprotein this residue mediates important interactions between Mpro and nsp7-10 [46]

Fixed mutations tend to have beneficial or neutral effects

Amino-acid mutations that have fixed in at least one viral clade are estimated to mostly have neutral or beneficial effects, whereas most other mutations are deleterious (Figure S2). This fact is unsurprising: viral lineages that expand into new clades do so because they have acquired beneficial mutations while avoiding deleterious ones [47, 48, 49]. But the fact that the beneficial effects of fixed mutations are correctly estimated by our approach, which simply counts mutation occurrences and does not incorporate information on lineage size, demonstrates such mutations occur independently in many viral lineages that are more successful than average.

Most fixed mutations are estimated to be beneficial regardless of whether estimates are made using all viral clades, or just clades that did not fix the mutation (Figure S3). However, a few beneficial fixed mutations show epistatic entrenchment [38, 50] in the sense that they are not particularly beneficial in clades in which they did not fix (Figure S3). The most striking example is S373P in spike, which has experimentally been shown to be neutral or slightly deleterious in pre-Omicron clades, but strongly beneficial in the Omicron clades in which it fixed [45, 36].

Interactive exploration of amino-acid fitnesses

To enable easy access to the mutation-effect estimates, we created an interactive plots to enable exploration of the data for each protein. A static view of one of these plots is in Figure 5; see https://jbloomlab.github.io/SARS2-mut-fitness for interactive versions for all proteins. These plots enable both high-level inspection of functional constraint across each protein, and detailed interrogation of the effects of specific mutations.

Discussion

Enough SARS-CoV-2 viruses have now been sequenced that many independent occurrences of every tolerated single-nucleotide mutation have been observed along the viral phylogeny. Here we have described a new approach that leverages this fact to estimate the effects of these mutations. In essence, we treat natural evolution as a deep mutational scan, with the millions of publicly available SARS-CoV-2 sequences providing a readout of this experiment. The key is simply to calculate how many times each mutation has been "tested" along the history

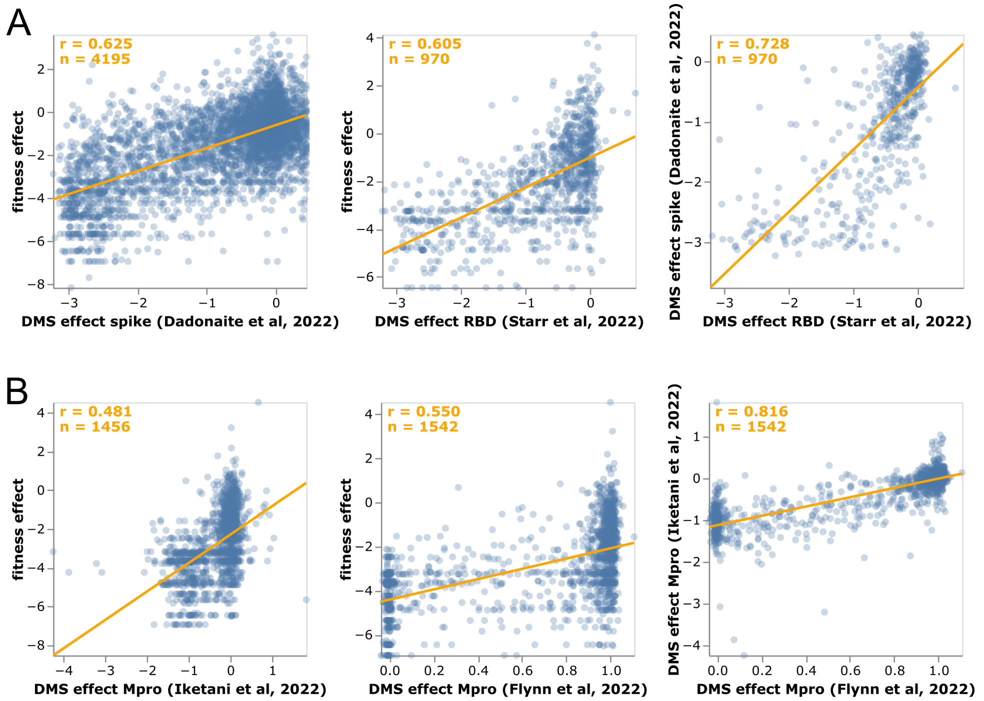


Figure 4 Correlation of mutation-effect estimates with experimental deep mutational scanning measurements for **(A)** the full spike [9] or its RBD [45], and **(B)** Mpro [21, 22]. Each point is an amino-acid mutation, the orange line is a least-squares regression, and orange text in the upper left shows the number of mutations and Pearson correlation coefficient. Each sub-panel shows a different set of mutations (depending on which mutations were measured in that experiment). See https://jbloomlab.github.io/SARS2-mut-fitness/dms_S_corr.html and https://jbloomlab.github.io/SARS2-mut-fitness/dms_nsp5_corr.html for plots that enable subsetting on just mutations shared across all datasets and other interactive options such as mousing over points to see mutation identities. The experiments in [45, 21] measure multiple phenotypes and these plots show the effect of each mutation averaged across these phenotypes; see https://jbloomlab.github.io/SARS2-mut-fitness/dms_S_all_corr.html and https://jbloomlab.github.io/SARS2-mut-fitness/dms_nsp5_all_corr.html for plots that show each phenotype separately.

of sampled viral sequences, and compare that expectation to the actual observations of the mutation among viruses sufficiently fit to have been sequenced in actual human infections.

The resulting estimates of mutational effects are robust to subsetting on specific viral clades or geographies, and correlate well with experimental measurements. In broad strokes, the mutation effects illuminate patterns of constraint: for instance, there is strong selection on structural and non-structural proteins, but only limited purifying selection on the accessory proteins.

However, the real value of our approach is in the detailed maps of effects of specific mutations to all viral proteins, including proteins with poorly understood functions not easily characterized in the lab. These maps will be of value for designing drugs that target constrained sites, interpreting the consequences of mutations observed during viral surveillance, and guiding experiments to mechanistically characterize protein function.

There are several caveats to our approach. First, because the number of observations of any given mutation is small compared to the millions of SARS-CoV-2 sequences being analyzed, our approach requires careful quality control of publicly available sequences to remove those affected by sequencing errors. Second, we assume the rate of each type of nucleotide mutation is uniform across the viral genome, and neglect higher-order context that may influence mutation rate [51, 52]. Likewise, we neglect constraint on nucleotide identity beyond the encoded protein sequence [53, 54]. Third, the exact relationship between the statistics we calculate and viral fitness depend on the fraction of all infections that are sequenced (sampling intensity) and viral population dynamics. Although we derive this relationship, we do not adjust for sampling intensity and population dynamics when estimating mutation effects. Fourth, we make a single estimate for each mutation across all SARS-CoV-2, neglecting the epistasis that can affect some mutations [35, 36]. Finally there are a few technical caveats to how we count mutations that are discussed in the Methods section.

Conceptually, our approach differs from prior methods that aim to identify beneficial SARS-CoV-2 mutations associated with viral clades that increase in frequency [23, 24, 25]. Those methods draw information primarily from what happens downstream of a mutation. In contrast, we treat all mutations equivalently regardless of whether they are on a tip node or at the base of a large clade. Our approach is better for estimating effects of deleterious or nearly neutral mutations, but clade-growth methods may be better for beneficial mutations. In particular, clade size carries information beyond that contained in mutation counts alone (Figure S4). Hopefully future work can combine mutation-counting and clade-growth methods for even better estimates of SARS-CoV-2 mutation effects. Note our approach is conceptually similar to estimating fitness costs of HIV or polio mutations from mutation-selection balance in deep sequencing of intra-population viral quasispecies [55, 56], except we analyze mutation occurrences rather than frequencies to account for the phylogenetic structure and genetic hitchhiking that characterize global SARS-CoV-2 evolution.

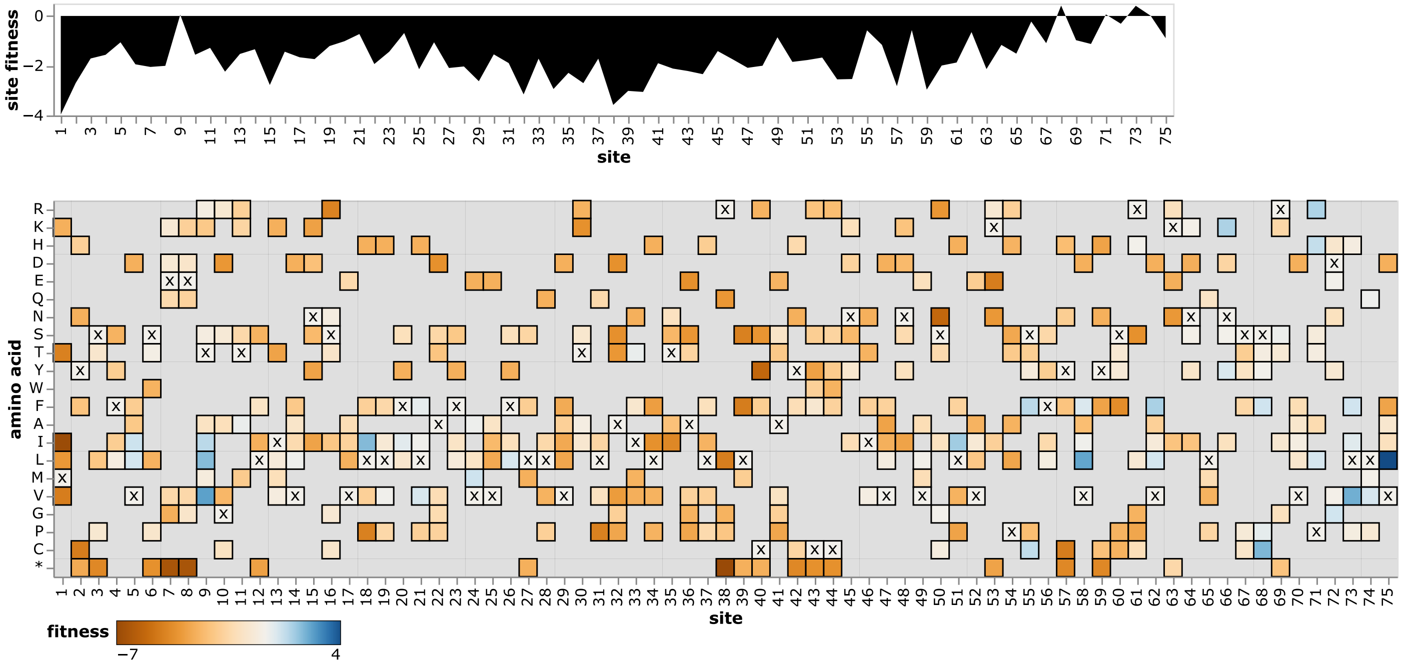


Figure 5 Effects of amino-acid mutations to E protein. The area plot at top shows the average effects of mutations at each site, and the heatmap shows the effects of specific amino acids, with **x** denoting the amino-acid identity in the Wuhan-Hu-1 strain. See https://jbloomlab.github.io/SARS2-mut-fitness/E.html for an interactive version of this plot that enables zooming, mouseovers, adjustment of the minimum expected count threshold, and layering of stop codon effects on the site plot. See https://jbloomlab.github.io/SARS2-mut-fitness for comparable interactive plots for all SARS-CoV-2 proteins.

The power of the approach we have described will increase with more viral sequencing. SARS-CoV-2 is the first virus with enough sequences that every tolerated mutation is observed multiple independent times. As costs drop, it is easy to imagine a future with even more viral sequences. As this occurs, viral genomic sequencing—which has traditionally been used primarily to track evolution and spread—will also become an increasingly precise tool to determine the effects of specific mutations.

Methods

Code and data availability

See the GitHub repository at https://github.com/jbloomlab/SARS2-mut-fitness for the computer code and processed data (eg, fitness estimates and mutation counts). That repository contains a README with links to specific data files as well as a description of the computational pipeline. See https://github.com/jbloomlab/SARS2-mut-fitness/blob/main/results/aa_fitness/aa_fitness.csv final estimates of amino-acid fitnesses across all clades; other intermediate data files are also provided in the GitHub repository. The specific version of the repository used for this paper is tagged as "bioRxiv-v1" on GitHub (https://github.com/jbloomlab/SARS2-mut-fitness/tree/bioRxiv-v1) The pipeline is fully reproducible, and is run using `snakemake` [57] with interactive plots rendered using `altair` [58].

The interactive plots are rendered at https://jbloomlab.github.io/SARS2-mut-fitness via GitHub pages.

Counting mutations along the phylogenetic tree

We counted occurrences of each mutation in each viral clade using the UShER pre-built mutation-annotated tree [28, 29, 30] from Dec-18-2022 (http://hgdownload.soe.ucsc.edu/goldenPath/wuhCor1/UShER_SARS-CoV-2/2022/12/18/public-2022-12-18.all.masked.nextclade.pangolin.pb.gz), which contains all ∼6.5-million SARS-CoV-2 sequences that are available in public databases. To make these counts at a per-clade level, we first subsetted the mutation-annotated tree on all sequences for each Nexstrain clade [59], retained only clades with at least 10^4 sequences, and then used the matUtils program distributed with UShER to extract the nucleotide mutations on every branch of the each clade-subsetted mutation-annotated tree. For the analyses by geographic location (Figure 2), we subsetted on all sequences that began with "USA" or "England" as these were the two locations with the most publicly available sequences.

We then performed quality control by ignoring any branch that met any of the following criteria:

- it had more than four nucleotide mutations;
- it contained more than one nucleotide mutation that was a reversion to the Wuhan-Hu-1 reference sequence;
- it contained more than one nucleotide mutation that was a reversion to the founder sequence for that clade as provided at https://raw.githubusercontent.com/neherlab/SC2_variant_rates/7e738194a8c6592082f1caa9a6ca70cb68289790/data/clade_gts.json by [34];
- it contained more than one nucleotide mutation to the same codon.

The rationale for the first exclusion is that highly mutated branches are often indicative of sequencing errors, and the rationale for the second and third exclusions is that excess reversions can arise from base-calling pipelines that erroneously call low-coverage sites as reference. We ignore branches with multiple nucleotide mutations to the same codon (this is very rare) because as detailed below our method is only designed to make estimates for mutations that represent single-nucleotide changes from the clade founder. Note also that the mutation-annotated tree does not include insertion or deletion mutations, and so we only consider (and make estimates for) point mutations.

We then specified for exclusion certain mutations and sites that are prone to sequencing or base-calling errors. Specifically, we excluded

- the sites specified in Table S1 of [60] as being error prone;
- sites 5629, 6851, 7328, 28095, and 29362 since they had very high error rates in some clades;
- the problematic sites listed at https://github.com/W-L/ProblematicSites_SARS-CoV2, which are masked in the pre-built mutation-annotated tree;
- for each clade, the clade-specific sites listed in https://github.com/jbloomlab/SARS2-mut-fitness/blob/main/data/usher_masked_sites.yaml, which are masked in the pre-built mutation-annotated tree;
- for each clade, any mutation that was a reversion from the clade

founder to the Wuhan-Hu-1 reference, and the reverse complements of these mutations.

The last exclusion criteria is because some bioinformatics pipelines called low-coverage sites as reference.

See https://github.com/jbloomlab/SARS2-mut-fitness/blob/main/results/mutation_counts/aggregated.csv for the final counts of each nucleotide mutation in each clade; note that this file also contains excluded mutations.

Calculation of expected counts

To calculate the expected counts for each nucleotide mutation, we analyzed just the four-fold degenerate sites in each clade in an approach paralleling that of [31]. Specifically, we identify all non-excluded four-fold degenerate sites in each clade founder. We then count nucleotide mutations just at those sites in each clade, and calculate the expected per-site number of mutations from nucleotide x to y as the total number of x to y mutations at four-fold degenerate sites divided by the number of four-fold degenerate sites with x as the parental identity. This analysis is done at the clade level for two reasons: referencing mutations to the clade founder (rather than the Wuhan-Hu-1 reference) limits problem with the approach that would arise at sites that substitute multiple times in the history of a sequence (since each clade is a relatively high-identity group multiple mutations at the same site within a clade are very rare), and because it is know that SARS-CoV-2 mutation rates vary somewhat among clades [31, 32]. We only retain clades with at least 5000 mutations at four-fold degenerate sites in order to avoid inaccurate estimates of expected counts due to low sampling of mutations.

Mutational effects from actual versus expected counts

To estimate the effects of mutations, we simply compare the expected counts of each nucleotide mutation to the actual counts in the pre-built mutation-annotated tree. See https://github.com/jbloomlab/SARS2-mut-fitness/blob/main/results/expected_vs_actual_mut_counts/expected_vs_actual_mut_counts.csv for these expected versus actual counts on a per-clade basis; note that this file also includes counts at excluded sites.

To estimate the effects of mutations, we first sum the counts of all non-excluded nucleotide mutations that encode each amino-acid mutation to convert the nucleotide counts to amino-acid counts. In doing this, we exclude any mutations that are not from the clade-founder codon identity: in other words, we ignore sequences with histories that involve multiple mutations at the same codon in the same clade (this is a caveat of the approach, although because each clade is relatively high identity it does not have a major effect). For the overall estimates reported in this paper, we also sum these counts across all retained clades; for the analyses in Figure 2 we also make estimates without summing across clades and only for counts from sequences from specific geographic locations. We then compute the estimated fitness Δf of each mutation as simply the natural logarithm of the ratio of actual to expected counts after adding a pseudocount of $P - 0.5$ to each count, namely $\Delta f = \log\left(\frac{n_{actual}+P}{n_{expected}+P}\right)$.

Note that these mutation-effect estimates will have more statistical noise the smaller the value of the expected counts for each mutation. Therefore, we also track the expected counts alongside the estimates. In this paper, we only show estimates for mutations with expected counts of at least 10 unless otherwise noted. However, the figures link to interactive legends that allow adjustment of this threshold: larger values (eg, 20 or more) will lead to slightly more accurate estimates but drop some mutations, lower values can be used if you need a noisier estimate for a mutation that has less then 10 expected counts.

See https://github.com/jbloomlab/SARS2-mut-fitness/blob/main/results/aa_fitness/aamut_fitness_all.csv for the estimates of amino-acid mutation effects across all clades, and see https://github.com/jbloomlab/SARS2-mut-fitness/blob/main/results/aa_fitness/aamut_fitness_by_clade.csv for the clade-specific estimates. The all-clade estimates of mutation effects are what are shown in Figure 3.

For the clade correlations plotting in Figure 2, we only include clades with at least 5×10^5 expected counts across all sites, as only these clades have enough counts for reasonable per-clade estimates.

Mutation effects to amino-acid fitnesses

For the final estimates of amino-acid fitnesses shown in the heatmaps such as in Figure 5, we need a single estimate for each amino acid. This is straightforward for sites that have the same amino-acid identity in all clade founders: the "wildtype" residue shared across all clades has a fitness of zero, and all other amino acids have fitnesses equal to the effect of mutating from the "wildtype" to that amino acid. However, for sites that change amino-acid identity between clade founders, things are more complicated and we need to take the extra step below.

For each clade have estimated the change in fitness Δf_{xy} caused by mutating a site from amino-acid x to y, where x is the amino acid in the clade founder sequence. For each such mutation, we also have n_{xy} which is the number of expected mutations from the clade founder amino acid x to y. These n_{xy} values are important because they give some estimate of our "confidence" in the Δf_{xy} values: if a mutation has high expected counts (large n_{xy}) then we can estimate the change in fitness caused by the mutation more accurately, and if n_{xy} is small then the estimate will be much noisier.

However, we would like to aggregate the data across multiple clades to estimate amino-acid fitness values at a site under the assumption that these are constant across clades. Things get complicated if not all clade founders have the same amino acid identity at a site. For instance, let's say at our site of interest, the clade founder amino acid is x in one clade and z in another clade. For each clade we then have a set of Δf_{xy} and n_{xy} values for the first clade (where y ranges over the 20 amino acids, including stop codon, that aren't x), and another set of up to 20 Δf_{zy} and n_{zy} values for the second clade (where y ranges over the 20 amino acids that aren't z).

From these sets of mutation fitness changes, we'd like to estimate the fitness f_x of each amino acid x, where the f_x values satisfy $\Delta f_{xy} = f_y - f_x$ (in other words, a higher f_x means higher fitness of that amino acid). When there are multiple clades with different founder amino acids at the site, there is no guarantee that we can find f_x values that precisely satisfy the above equation since there are more Δf_{xy} values than f_x values and the Δf_{xy} values may have noise (and is some cases even real shifts among clades due to epistasis). Nonetheless, we can try to find the f_x values that come closest to satisfying the above equation.

First, we choose one amino acid to have a fitness value of zero, since the scale of the f_x values is arbitrary and there are really only 20 unique parameters among the 21 f_x values (there are 21 amino acids since we consider stops, but we only measure differences among them, not absolute values). Typically if there was just one clade, we would set the wildtype value of $f_x = 0$ and then for mutations to all other amino acids y we would simply have $f_y = \Delta f_{xy}$. However, when there are multple clades with different founder amino acids, there is no longer a well defined "wildtype". So we choose the most common non-stop parental amino-acid for the observed mutations and set that to zero. In other words, we find x that maximizes $\sum_y n_{xy}$ and set that f_x value to zero.

Next, we choose the f_x values that most closely match the measured mutation effects, weighting more strongly mutation effects with higher expected counts (since these should be more accurate). Specifically, we define a loss function as

$$L = \sum_x \sum_{y \neq x} n_{xy} \left(\Delta f_{xy} - \left[f_y - f_x\right]\right)^2$$

where we ignore effects of synonymous mutations (the $x \neq y$ term in second summand) because we are only examining protein-level effects. We then use numerical optimization to find the f_x values that minimize that loss L.

Finally, we would still like to report an equivalent of the n_{xy} values for the Δf_{xy} values that give us some sense of how accurately we have estimated the fitness f_x of each amino acid. To do that, we tabulate $N_x = \sum_y \left(n_{xy} + n_{yx}\right)$ as the total number of mutations either from or to amino-acid x as the "count" for the amino acid. Amino acids with larger values of N_x should have more accurate estimates of f_x.

See https://github.com/jbloomlab/SARS2-mut-fitness/blob/main/results/aa_fitness/aa_fitness.csv for these overall amino-acid fitness estimates.

Site numbering and protein naming

All sites are numbered according to the sequential Wuhan-Hu-1 reference numbering scheme, using the reference sequence at http://hgdownload.soe.ucsc.edu/goldenPath/wuhCor1/bigZips/wuhCor1.fa.gz. The protein annotations are taken from the associated GTF at http://hgdownload.soe.ucsc.edu/goldenPath/wuhCor1/bigZips/genes/ncbiGenes.gtf.gz. Those protein annotations refer to the polyproteins encoding the non-structural proteins as ORF1a and ORF1ab. To convert to from ORF1ab numbering/naming to the nsp-based naming (eg, nsp1, nsp2, etc) we use the conversions specified under "orf1ab_to_nsps" in https://github.com/jbloomlab/SARS2-mut-fitness/blob/main/config.yaml, which are in turn taken from Theo Sanderson's annotations at https://github.com/theosanderson/Codon2Nucleotide/blob/main/src/App.js.

Comparison to deep mutational scanning

Deep mutational scanning data were taken from published studies [9, 45, 21, 22], using the data at the links specified under the "dms_datasets" key in https://github.com/jbloomlab/SARS2-mut-fitness/blob/main/config.yaml. For the spike deep mutational scanning [9] we only included mutations with "times seen" values of at least three in the deep mutational scanning. The RBD data [45] include measurements for two phenotypes (ACE2 affinity and RBD expression), and one of the Mpro studies [21] includes measurements for three different phenotypes in yeast (growth, FRET, and transcription factor activity). In both cases, Figure 4 shows the effect averaged across all phenotypes measured by the study. For plots that break the correlations out by phenotype, see https://jbloomlab.github.io/SARS2-mut-fitness/dms_S_all_corr.html and https://jbloomlab.github.io/SARS2-mut-fitness/dms_nsp5_all_corr.html.

Derivation of relationship between actual to expected count ratio and viral fitness

The ratio of actual to expected counts that we calculate in this paper is related to the probability that we observe a viral lineage containing an occurrence of a specific mutation among sequenced human SARS-CoV-2. This probability depends on three factors: the fitness effect of the mutation, the fraction of all SARS-CoV-2 viruses that are sequenced (sampling intensity), and the growth dynamics of the viral population. In the supplementary appendix, we derive the approximate relationship between this probability as a function of the fitness cost s and sampling intensity ϵ for deleterious mutations for both a constant and exponentially growing viral population.

We show that for a constant viral population size, the probability of observing a lineage containing a deleterious mutation with cost s is roughly $\frac{\epsilon}{s+\epsilon}$ when $s^2 > \epsilon$, and more weakly dependent on s for smaller fitness costs (when $s^2 < \epsilon$). The intuitive explanation is that the average size of a mutant lineage with fitness cost s is $1/s$ and we basically ask whether we sample the lineage before it disappears. If we sample more intensely (larger ϵ), whether a lineage gets sampled depends primarily on the stochastic dynamics and little on the fitness effect. With a typical sampling intensity for SARS-CoV-2 between 1/1000 and 1/100, this means our approach is sensitive to fitness effects larger than a few percent per serial interval; mutations with fitness costs smaller than that will not show an appreciable difference from neutral mutations in their ratio of actual to expected accounts.

In an exponentially growing population, the probability of observing a mutant lineage with fitness cost s again scales as $\sim \frac{\epsilon}{\epsilon+s}$ if $sT > 1$, where T is the time over which the variant has expanded. If T is $\sim$ months, that is 20 generations, which again corresponds to s of at least a few percent for $sT > 1$. For mutations with smaller fitness costs, the dependence scales more as $\sim \epsilon\,(1 - sT)$.

Overall, these calculations indicate that for multiple different growth dynamics of the viral population, the ratio of expected to actual counts will scale inversely with the fitness cost of deleterious mutations for mutations with costs that exceed a few percent. Note that the approach we use in this paper does not account for variation in sampling intensity across space or time, does not attempt to adjust for changes in viral growth dynamics over time, uses the heuristic formula of calculating the effect as the log ratio of counts, and applies this same formula to all mutations regardless of whether they are deleterious, neutral, or beneficial. A more complete derivation might try to calculate the fitness effects from the full distribution of lineage sizes more rigorously and incorporate information about the sampling intensity and viral growth dynamics. However, such a derivation (if possible at all) is beyond the scope of this study, and we also note that good empirical data is generally lacking to precisely account for sampling intensity and viral growth dynamics over the full span of time and space from which the sequences we analyze are drawn. The key point of the derivations for our current study is simply that our approach should be sensitive to detecting the effects of mutations with fitness costs greater than a few percent.

Acknowledgments

We thank the UShER team for promptly answering and addressing GitHub issues related to use of this package and its pre-built mutation-annotated tree. The work of JDB was supported in part by the NIH/NIAID under grants U19AI171399 and R01AI141707, and contract 75N93021C00015. JDB is an Investigator of the Howard Hughes Medical Institute.

Competing interests

JDB is on the scientific advisory boards of Apriori Bio, Aerium Therapeutics, Invivyd, the Vaccine Company, and Oncorus. JDB receives royalty payments as an inventor on Fred Hutch licensed patents related to deep mutational scanning of viral proteins.

Literature Cited

[1] Harvey WT, Carabelli AM, Jackson B, Gupta RK, Thomson EC, Harrison EM, et al. SARS-CoV-2 variants, spike mutations and immune escape. Nature Reviews Microbiology. 2021;19(7):409–424.

[2] Abdool Karim SS, de Oliveira T. New SARS-CoV-2 variants—clinical, public health, and vaccine implications. New England Journal of Medicine. 2021;384(19):1866–1868.

[3] DeGrace MM, Ghedin E, Frieman MB, Krammer F, Grifoni A, Alisoltani A, et al. Defining the risk of SARS-CoV-2 variants on immune protection. Nature. 2022;605(7911):640–652.

[4] Moghadasi SA, Heilmann E, Khalil AM, Nnabuife C, Kearns FL, Ye C, et al. Transmissible SARS-CoV-2 variants with resistance to clinical protease inhibitors. bioRxiv. 2022;DOI 10.1101/2022.08.07.503099.

[5] Iketani S, Mohri H, Culbertson B, Hong SJ, Duan Y, Luck MI, et al. Multiple pathways for SARS-CoV-2 resistance to nirmatrelvir. Nature. 2022;DOI 10.1038/s41586-022-05514-2.

[6] Hiscox JA, Khoo SH, Stewart JP, Owen A. Shutting the gate before the horse has bolted: is it time for a conversation about SARS-CoV-2 and antiviral drug resistance? Journal of Antimicrobial Chemotherapy. 2021;76(9):2230–2233.

[7] Weisblum Y, Schmidt F, Zhang F, DaSilva J, Poston D, Lorenzi JC, et al. Escape from neutralizing antibodies by SARS-CoV-2 spike protein variants. eLife. 2020;9:e61312.

[8] Starr TN, Greaney AJ, Hilton SK, Ellis D, Crawford KH, Dingens AS, et al. Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. Cell. 2020;182(5):1295–1310.

[9] Dadonaite B, Crawford KH, Radford CE, Farrell AG, Timothy CY, Hannon WW, et al. A pseudovirus system enables deep mutational scanning of the full SARS-CoV-2 spike. bioRxiv. 2022;DOI 10.1101/2022.10.13.512056.

[10] Greaney AJ, Starr TN, Gilchuk P, Zost SJ, Binshtein E, Loes AN, et al. Complete mapping of mutations to the SARS-CoV-2 spike receptor-binding domain that escape antibody recognition. Cell Host & Microbe. 2021;29(1):44–57.

[11] Cao Y, Jian F, Wang J, Yu Y, Song W, Yisimayi A, et al. Imprinted SARS-CoV-2 humoral immunity induces convergent Omicron RBD evolution. Nature. 2022;DOI 10.1038/s41586-022-05644-7.

[12] Greaney AJ, Starr TN, Bloom JD. An antibody-escape estimator for mutations to the SARS-CoV-2 receptor-binding domain. Virus Evolution. 2022;8(1):veac021.

[13] Tzou PL, Tao K, Pond SLK, Shafer RW. Coronavirus Resistance Database (CoV-RDB): SARS-CoV-2 susceptibility to monoclonal antibodies, convalescent plasma, and plasma from vaccinated persons. Plos one. 2022;17(3):e0261045.

[14] Starr TN, Czudnochowski N, Liu Z, Zatta F, Park YJ, Addetia A, et al. SARS-CoV-2 RBD antibodies that maximize breadth and resistance to escape. Nature. 2021;597(7874):97–102.

[15] Rappazzo CG, Tse LV, Kaku CI, Wrapp D, Sakharkar M, Huang D, et al. Broad and potent activity against SARS-like viruses by an engineered human monoclonal antibody. Science. 2021;371(6531):823–829.

[16] Cao Y, Jian F, Zhang Z, Yisimayi A, Hao X, Bao L, et al. Rational identification of potent and broad sarbecovirus-neutralizing antibody cocktails from SARS convalescents. Cell reports. 2022;41(12):111845.

[17] Thorne LG, Bouhaddou M, Reuschl AK, Zuliani-Alvarez L, Polacco B, Pelin A, et al. Evolution of enhanced innate immune evasion by SARS-CoV-2. Nature. 2022;602(7897):487–495.

[18] Syed AM, Taha TY, Tabata T, Chen IP, Ciling A, Khalid MM, et al. Rapid assessment of SARS-CoV-2–evolved variants using virus-like particles. Science. 2021;374(6575):1626–1632.

[19] McGrath M, Xue Y, Dillen C, Oldfield L, Assad-Garcia N, Zaveri J, et al. SARS-CoV-2 Variant Spike and accessory gene mutations

alter pathogenesis. Proceedings National Academy of Sciences USA. 2022;119:e2204717119.

[20] Tao K, Tzou PL, Nouhin J, Bonilla H, Jagannathan P, Shafer RW. SARS-CoV-2 antiviral therapy. Clinical microbiology reviews. 2021;34(4):e00109–21.

[21] Flynn JM, Samant N, Schneider-Nachum G, Barkan DT, Yilmaz NK, Schiffer CA, et al. Comprehensive fitness landscape of SARS-CoV-2 Mpro reveals insights into viral resistance mechanisms. eLife. 2022;11:e77433. doi:10.7554/eLife.77433.

[22] Iketani S, Hong SJ, Sheng J, Bahari F, Culbertson B, Atanaki FF, et al. Functional map of SARS-CoV-2 3CL protease reveals tolerant and immutable sites. Cell Host & Microbe. 2022;30(10):1354–1362.

[23] Obermeyer F, Jankowiak M, Barkas N, Schaffner SF, Pyle JD, Yurkovetskiy L, et al. Analysis of 6.4 million SARS-CoV-2 genomes identifies mutations associated with fitness. Science. 2022;376(6599):1327–1332. doi:10.1126/science.abm1208.

[24] Lee B, Sohail MS, Finney E, Ahmed SF, Quadeer AA, McKay MR, et al. Inferring effects of mutations on SARS-CoV-2 transmission from genomic surveillance data. medRxiv. 2022;10.1101/2021.12.31.21268591v1:2021.12.31.21268591. doi:DOI 10.1101/2021.12.31.21268591.

[25] Maher MC, Bartha I, Weaver S, Di Iulio J, Ferri E, Soriaga L, et al. Predicting the mutational drivers of future SARS-CoV-2 variants of concern. Science translational medicine. 2022;14(633):eabk3445.

[26] Rodriguez-Rivas J, Croce G, Muscat M, Weigt M. Epistatic models predict mutable sites in SARS-CoV-2 proteins and epitopes. Proceedings of the National Academy of Sciences. 2022;119(4):e2113118119.

[27] Thadani NN, Gurev S, Notin P, Youssef N, Rollins NJ, Sander C, et al. Learning from pre-pandemic data to forecast viral antibody escape. bioRxiv. 2022;DOI 10.1101/2022.07.21.501023.

[28] McBroome J, Thornlow B, Hinrichs AS, Kramer A, De Maio N, Goldman N, et al. A daily-updated database and tools for comprehensive SARS-CoV-2 mutation-annotated trees. Molecular Biology and Evolution. 2021;38(12):5819–5824.

[29] Turakhia Y, Thornlow B, Hinrichs AS, De Maio N, Gozashti L, Lanfear R, et al. Ultrafast Sample placement on Existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. Nature Genetics. 2021;53(6):809–816.

[30] Lanfear R. A global phylogeny of SARS-CoV-2 sequences from GISAID. Zenodo. 2020;DOI 10.5281/zenodo.3958883.

[31] Bloom JD, Beichman AC, Neher RA, Harris K. Evolution of the SARS-CoV-2 mutational spectrum. bioRxiv. 2022;DOI 10.1101/2022.11.19.517207.

[32] Ruis C, Peacock TP, Polo LM, Masone D, Alvarez MS, Hinrichs AS, et al. Mutational spectra distinguish SARS-CoV-2 replication niches. bioRxiv. 2022;DOI 10.1101/2022.09.27.509649.

[33] De Maio N, Walker CR, Turakhia Y, Lanfear R, Corbett-Detig R, Goldman N. Mutation rates and selection on synonymous mutations in SARS-CoV-2. Genome Biology and Evolution. 2021;13(5):evab087.

[34] Neher RA. Contributions of adaptation and purifying selection to SARS-CoV-2 evolution. Virus Evolution. 2022;8(2):veac113.

[35] Starr TN, Greaney AJ, Hannon WW, Loes AN, Hauser K, Dillen JR, et al. Shifting mutational constraints in the SARS-CoV-2 receptor-binding domain during viral evolution. Science. 2022;377:420–424.

[36] Moulana A, Dupic T, Phillips AM, Chang J, Nieves S, Roffler AA, et al. Compensatory epistasis maintains ACE2 affinity in SARS-CoV-2 Omicron BA. 1. Nature Communications. 2022;13:7011.

[37] Pollock DD, Thiltgen G, Goldstein RA. Amino acid coevolution induces an evolutionary Stokes shift. Proceedings of the National Academy of Sciences. 2012;109(21):E1352–E1359.

[38] Shah P, McCandlish DM, Plotkin JB. Contingency and entrenchment in protein evolution under purifying selection. Proceedings of the National Academy of Sciences. 2015;112(25):E3226–E3235.

[39] Lee JM, Huddleston J, Doud MB, Hooper KA, Wu NC, Bedford T, et al. Deep mutational scanning of hemagglutinin helps predict evolutionary fates of human H3N2 influenza variants. Proceedings of the National Academy of Sciences. 2018;115(35):E8276–E8285.

[40] Sun K, Tempia S, Kleynhans J, von Gottberg A, McMorrow ML, Wolter N, et al. Rapidly shifting immunologic landscape and severity of SARS-CoV-2 in the Omicron era in South Africa. Nature Communications. 2023;14(1):246.

[41] V'kovski P, Kratzel A, Steiner S, Stalder H, Thiel V. Coronavirus biology and replication: implications for SARS-CoV-2. Nature Reviews Microbiology. 2021;19(3):155–170.

[42] Su YC, Anderson DE, Young BE, Linster M, Zhu F, Jayakumar J, et al. Discovery and genomic characterization of a 382-nucleotide deletion in ORF7b and ORF8 during the early evolution of SARS-CoV-2. mBio. 2020;11(4):e01610–20.

[43] Silvas JA, Vasquez DM, Park JG, Chiem K, Allué-Guardia A, Garcia-Vilanova A, et al. Contribution of SARS-CoV-2 accessory proteins to viral pathogenicity in K18 human ACE2 transgenic mice. Journal of Virology. 2021;95(17):e00402–21.

[44] Liu Y, Zhang X, Liu J, Xia H, Zou J, Muruato AE, et al. A live-attenuated SARS-CoV-2 vaccine candidate with accessory protein deletions. Nature Communications. 2022;13(1):1–14.

[45] Starr TN, Greaney AJ, Stewart CM, Walls AC, Hannon WW, Veesler D, et al. Deep mutational scans for ACE2 binding, RBD expression, and antibody escape in the SARS-CoV-2 Omicron BA. 1 and BA. 2 receptor-binding domains. PLoS Pathogens. 2022;18(11):e1010951.

[46] Yadav R, Courouble VV, Dey SK, Harrison JJE, Timm J, Hopkins JB, et al. Biochemical and structural insights into SARS-CoV-2 polyprotein processing by Mpro. Science Advances. 2022;8(49):eadd2191.

[47] Łuksza M, Lässig M. A predictive fitness model for influenza. Nature. 2014;507(7490):57–61.

[48] Koelle K, Rasmussen DA. The effects of a deleterious mutation load on patterns of influenza A/H3N2's antigenic evolution in humans. Elife. 2015;4:e07361.

[49] Huddleston J, Barnes JR, Rowe T, Xu X, Kondor R, Wentworth DE, et al. Integrating genotypes and phenotypes improves long-term forecasts of seasonal influenza A/H3N2 evolution. Elife. 2020;9:e60067.

[50] Starr TN, Flynn JM, Mishra P, Bolon DN, Thornton JW. Pervasive contingency and entrenchment in a billion years of Hsp90 evolution. Proceedings of the National Academy of Sciences. 2018;115(17):4453–4458.

[51] Sadykov M, Mourier T, Guan Q, Pain A. Short sequence motif dynamics in the SARS-CoV-2 genome suggest a role for cytosine deamination in CpG reduction. Journal of Molecular Cell Biology. 2021;13(3):225–227.

[52] Beale RC, Petersen-Mahrt SK, Watt IN, Harris RS, Rada C, Neuberger MS. Comparison of the differential context-dependence of DNA deamination by APOBEC enzymes: correlation with mutation spectra in vivo. Journal of Molecular Biology. 2004;337(3):585–596.

[53] Huston NC, Wan H, Strine MS, Tavares RdCA, Wilen CB, Pyle AM. Comprehensive in vivo secondary structure of the SARS-CoV-2 genome reveals novel regulatory motifs and mechanisms. Molecular Cell. 2021;81(3):584–598.

[54] Kuo L, Masters PS. Functional analysis of the murine coronavirus genomic RNA packaging signal. Journal of Virology. 2013;87(9):5182–5192.

[55] Zanini F, Puller V, Brodin J, Albert J, Neher RA. In vivo mutation rates and the landscape of fitness costs of HIV-1. Virus Evolution. 2017;3(1):vex003. doi:10.1093/ve/vex003.

[56] Acevedo A, Brodsky L, Andino R. Mutational and fitness landscapes of an RNA virus revealed through population sequencing. Nature. 2014;505(7485):686–690.

[57] Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, et al. Sustainable data analysis with Snakemake. F1000Research. 2021;10.

[58] VanderPlas J, Granger B, Heer J, Moritz D, Wongsuphasawat K, Satyanarayan A, et al. Altair: interactive statistical visualizations for Python. Journal of Open Source Software. 2018;3(32):1057.

[59] Aksamentov I, Roemer C, Hodcroft EB, Neher RA. Nextclade: clade assignment, mutation calling and quality control for viral genomes. Journal of Open Source Software. 2021;6(67):3773.

[60] Turakhia Y, De Maio N, Thornlow B, Gozashti L, Lanfear R, Walker CR, et al. Stability of SARS-CoV-2 phylogenies. PLoS Genetics. 2020;16(11):e1009175.

Supplementary appendix deriving relationship between fitness cost and ratio of expected to actual counts

With millions of SARS-CoV-2 sequences shared publicly, almost all mutations that are tolerated by the virus are observed dozens to hundreds of times. Where on the tree and how often on the tree we observe specific mutations has information about the effects of these mutations on viral spread. The mutation rate depends on the nucleotides involved and possibly on the sequence context and other viral determinants, but for the purpose of this derivation, we will assume the neutral rate μ is known. If the mutation is neutral, the total number of times the mutation is observed on the tree is μT, where T is the total length of the tree (assuming that the mutation never reached high frequency which is true for almost all mutations, particularly when mutations are counted on a per-clade basis relative too the clade founder as done above).

If a mutation reduces fitness, the lineages descending from branches on which this mutation happened will spread more slowly than those without this mutation. As a result, the down-stream subclades are smaller and more short lived, which in turn means that they will be less likely to be sampled and represented in the tree. To infer a mutation's effect on fitness, we need to calculate how the probability of observation depends on this fitness effect.

For a mutation to be represented in the tree, one of its descendants has to be sampled and sequenced. If the total number of descendants is w and the sampling fraction is ϵ, the probability that the mutation is present in the tree is

$$P = 1 - e^{-w\epsilon} \tag{1}$$

W is a random number that depends on the realization of the transmission process, which is commonly modeled by a branching process with birth rate b and death rate d. The death rate here corresponds to clearing an infection, the birth rate to onward transmission. The latter is affected by the fitness cost of the mutation.

To obtain insight how the probability of observing a lineage depends on parameters, we calculate the probability $p(w, T|t)$ that a lineage had an integrated size $w = \int_t^T k(t')\, dt'$, where t is the birth time of the lineage, T is the current time, and $k(t')$ is the size of the lineage at time t'. To calculate $p(w, T|k)$, we generalize it slightly to $p(W, T|k, t)$, where k is the number of individuals at the start time t. This quantity obeys the following "first-step" equation:

$$-(\partial_t - k\partial_w)p(w, T|k, t) = -k(b+d)p(w, T|k, t) + kbp(w, T|k+1, t) + kdp(w, T|k-1, t) \tag{2}$$

We will solve for the Laplace transform $\hat{p}(z, T|k, t) = \int_0^\infty dw\, e^{-wz} p(w, T|k, t) = \hat{p}^k(z, T|1, t)$. Using the following identity for the derivative of the Laplace transform

$$\int_0^\infty e^{-wz}\partial_w p\, dw = [e^{-wz}p]_0^\infty - \int_0^\infty p\partial_w e^{-wz}\, dw = 0 + z\int_0^\infty pe^{-wz} = z\hat{p} \tag{3}$$

and setting $k = 1$, we have

$$-\partial_t \hat{p}(z, T|t) = -(b+d+z)\hat{p}(z, T|t) + b\hat{p}^2(z, T|t) + d \tag{4}$$

This simplifies further to if we substitute $\phi(z, T|t) = 1 - \hat{p}(z, T|t)$.

$$\begin{aligned} \partial_t \phi(z, T|t) &= -(b+d+z)(1-\phi(z, T|t)) + b(1-\phi(z, T|t))^2 + d \\ &= -z - (b-d-z)\phi(z, T|t) + b\phi(z, T|t)^2 \end{aligned} \tag{5}$$

where it is important to note that the derivative is with respect to the first time point and the interval $T - t$ is shrinking with increasing t.

Constant birth and death rate

If the fitness effect of the mutation in question is detrimental and the overall population is constant (background $b_0 = d_0$), all mutant lineages will eventually die out and we can consider large $T - t$ and the long time asymptotic $\partial_t \phi(z, T|t) = 0$. Further define $b = b_0 - s$ and $d = d_0$ where s is the fitness cost of the mutation (so larger values indicate a greater fitness cost). The steady state generating function is then

$$0 = -z - (b-d-z)\phi(z) + b\phi(z)^2 \tag{6}$$

with solution

$$\begin{aligned} \phi(z) &= -\frac{s+z}{2(b_0-s)} \pm \frac{\sqrt{(s+z)^2 + 4z(b_0-s)}}{2(b_0-s)} \\ &\approx -\frac{s+z}{2b_0} \pm \frac{\sqrt{(s+z)^2 + 4zb_0}}{2b_0} \\ &\approx \begin{cases} \frac{z}{s+z} & (s+z)^2 \gg 4zb_0 \\ \sqrt{\frac{z}{b_0}}\left(1 + \frac{(s+z)^2}{8zb_0}\right) - \frac{s+z}{2b_0} & (s+z)^2 \ll 4zb_0 \end{cases} \end{aligned} \tag{7}$$

Since $\phi(z) = 1 - \int e^{-wz}p(w)\, dw$, $\phi(\epsilon)$ is exactly the probability that a lineage is sampled when the entire population is sampled at rate ϵ. We thus expect two regimes: if the square of the fitness cost exceeds the sampling intensity (typically at 1% or less), the probability of sampling a lineage is essentially inversely proportional to the fitness cost. The sampling probability of lineages with smaller costs effects depends less strongly on s. Their sampling mostly comes down to stochasticity independent of the fitness cost.

Growing populations

In many scenarios relevant for lineages that arise during a viral outbreak, the background population isn't constant but is undergoing a rapid exponential expansion. The background birth rate b_0 is bigger than d_0 in this case. Since the population is growing, deleterious mutations can increase in frequency deterministically and we can not send the t to infinity as before. Instead, we need to integrate

$$\partial_t \phi(z,T|t) = -z - (b-d-z)\phi(z,T|t) + b\phi(z,T|t)^2 \tag{8}$$

backwards in time starting from $\phi(z,T|T) = 0$ at $t = T$. While $\phi(z,T|t)$ is small and the quadratic term can be neglected, this is approximately solved by

$$\begin{aligned}
\phi(z,T|t) &= ze^{\int_t^T (b-d-z)dt'} \int_t^T e^{-\int_\tau^T (b-d-z)dt'} d\tau \\
&= ze^{(b-d-z)(T-t)} \int_t^T e^{-(b-d-z)(T-\tau)} d\tau \\
&= ze^{(b-d-z)(T-t)} \left[1 - e^{-(b-d-z)(T-t)}\right] / (b-d-z) \\
&= \frac{z}{b-d-z}\left[e^{(b-d-z)(T-t)} - 1\right] = \frac{z}{\gamma_0 - s - z}\left[e^{\gamma_0(T-t)-(s+z)(T-t)} - 1\right]
\end{aligned} \tag{9}$$

where γ_0 is the growth rate of the background population.

At longer times when $ze^{\gamma_0(T-t)} \sim 1$ and ϕ is no longer small, ϕ tends towards a constant value determined by the same quadratic equation as above. This limit is neither interesting or relevant for the present purpose, since there are very few lineages that emerged early enough to have saturated ϕ. Instead, we need to average ϕ (the linear approximation) over all the time points when the lineage could have arisen.

$$\begin{aligned}
\langle\phi\rangle &\sim \int_t^T dt' \, e^{-\gamma_0(T-t')} \frac{z(e^{\gamma_0(T-t')-(s+z)(T-t')} - 1)}{(\gamma_0 - s - z)} \\
&= \int_t^T dt' \, \frac{z(e^{-(s+z)(T-t')} - e^{-\gamma_0(T-t')})}{(\gamma_0 - s - z)} \\
&\approx \begin{cases} \frac{z}{\gamma_0-s-z}[\frac{1}{z+s} - \frac{1}{\gamma_0}] & s(T-t) \gg 1 \\ \frac{z}{\gamma_0-s-z}[(T-t) - \frac{(s+z)(T-t)^2}{2} - \frac{1}{\gamma_0}] & s(T-t) < 1 \end{cases}
\end{aligned} \tag{10}$$

This derivation assumed that $\gamma_0(T-t) \gg 1$, i.e. that the overall population size has expanded substantially. The most relevant fitness effects will be those with $s(T-t) > 1$, that is the fitness effect has strong effect on variant frequency, but $s < \gamma_0$ such that the variant is still spreading and can give rise to large lineages in an expanding variant. In this case, the above simplifies to

$$\langle\phi\rangle \approx \frac{z}{\gamma_0(z+s)} \tag{11}$$

In a variant that has been growing with rate γ_0 for a time $\tau = T - t$ and sampled with $z = \epsilon$, we thus expect that the number of times we observe separate mutant lineages depends on s as

$$\langle\phi\rangle \approx \frac{\epsilon}{\gamma_0(\epsilon+s)} \tag{12}$$

This has a very similar behavior as the solution for constant population size, which suggests that the overall dependence on s is robust and we can assume that the number of times a mutation is observed is inversely proportional to its effect on fitness. The same basic dependency is observed at steady state in a quasi-species [55]. In a constant population, this relationship breaks down for dense sampling $\epsilon > \sqrt{s}$. In growing population, the approximation fails if the product of fitness effect and the time over which the variant has grown, $s\tau$, is small, i.e., if the fitness cost does not affect variant frequency strongly. In these cases, there is still a dependence on s, but it is weaker.

Supplementary figures

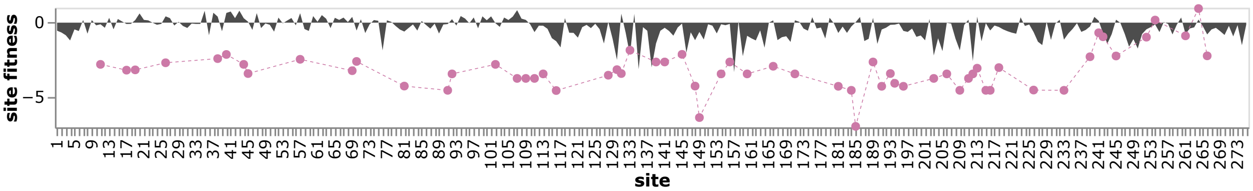


Figure S1 Effects of stop-codon and amino-acid mutations across ORF3a. The black area plot shows the mean effect of all amino-acid mutations at each site, and the purple points show the effects of stop codon mutations. There is strong selection against stop codons (negative effects) for all but the C-terminus of ORF3a, but only a few positions show strong selection against amino-acid substitutions. This plot shows only mutations with 20 expected counts. See https://jbloomlab.github.io/SARS2-mut-fitness/ORF3a.html for an interactive version of this plot along with zoomable heatmap of the effects of specific amino-acid substitutions.

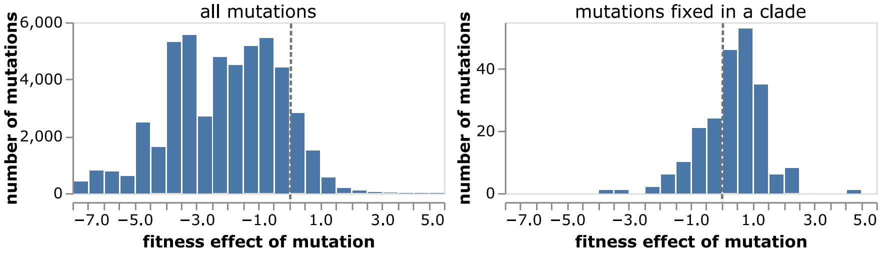


Figure S2 Distribution of fitness effects of all amino-acid mutations relative to Wuhan-Hu-1, and just those mutations that fixed in at least one clade of SARS-CoV-2 (using the Nextstrain clade definitions). The vertical dashed line at zero indicates the effect of a neutral mutation. See https://jbloomlab.github.io/SARS2-mut-fitness/clade_fixed_muts_hist.html for an interactive version of this plot that allows adjustment of the minimum expected count threshold.

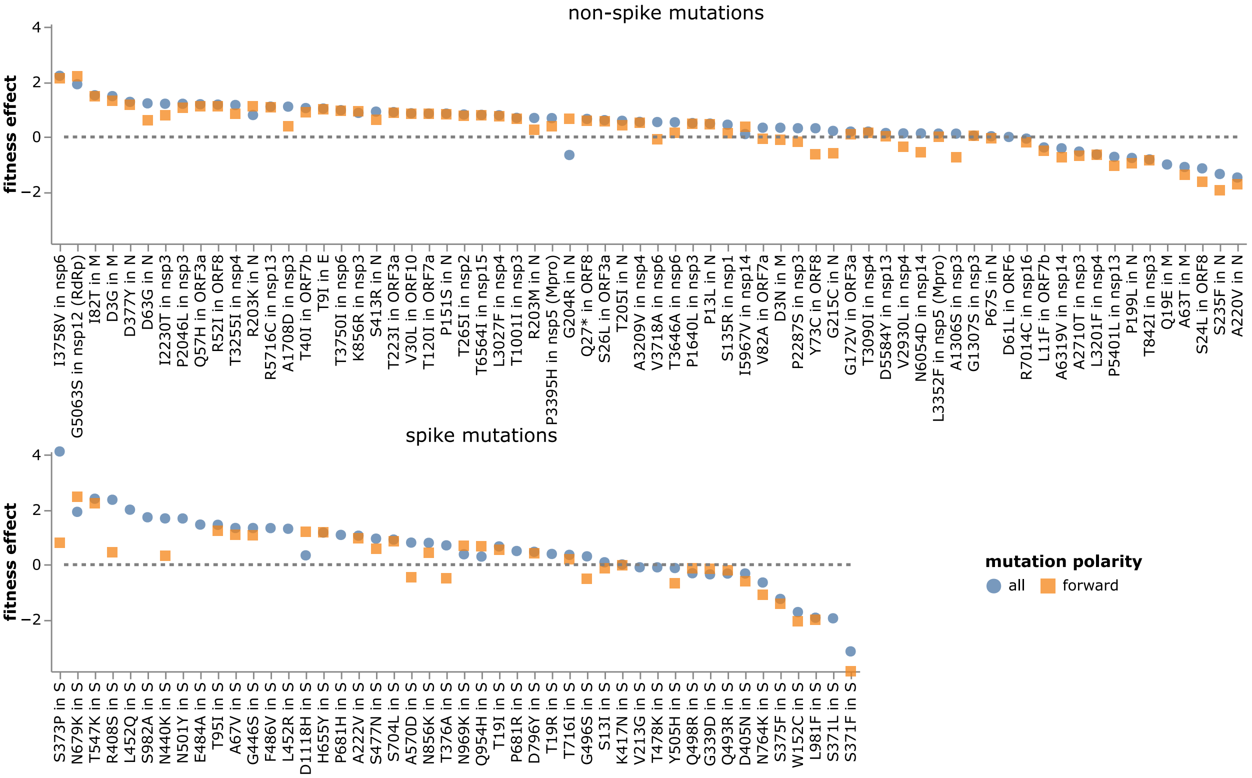


Figure S3 Effects of individual mutations that fixed in at least one clade of SARS-CoV-2, faceted by whether they are in spike or another protein. "Mutation polarity" indicates if the point shows the effect of the mutation estimated using all viral clades (including those that have fixed the mutation), or just from direct forward occurrences of the mutation in clades in which it has not yet fixed. Some mutations are estimated to be more favorable when including clades in which they have fixed (blue circles) in addition to just clades in which it has not yet fixed (orange squares)—when this occurs, it suggests epistatic entrenchment of the mutations [38, 50]. Note that clades in which a mutation has already fixed contribute to estimates of its fitness via estimates of the effect of its reversion and via estimates of the effects of mutations to other amino acids at the same site. See https://jbloomlab.github.io/SARS2-mut-fitness/clade_fixed_muts.html for an interactive version of this plot.

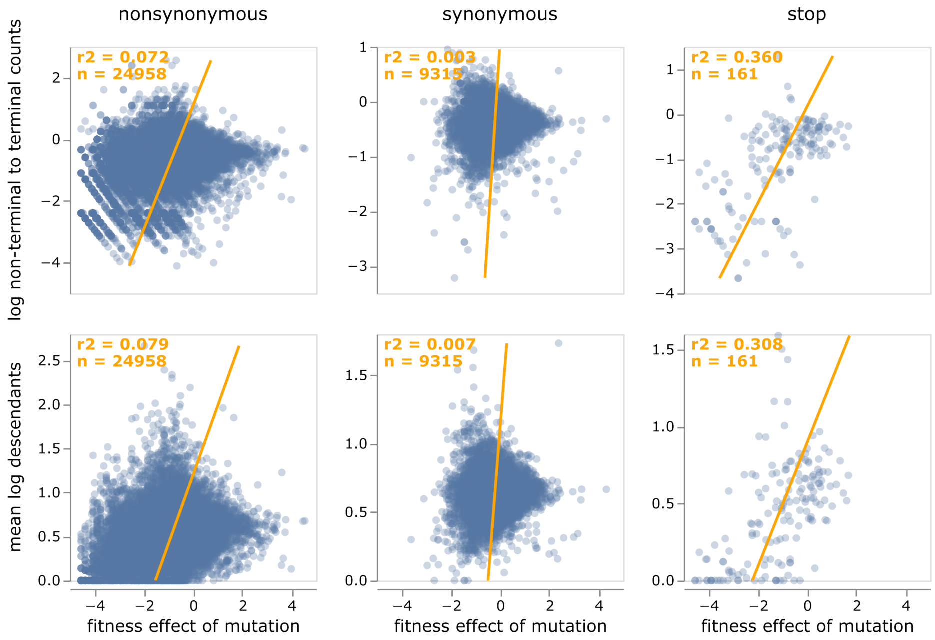


Figure S4 Relationship between fitness effects of mutations and two measures of the number of descendants. At top is shown the log ratio of counts of the mutation on non-terminal (internal) to terminal (tip) branches; larger values indicate mutations more likely to be found in viruses that leave descendants. At bottom is shown the mean log number of tip descendants that share all the mutations on each branch containing the mutation of interest; larger values again indicate mutations more likely to be found in viruses that leave more descendants. Each point is an amino-acid mutation, the orange line is a least-squares regression, and the orange text in the upper left give the number of mutations and the Pearson correlation coefficient. This plot shows only mutations with at least 10 expected counts and 5 actual counts. See https://jbloomlab.github.io/SARS2-mut-fitness/fitness_vs_terminal.html for an interactive version of this plot that allows filtering by the number of actual or expected counts, or by gene. The number of descendants is calculated using the "leaves_sharing_mutations" variable of the UShER mutation-annotated tree.