PLOS PATHOGENS

REVIEW

# Developing an appropriate evolutionary baseline model for the study of SARS-CoV-2 patient samples

**John W. Terbot, II[1,2], Parul Johri[2], Schuyler W. Liphardt[1], Vivak Soni[2], Susanne P. Pfeifer[2], Brandon S. Cooper[1]*, Jeffrey M. Good[1]*, Jeffrey D. Jensen[2]***

**1** University of Montana, Division of Biological Sciences, Missoula, Montana, United States of America, **2** Arizona State University, School of Life Sciences, Center for Evolution & Medicine, Tempe, Arizona, United States of America

* brandon.cooper@mso.umt.edu (BSC); jeffrey.good@mso.umt.edu (JMG); jeffrey.d.jensen@asu.edu (JDJ)





**Citation:** Terbot JW, II, Johri P, Liphardt SW, Soni V, Pfeifer SP, Cooper BS, et al. (2023) Developing an appropriate evolutionary baseline model for the study of SARS-CoV-2 patient samples. PLoS Pathog 19(4): e1011265. https://doi.org/10.1371/journal.ppat.1011265

**Editor:** Tom C. Hobman, University of Alberta, CANADA

**Published:** April 5, 2023



**Funding:** Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health (NIH) under Award Numbers P20GM102546 and P30GM140963. NIH awards R35GM124701 (BSC) and R35GM139383 (JDJ) also supported this work. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing interests:** The authors have declared that no competing interests exist.

## Abstract

Over the past 3 years, Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) has spread through human populations in several waves, resulting in a global health crisis. In response, genomic surveillance efforts have proliferated in the hopes of tracking and anticipating the evolution of this virus, resulting in millions of patient isolates now being available in public databases. Yet, while there is a tremendous focus on identifying newly emerging adaptive viral variants, this quantification is far from trivial. Specifically, multiple co-occurring and interacting evolutionary processes are constantly in operation and must be jointly considered and modeled in order to perform accurate inference. We here outline critical individual components of such an evolutionary baseline model—mutation rates, recombination rates, the distribution of fitness effects, infection dynamics, and compartmentalization—and describe the current state of knowledge pertaining to the related parameters of each in SARS-CoV-2. We close with a series of recommendations for future clinical sampling, model construction, and statistical analysis.

## Introduction

Current evidence suggests that Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) emerged from Hubei Province, China, in late 2019 [1]. The virus has since spread through global human populations in several waves, infecting an estimated 612 million individuals and causing Coronavirus Disease 2019 (COVID-19), resulting in 6.55 million recorded deaths as of October 2022—though total estimates suggest 18.2 million fatalities in the first 2 years alone (January 2020 to December 2021; [2]). The situation has been exacerbated by the continued emergence of new variants of concern (VOCs) that continue to disrupt basic human health and activity (Fig 1), even after the development of vaccines that have significantly lessened COVID-19 severity. In particular, the Delta variant first identified in late 2020 in India was found to be highly transmissible; more recently, the emergence of Omicron —thought to have arisen from a long-term infection of an immunosuppressed individual [3]—led to even more rapid spread.

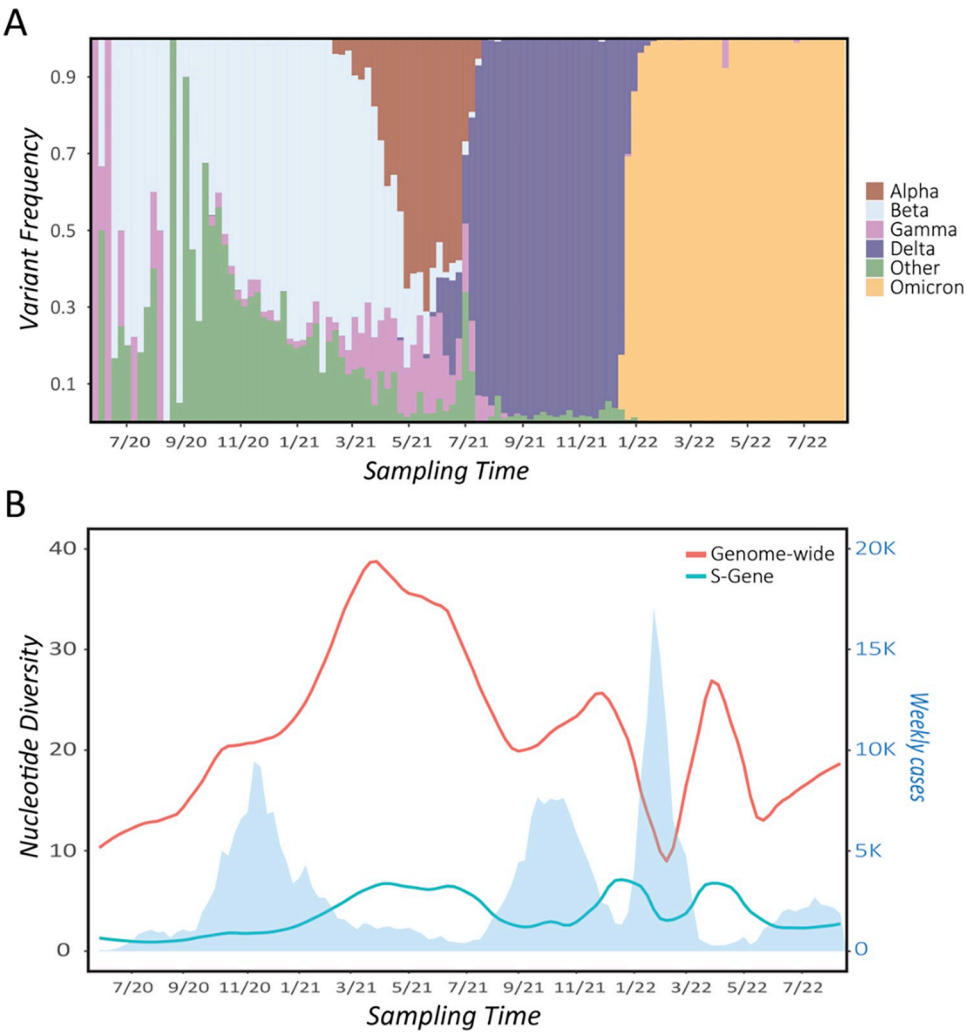


**Fig 1. SARS-CoV-2 variant frequencies and genetic diversity through time, given for the state of Montana as illustration. (A)** Frequency of major WHO-defined variants of concern (VOCs), binned by week of sampling as derived from GISAID metadata. **(B)** Average pairwise nucleotide differences between consensus sequence genomes isolated from patient samples genome-wide (orange line) and within the S-gene encoding the spike protein (blue line). As shown, local spread of major VOCs induced corresponding drops in overall genetic diversity across consensus sequences and within the S-gene. Note that while Beta dominated early in Montana, multiple variants cocirculated at appreciable frequency in 2020 to 2021, and the dominant observed strain differed by location. Consensus sequences were downloaded from GISAID ($n$ = 21,799), binned by week, and aligned to the SARS-CoV-2 reference sequence using Nextalign. Diversity was calculated in R, using the package PopGenome. Local polynomial regression fitting (method loess) was used in the R package ggplot2 to model diversity through time with the formula y ~ x. Case data (given by the light blue shading) were downloaded from the CDC.

https://doi.org/10.1371/journal.ppat.1011265.g001

In response, global surveillance efforts have emerged to anticipate and track the ongoing evolution of this deadly virus using whole genome high-throughput sequencing. In fact, SARS-CoV-2 has rapidly become one of the most heavily resequenced genomes ever, with more than 13 million patient isolates to date. At the same time, several components of SARS-CoV-2 population biology and molecular evolution remain unresolved, potentially limiting both the accurate detection of emerging VOCs and the design of novel therapeutics (e.g., [4]). Here, we discuss the co-occurring evolutionary processes that must be jointly modeled to accurately study SARS-CoV-2 evolution. This includes outlining relevant SARS-CoV-2

mutation and recombination rate estimates, which govern the input of new genetic variation and the potential generation of novel combinations of variation. We additionally assess current information pertaining to the distribution of fitness effects, infection dynamics, and patterns of compartmentalization that, together with mutation and recombination, comprise a set of critical components of an evolutionary baseline model.

### A brief overview of key considerations

Fundamentally, when studying genome evolution, one should first begin with the realization that populations are simultaneously shaped by multiple evolutionary processes, including mutation, recombination, natural selection, and genetic drift. For this reason, it is often not feasible to accurately study any individual process in isolation without considering the input of all factors in shaping observed levels and patterns of genomic variation. For example, searching for genomic loci that have recently experienced positive selection—potentially leading to organismal adaptation—is particularly fraught as positive selection is expected to be rare and episodic relative to other constantly acting evolutionary processes (see reviews of [5,6]). Thus, the ability to successfully identify positively selected loci and differentiate this process from others is highly dependent on multiple underlying parameters. For example, it has previously been demonstrated that population bottlenecks, as well as high neutral mutational inputs, may be mistaken for the action of positive selection (e.g., [7–10]). For these reasons, the construction of an appropriate baseline model consisting of constantly acting evolutionary processes (e.g., mutation, genetic drift as modulated by population history, purifying selection, background selection, and so on) is critical to the successful detection and quantification of positive selection [11]. Importantly, as many of these underlying processes will be both heterogenous and estimated with uncertainty across the genome in question (e.g., mutation rates vary across a genome, and there is additionally measurement uncertainty in the underlying rate at any given site), the range of feasible parameter values needs to be modeled and considered within the context of this baseline.

All of these considerations also apply when the organism in question is a virus and the population concerned is contained within a patient. Human pathogens are often characterized by extreme mutational inputs, drastic population size changes relating to infection, reinfection, immune response, and clinical therapeutics, as well as often severe selective constraints owing to their coding-dense genomes [12]. Prior efforts have made significant progress in developing evolutionary baseline models for other common viruses (reviewed in [13]), and we here consider such model construction for the study of within-host SARS-CoV-2 populations. We outline critical individual components of a baseline model—mutation rates, recombination rates, the distribution of fitness effects, infection dynamics, and compartmentalization—and describe the current state of knowledge pertaining to the related parameters of each. Finally, we close with a series of recommendations for future clinical sampling, baseline model construction, and statistical analysis.

## Mutation rates

As the evolutionary source of new genetic variation, the first component of this baseline model necessarily involves the rate of input of new mutations. Numerous methods exist for estimating these rates, though the first complication that arises in comparing among studies is that some follow a convention of reporting rates per year or per day, while others per viral cycle. To compare among such estimates for SARS-CoV-2, we consider the maximum number of sequential viral cycles thought to occur in a single year for use as a conversion factor. Specifically, based on cycle times in SARS-CoV-1, the time required to enter a host cell is

**Table 1. Estimated mutation rates in SARS-CoV-2, and the related CoV-1 and MHV.**

| Virus | Source | Date | Original Unit | Estimated Rate / Cycle | SEM | Citation |
|---|---|---|---|---|---|---|
| MHV | in vitro | June 2004 | mut/nt/cycle | $3.5 \times 10^{-6}$ mut/nt/cycle | n/a | [15,16] |
| SARS-CoV-1 | divergence | June 2004 | mut/nt/yr* | $0.93–2.76 \times 10^{-6}$ mut/nt/cycle | n/a | [17] |
| SARS-CoV-2 | divergence | February 2021 | mut/nt/day* | $0.79 \times 10^{-6}$ mut/nt/cycle | n/a | [20] |
| SARS-CoV-2 | divergence | February 2020 | mut/nt/yr* | $1.38 \times 10^{-6}$ mut/nt/cycle | 0.35** | [21] |
| SARS-CoV-2 | divergence | February 2020 | mut/nt/yr* | $1.52 \times 10^{-6}$ mut/nt/cycle | 0.36** | [21] |
| SARS-CoV-2 | divergence | June 2020 | mut/nt/yr* | $2.79 \times 10^{-6}$ mut/nt/cycle | 0.53** | [22] |
| SARS-CoV-2 | divergence | February 2021 | mut/nt/yr* | $1.71 \times 10^{-6}$ mut/nt/cycle | 0.23** | [23] |
| SARS-CoV-2 | divergence | August 2022 | mut/nt/yr* | $0.72 \times 10^{-6}$ mut/nt/cycle | n/a | [26] |
| SARS-CoV-2 | in vitro | March 2022 | mut/nt/cycle | $1.3 \times 10^{-6}$ mut/nt/cycle | 0.1 | [24] |

*Evolutionary rates reported as per year or per day mutation rates were converted to per cycle mutation rate using the approximation of 861.64 viral cycles per year.
**SEMs (standard errors of the mean) estimated based on reported 95% HPD (highest posterior density) intervals.

https://doi.org/10.1371/journal.ppat.1011265.t001

approximately 10 minutes and the eclipse period of the virus (i.e., the time between host cell entry and the generation of new infectious particles) is approximately 10 hours [14], leading to a conversion factor of 861.64 viral cycles/year (i.e., the number of 610-minute cycles in a calendar year). While such a conversion is necessary in order to compare between existing estimates, variation in these entry and eclipse times would naturally result in a modified scaling factor.

Molecular clock-based substitution rates previously observed in other betacoronaviruses provide the first insight into relevant mutation rates (Table 1); however, such clock-based rates ought to be regarded as more closely approximating the rate of neutral mutations, rather than the rate of all mutations. Using such a phylogenetic approach, this rate has been estimated at $3.5 \times 10^{-6}$ mutations/nucleotide (nt)/cycle for murine hepatitis virus (MHV) [15,16] and at 0.80 to $2.87 \times 10^{-3}$ mutations/nt/year for SARS-CoV-1 [17], corresponding to 0.93 to $3.33 \times 10^{-6}$ mutations/nt/cycle following our above conversion. Given the abundance of sequence data in public databases (e.g., [18,19]), comparable rate estimates have also recently been reported for SARS-CoV-2. For example, using SARS-CoV-2 sequences generated prior to April 2020 and estimating genetic distance to closely related bat and pangolin coronaviruses, an initial clock-based estimate was inferred to be $1.87 \times 10^{-6}$ mutations/nt/day (approximately $7.92 \times 10^{-7}$ mutations/nt/cycle) [20]. Other estimates have focused on rates of divergence over time within humans. An early analysis based on 73 genomes collected between December 2019 and February 2020 estimated a mutation rate of 1.19 to $1.31 \times 10^{-3}$ mutations/nt/year [21], or 1.38 to $1.52 \times 10^{-6}$ mutations/nt/cycle. A subsequent study based on 137 genomes estimated a rate of $2.4 \times 10^{-3}$ mutations/nt/year (95% CI: 1.5 to $3.3 \times 10^{-3}$ mutations/nt/year; [22]), or $2.79 \times 10^{-6}$ mutations/nt/cycle (95% CI: 1.74 to $3.83 \times 10^{-6}$ mutations/nt/cycle)—a rate similar to other estimates from the same period using similar methods [23].

An alternative approach for estimating mutation rates instead relies on serial sampling within an experimental setting or from individual patients. This serial sampling approach has the advantage of not suffering the data reduction associated with comparisons between consensus sequences (see Data consideration section below) and should capture a wider (but still limited) range of newly emerging mutations. For example, sampling over 15 days from lines of African green monkey (*Chlorocebus aethiops*) kidney cell cultures inoculated with a clinical isolate led to an observed rate between 2.9 and $3.7 \times 10^{-6}$ mutations/nt/cycle across the genome [24]. However, after excluding the nsp3, nsp6, and S genes—which showed evidence of selection potentially biasing the mutation accumulation-based inference—estimates were

reduced to $1.3 \pm 0.2 \times 10^{-6}$ mutations/nt/cycle (and see [24] for a more mechanistic discussion of these contributing variants).

In addition to the overall mutation rate, consideration needs to be given to the extent to which mutation rates may vary across the SARS-CoV-2 genome. The causes of any apparent mutation rate heterogeneity also require attention given that differential selection across genomic regions may be responsible for these patterns rather than a variable underlying rate itself. For example, based on comparisons between 4-fold degenerate sites relative to other sites in the genome, it was estimated that the genuine mutation rate may be 49% to 67% higher than that estimated based on common segregating variants [25]. Additionally, a continuous reduction in the nonsynonymous (but not synonymous) mutation rate has been observed throughout the pandemic [26]—a trend likely indicating changes in selection pressures rather than mutation rates. For this reason, a consideration of the proportion of new mutations that are deleterious, neutral and beneficial (see Distributions of fitness effects section below) will be important for understanding the total rate of new mutation and distinguishing it from rates estimated from segregating variation.

In this regard, it is important to also consider the genomic composition of SARS-CoV-2, which is approximately 29.8 kilobases (kb) and contains 25 to 30 distinct regions that encode gene products. As expected, genic regions generally appear highly constrained, while a subset of genes, including the spike gene (S), appear to evolve comparatively rapidly. Mechanistically, RNA-dependent RNA polymerases are known to be error-prone [4,27,28], which in coronaviruses is partly compensated by a 3′ to 5′ proofreading exonuclease, nsp14 [29]. The impact of this sort of "evolutionary layering" remains in need of further study [4,30]. Notably, mutations within the replicative machinery (nsp12 and nsp14) can directly lead to increases in mutation rates. For example, using inoculated kidney cells from the African green monkey, mutator lines have been observed to develop which accumulated mutations at roughly an order of magnitude greater rate [24]. These mutator lines were found to have several nonsynonymous mutations in their replication machinery unique to those lineages (8 in nsp12, and 9 within nsp14). Such observations have in fact led to the suggestion that drug-induced mutational meltdown may itself represent a viable patient-treatment strategy (e.g., [4,31]), as has previously been suggested in other viruses (e.g., [32]).

## Recombination rates

Though mutation is the source of new genetic variation, recombination is an important process for generating novel combinations of variation. Additionally, by disrupting selective interference between and among mutations (i.e., Hill–Robertson interference; [33,34]), recombination may improve the efficacy of both positive and purifying selection.

Recombination has been observed in many viruses, including coronaviruses [4,35–38]. Indeed, recombination events between different coronavirus lineages are thought to have been essential for the evolution of SARS-CoV-2. Specifically, several critical recombinant regions were identified in the genome; 3 within the spike protein and 1 associated with each of the RNA-dependent RNA polymerase [39,40]. An early report using a sequence similarity search of a local database of SARS-CoV-2 samples and other coronaviruses concluded that there were sequences derived from coronaviruses of bats and, potentially, pangolins [20]. Phylogenetic approaches using sliding window analyses have found a similar mosaic origin for SARS-CoV-2 [41–44]. Moreover, recombination between SARS-CoV-2 lineages was detected as early as April 2020, using analyses based on linkage disequilibrium [20]. A subsequent tree-based analysis constructed from 1.6 million sequences identified 606 putative recombination events, suggesting that 2.7% of circulating strains likely had a recombinant origin [45]. Importantly, the accurate computational estimation of these rates requires the analysis of within-host

polymorphism data, rather than consensus sequence data as is common, suggesting that this frequency may be underestimated [46].

A major limiting factor for recombination detection is that it requires a coinfection by 2 or more strains [47,48]. Moreover, this coinfection needs to not only occur within a single host, but the viral strains also need to infect the same host cell. Coinfections may be relatively rare—estimated to occur in 0.18% to 0.61% of samples [48–50]—though this rate would be expected to increase alongside increases in viral occurrence [47]. Additionally, within-host dynamics will impact the duration of the infection—the longer an infection persists, the greater the opportunity for a secondary infection to be acquired [48].

The analysis of recombination breakpoints has identified possible hotspots within the SARS-CoV-2 genome [51], suggesting that, as with mutation, rate heterogeneity across the genome may be important. These hotspots are often associated with transcription regulatory sequences (TRSs) found nearby various open reading frames (ORFs) and are associated with the template switching process, which produces sgRNA. Notably, recombination junctions that were associated with TRSs were less likely to produce defective viral genomes [52]. Microhomologies of 2 to 7 nucleotides between recombination junctions of SARS-CoV-2 and MHV were also identified, potentially suggesting some level of conservation [52]. Finally, the recombination rate may also be related to the proofreading enzyme found in coronaviruses. When this gene was knocked out in lines of MHV, there was a significant reduction in recombination rates in addition to altered recombination patterns [52].

The estimated rates of recombination for coronaviruses, including SARS-CoV-2, occur within a considerable range. Work in MHV estimated the recombination rate of that virus to be roughly $1 \times 10^{-5}$ between consecutive sites [53]. A recent study in SARS-CoV-2, seeking to detect recombination using a parsimony approach [54], had a sensitivity that allowed recombination detection given at least a minimal rate of $1 \times 10^{-6}$ recombination events/site/cycle [55]. Recombination was successfully detected by this method, providing a possible minimum for the recombination rate. An alternate approach using a Markov chain Monte Carlo method to infer recombination networks under a template-switching model of recombination [56] inferred a rate of $2 \times 10^{-6}$ events/site/year. Using the same conversion factor used previously for mutation rates, this would suggest an approximate recombination rate of $2.32 \times 10^{-9}$ events/site/cycle—substantially below the detection floor of the parsimony-based method referenced above, likely owing to considerations related to coinfection rates. Thus, these estimates, combined with the uncertain rates of coinfection, suggest considerable ambiguity in this parameter space. However, current evidence would imply recombination rates in the range of $1 \times 10^{-5}$ to $1 \times 10^{-6}$ events/site/cycle in coinfections (i.e., pertaining to $<1\%$ of total infections).

## Distributions of fitness effects

The SARS-CoV-2 genome, albeit sizeable for an RNA virus, is nonetheless relatively small and highly compact, with $>95\%$ of the genome thought to be functionally significant [57,58]. Given this paucity of nonfunctional sequence, and that many ORFs are overlapping, it will likely be challenging to identify neutrally evolving sites in the SARS-CoV-2 genome in sufficient numbers to allow for common neutrality-based inference approaches. That said, synonymous sites are possible candidates, with a recent study [26] suggesting that the rate of synonymous substitution appears to be roughly constant over time, indicating that synonymous sites in SARS-CoV-2 may be evolving nearly neutrally.

However, while the frequency distribution of neutral alleles is shaped by demographic events (e.g., the population bottleneck and subsequent growth associated with infecting a host)—allowing us to make inferences about population sizes [59]—it is also impacted by the linked

effects of selection acting on nearby functionally important sites ([60,61]; and see review of [6]). For this reason, it has been shown to be necessary to jointly infer the distribution of fitness effects (DFE) together with population history when analyzing coding-dense genomes [62–64]. In other words, while neutral mutations surely exist, fully degenerate sites that are unlinked to constrained sites may not.

There has been some work to date to characterize the DFE of observed mutations sampled from SARS-CoV-2 patients, using a variety of approaches. Accounting for the background rate of growth/transmission for each geographic region in the United States separately, Kepler and colleagues [65] used a maximum likelihood phylodynamic method to infer the fitness effects of segregating amino acid variants from 88,000 viral genomes. Under this framework, the fitness of each lineage on the tree was estimated using a birth–death process (such that fitness of the parent and child branch was correlated). As expected, most segregating variants were found to be effectively neutral, with a small minority being mildly deleterious and approximately 20% being putatively beneficial (Fig 2). More specifically, prior to 2020, approximately 14% (7 out of 51) of all amino acid variants segregating were inferred to be significantly beneficial with an average fitness advantage of 1.15 relative to the wild type. Strongly beneficial mutations are expected to fix rapidly in populations, and strongly deleterious mutations are likewise expected to be purged quickly (conditional on fixation and loss, respectively); therefore, most segregating variants sampled at any given time are likely to be neutral or weakly selected, consistent with this inferred DFE of segregating variants (see Fig 2).

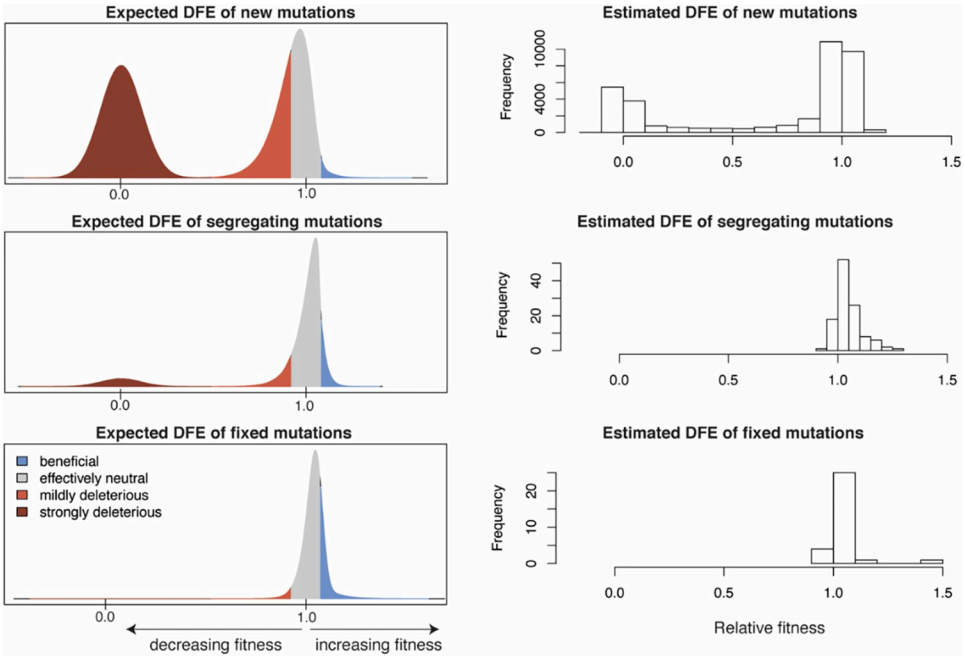


**Fig 2. The expected and estimated distributions of fitness effects (DFE) of mutations.** Left panels: a hypothetical expected DFE of new, segregating, and fixed mutations, reflecting the effects of selection at each stage. Effectively neutral mutations are shown in gray, beneficial mutations in blue, and deleterious mutation in shades of red. Right panels: the DFE of new, segregating, and fixed amino acid variants in SARS-CoV-2 as recently estimated by Flynn and colleagues [131], Kepler and colleagues [65], and Obermeyer and colleagues [132], respectively. From Flynn and colleagues, the normalized functional scores from 2 sets of biological replicates were pooled together. From Kepler and colleagues, the relative fitness of all single mutations from pre- and post-2020 studies were pooled together. From Obermeyer and colleagues, the DFE of fixed mutations was approximated by using 31 high-frequency variants (defined as those present in more than 100 lineages). Importantly, observed genomic variation will depend heavily on the underlying heterogeneity in both mutation rates and DFEs across the genome, among other factors [133].

https://doi.org/10.1371/journal.ppat.1011265.g002

More generally, multiple studies (e.g., [66–68]) have found significantly fewer segregating alleles at nonsynonymous relative to synonymous sites, likely reflecting the effects of purifying selection acting on functional changes. For example, Neher [26] evaluated patterns of segregating alleles and found that 15% to 20% of first and second codon positions exhibit no observed variation. This is consistent with previous random mutagenesis studies in other coding-dense RNA viruses, which have suggested that 20% to 40% of all new mutations are likely strongly deleterious [69,70]. Interestingly, some ORFs also appear to be experiencing mild selective constraints, as evidenced by comparison with synonymous sites [26], further suggesting that an appreciable class of mutations may also be weakly deleterious.

Thus, while considerable study is required to further characterize within-patient DFEs—particularly accounting for the simultaneous contributions of population size change, direct selection, and selection at linked sites [71]—current studies have provided first glimpses into the relative occupancy of different DFE classes. Furthermore, phylogenetic methods utilized to date neglect the effects of recombination, further highlighting the value of future applications of population genetic-based inference approaches. With regard to baseline model construction specifically, the uncertainty in the underlying DFE may be evaluated by modeling the effects of different possible densities in the 2 modes of the DFE, utilizing a range of generalized densities (e.g., [72]).

## Infection dynamics

Within-host population dynamics are important determinants of levels and patterns of variation, as well as of potential selective outcomes (Fig 3). Generally, a population bottleneck tends to be associated with initial patient infection, followed by rapid population growth (see review of [13]). The size of the transmission bottleneck is an important factor in determining how much within-host genetic variation is initially present, on which selection may act [73]. A narrow transmission bottleneck can result in a severe loss of genetic variation (known as a founder effect), with low-frequency within-host variants being stochastically lost from the population, largely regardless of their fitness effects. Conversely, if the transmission bottleneck is wide, then there may be numerous viral particles founding the initial infection, increasing the chance that beneficial variants are maintained.

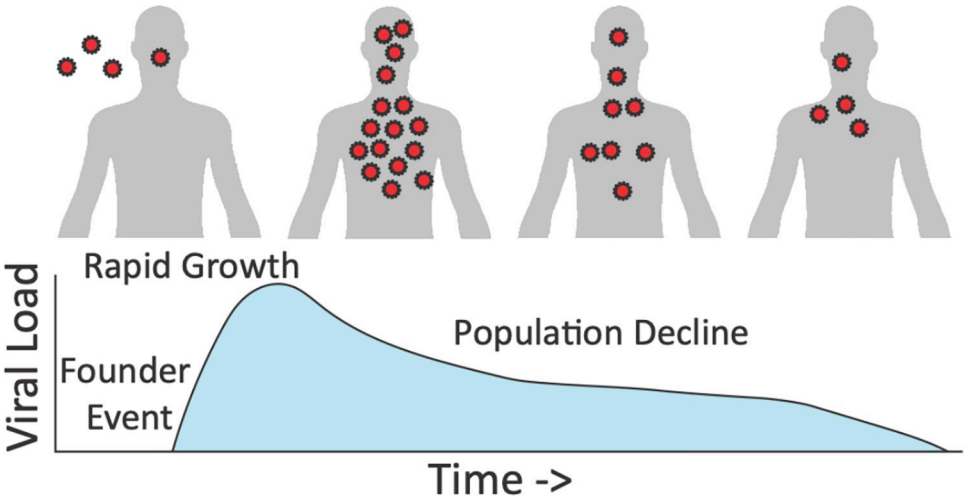


**Fig 3. A schematic of a simple intra-host demographic model potentially underlying infection dynamics.** At the time of infection, the virus population will initially be characterized by a population bottleneck associated with the founder event. A successful infection will next be characterized by rapid population growth associated with high viral loads, and reducing sizes and loads as the patient begins to clear the infection. The details of this infection history will greatly shape the observed levels and patterns of intra-host diversity. Viral load schematic modified from [134].

https://doi.org/10.1371/journal.ppat.1011265.g003

Several studies [66,74–80] have attempted to infer the severity of the SARS-CoV-2 infection bottleneck using the beta-binomial method [81], though estimates have ranged considerably, from 1 to 8 infecting particles [66] to more than 1,000 [74]. It has also been shown that the frequency threshold for detecting variants impacts bottleneck estimates [77–79], with an increased frequency threshold resulting in decreased bottleneck size estimates. There are fewer studies that have considered the increase in population size following the transmission bottleneck. Du and colleagues [81] performed a systematic review of viral dynamic parameters used in within-host models, estimating a mean viral load at symptom onset of 4.78 (95% CI: 2.93, 6.62) log(copies/ml) across 3 models, compared with a mean viral load at point of infection of −1.00 (95% CI: −0.94, −0.05) log(copies/ml), which puts a wide range on the respective rate of postinfection expansion. Wang and colleagues [82] compared genetic diversity in intra-host populations from the respiratory and gastrointestinal tract, finding considerably less diversity in the former than in the latter. One possible intra-host migration route is from the respiratory tract to the gastrointestinal epithelia, suggesting that this reduction in diversity in the respiratory tract may be due to the infection bottleneck, followed by rapid recovery in the gastrointestinal tract (though see the below section on Compartmentalization as well).

The considerable variance in these estimates suggest that there is still important work to be done in inferring a demographic infection model and its underlying parameters for SARS-CoV-2. Such inference is inherently challenging due to the impact of selection on biasing demographic inference as discussed above (and see [83,84]), and indeed due to the impact of demography on biasing selection inference [62,85]. To account for this circular problem, methods that jointly infer demography and the DFE will be critical. Although numerous joint estimation approaches for demographic inference have been developed, the most appropriate approach will be dependent on the context in which it is applied (considerations of which are reviewed in [11]). Neutral demographic estimators require sufficiently large nonfunctional regions and high rates of recombination, such that assumptions of strict neutrality hold [86–89]. Specifically, these criteria ensure that variants can be chosen that are not experiencing background selection. For example, Renzette and colleagues [90] utilized a neutral demographic inference approach (*dadi*; [86]) to build and parameterize infection models in human cytomegalovirus (HCMV) (and see [91,92]). It is notable, however, that the HCMV genome is 236 kb in size—among the largest human viral genomes [93]—and has large noncoding regions [94]. By contrast, the smaller, largely functional SARS-CoV-2 genome [95,96] likely prohibits such neutral inference. For these reasons, recently proposed approximate Bayesian computation (ABC) approaches to estimate demography while accounting for background selection effects will likely be the most fruitful path forward [62].

Of additional importance in considering viral infection dynamics is the notion of progeny skew—i.e., the viral replication dynamics. A majority of population genetic inference approaches assume small progeny distributions—an assumption that is likely violated in many pathogens (see reviews of [12,97]). Helpfully, recent inference approaches have relaxed this assumption, demonstrating an ability to coestimate parameters related to the biology of progeny skew together with those of demographic and selective histories (e.g., [98,99]). Moreover, progeny skew has been incorporated into the joint ABC inference scheme noted above, demonstrating an ability to tailor such inference specifically to viral populations [64,100], and, importantly, to avoid the misinference resulting from a neglect of this consideration.

## Compartmentalization

Related to the population dynamics discussed above, which were largely concerned with population size changes associated with infection, the compartmentalization of viral populations

within an infected individual (i.e., within-host population structure)—either across tissue types or regions of a single tissue—can also play a key role in the intra- and inter-host evolutionary dynamics of a virus. For example, HIV is known to spread throughout the body during early stages of infection leading to distinct viral populations localized to certain organs or systems, resulting in populations evolving independently under unique evolutionary pressures [101]. Influenza A virus has been shown to compartmentalize within different lobes of the lungs resulting in genetically distinct populations, each with distinct evolutionary histories. Similarly, HCMV has been shown to have strong compartmentalization effects, with the plasma population facilitating a certain degree of gene flow between compartments [90,102–105].

It has been apparent since early in the pandemic that SARS-CoV-2 also demonstrates a certain level of intra-host compartmentalization, with the identification of unique variants not shared between the upper and lower respiratory tracts in patients presenting severe disease [106,107]. This is perhaps not surprising, given that previous studies have observed compartmentalization between upper and lower respiratory tracts in both SARS-CoV-1 and MERS [108,109]. Recent sequencing efforts across a larger number of samples, and to a greater coverage across the viral genome, have recapitulated this pattern of differentiation between the upper and lower respiratory tracts in SARS-CoV-2 [110]. Intra-host single nucleotide variants not shared between blood and upper respiratory tract samples were also recovered from a single immunocompromised patient in France [111]. SARS-CoV-2 compartmentalization across different organs has been studied to a lesser degree; however, organ-specific variants, including the observation of VOC mutations outside of lung tissue, have been reported [112]. For example, compartmentalization has been observed between the respiratory tract and the gastrointestinal tract, with unique variants recovered from different samples [82].

Early evidence suggests that the evolutionary dynamics within compartments may be at least initially shaped by stochastic processes for SARS-CoV-2—such as founder events resulting in genetically distinct compartmental populations—similar to patterns seen in other viruses [110,113]. Further, several recent studies have demonstrated that the ability of the Omicron variant to replicate in the lungs is severely reduced compared to that of the nasal tract, potentially suggesting compartmentalization effects in this regard as well [114,115]. As such, this within-host structuring owing to compartmentalization represents another important component of any underlying evolutionary baseline model.

## Data considerations

Apart from the considerations of contributing evolutionary processes discussed in the sections above, it is additionally important to consider the types of data needed and available to perform such evolutionary inference. As of October 2022, there were over 6.3 million SARS-CoV-2 genomes deposited in NCBI, and over 13.4 million genomes deposited in GISAID [116]. Consensus sequences (i.e., single-sequence representations of a patient's viral population) produced from individual samples comprise most of these data, and these sequences serve as the primary unit for most genetic analyses of SARS-CoV-2 lineage variation and evolution. While consensus sequences provide the opportunity to carry out analyses concerning inter-host viral variation across millions of infected patients, they completely obscure intra-host variants segregating within patients (Fig 4). This preventable loss of information precludes analysis of variation within hosts and can lead to incorrect evolutionary inference regarding the relative contributions of respective evolutionary processes during viral evolution [13,104]. More fundamentally, consensus-level variation is simply a summarized by-product of the multiple contributing evolutionary processes acting within hosts. At a minimum, we emphasize that

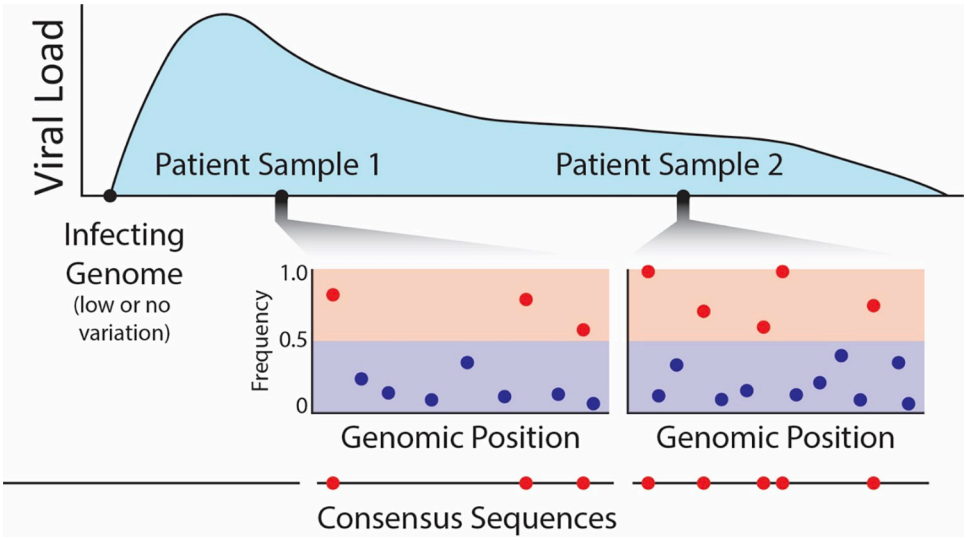


**Fig 4. An example of clinical sampling of a patient over the course of an infection, demonstrating how a consensus sequence-based summary neglects the great majority of intra-host variants (which are expected to primarily segregate at low frequencies [shown as blue circles]).** This ascertainment of high-frequency intra-host variants (shown as red circles) for subsequent inter-host comparison thus represents an unfortunate and unnecessary loss of information.

https://doi.org/10.1371/journal.ppat.1011265.g004

depositing the raw reads from which consensus sequences were constructed would be a helpful step in allowing subsequent researchers to examine individual host-level variation. Indeed, as is already standard in other scientific fields, depositing raw sequencing data to public repositories should be adopted as a best practice for viral genome surveillance in order to assure scientific transparency and reproducibility.

Relatedly, how and where SARS-CoV-2 genomes are deposited can create barriers to carrying out analyses of intra-host variation. For example, researchers may submit to NCBI and / or GISAID, and submissions may or may not include both raw reads and consensus sequences. This also makes it challenging to reliably cross-reference samples and underlying data between databases. Still, as with the yellow fever virus [117], influenza [118], norovirus [119], Ebola [120], and other RNA viruses [121,122], analyses of available raw-read data have revealed significant intra-host SARS-CoV-2 variation [82,110,123]. Notably, individual patients who share consensus-level variation may generally have very different intra-host variation [82], highlighting how analyses focused only on inter-host variation may fail to capture the vast majority of relevant SARS-CoV-2 variation and evolution.

Relatedly, evolutionary studies of intra-host and inter-host variation have historically produced conflicting results, whereby intra-host variation is much greater than inter-host variation. For example, analysis of HCMV populations within hosts has reported per-site nucleotide diversity values that that differ by an order of magnitude from estimates of between host per-site nucleotide diversity [104,124,125]. These analyses indicate that approximately 68% of HCMV genomic sites are polymorphic within hosts, while only 12% of sites are segregating among hosts. Some researchers have interpreted these patterns as reflecting that most viral polymorphisms sampled within hosts are strongly deleterious [126,127]. However, population genetic analyses have demonstrated that these observed differences within and among hosts are consistent with most observed viral polymorphisms being nearly neutral with regard to fitness—indeed, standard population genetic models seem to fully capture both observed intra-host and inter-host HCMV variation [91,104,125].

Applying similar approaches to SARS-CoV-2 will be crucial to better understand its evolution and global spread, but this first requires the development of a baseline model as discussed herein. This development, in turn, will be aided by sampling and sequencing SARS-CoV-2 populations from single patients at multiple time points (Fig 4). While still relatively rare, several studies have recently produced time-sampled whole-genome SARS-CoV-2 data [66,82,110,128]. For example, one study sampled 41 patients at 2 time points, collected on average 6 days apart, and observed both generation and loss of intra-host variation during this period [66]. Most importantly, additional time-series data will allow for the intra-host description of temporal changes in allele frequencies providing greatly improved resolution on the parameters underlying selection, population size change, and population structure [129,130], and will further provide the data necessary to perform needed inference to better quantify underlying rates of mutation and recombination.

## Concluding thoughts

Though this virus currently represents a unique global threat, it is also simply an organism like any other that can be studied using basic population genetic principles. While the many considerations here described may appear rather complex, it is our hope that these recommendations together with this compendium of recently obtained estimates will prove useful in future efforts to better illuminate the within-patient evolutionary dynamics of SARS-CoV-2. Without this framework to define hypotheses and accurately quantify contributing evolutionary processes, epidemiological and genetic data describing spatial and temporal variations in disease incidence and mutational frequencies will remain merely descriptive and thus will alone be unable to broach the key evolutionary questions of greatest relevance to public health.

## Acknowledgments

We would like to thank Jeremy Kamil and Timothy Kowalik for helpful comments on the manuscript, and Will Conner, David Xing, and the University of Montana Genomics Core for data contributions.

## References

1. Worobey M. Dissecting the early COVID-19 cases in Wuhan. Science. 2021; 374:1202–1204. https://doi.org/10.1126/science.abm4454 PMID: 34793199
2. COVID-19 Excess Mortality Collaborators. Estimating excess mortality due to the COVID-19 pandemic: a systematic analysis of COVID-19-related mortality, 2020–21. Lancet (London, England). 2022; 399:1513–1536. https://doi.org/10.1016/S0140-6736(21)02796-3 PMID: 35279232
3. Weigang S, Fuchs J, Zimmer G, Schnepf D, Kern L, Beer J, et al. Within-host evolution of SARS-CoV-2 in an immunosuppressed COVID-19 patient as a source of immune escape variants. Nat Commun. 2021; 12:6405. https://doi.org/10.1038/s41467-021-26602-3 PMID: 34737266
4. Jensen JD, Stikeleather RA, Kowalik TF, Lynch M. Imposed mutational meltdown as an antiviral strategy. Evolution. 2020; 74:2549–2559. https://doi.org/10.1111/evo.14107 PMID: 33047822
5. Bank C, Ewing GB, Ferrer-Admettla A, Foll M, Jensen JD. Thinking too positive? Revisiting current methods of population genetic selection inference. Trends Genet. 2014; 30:540–546. https://doi.org/10.1016/j.tig.2014.09.010 PMID: 25438719
6. Charlesworth B, Jensen JD. Effects of selection at linked sites on patterns of genetic variability. Annu Rev Ecol Evol Syst. 2021; 52:177–197. https://doi.org/10.1146/annurev-ecolsys-010621-044528
7. Barton NH. The effect of hitch-hiking on neutral genealogies. Genet Res. 1998; 72:123–133. https://doi.org/10.1017/S0016672398003462
8. Thornton KR, Jensen JD. Controlling the false-positive rate in multilocus genome scans for selection. Genetics. 2007; 175:737–750. https://doi.org/10.1534/genetics.106.064642 PMID: 17110489

9. Harris RB, Sackman A, Jensen JD. On the unfounded enthusiasm for soft selective sweeps II: Examining recent evidence from humans, flies, and viruses. PLoS Genet. 2018; 14:e1007859. https://doi.org/10.1371/journal.pgen.1007859 PMID: 30592709
10. Johri P, Stephan W, Jensen JD. Soft selective sweeps: Addressing new definitions, evaluating competing models, and interpreting empirical outliers. PLoS Genet. 2022; 18:e1010022. https://doi.org/10.1371/journal.pgen.1010022 PMID: 35202407
11. Johri P, Aquadro CF, Beaumont M, Charlesworth B, Excoffier L, Eyre-Walker A, et al. Recommendations for improving statistical inference in population genomics. PLoS Biol. 2022; 20:e3001669. https://doi.org/10.1371/journal.pbio.3001669 PMID: 35639797
12. Irwin KK, Laurent S, Matuszewski S, Vuilleumier S, Ormond L, Shim H, et al. On the importance of skewed offspring distributions and background selection in virus population genetics. Heredity. 2016; 117:393–399. https://doi.org/10.1038/hdy.2016.58 PMID: 27649621
13. Jensen JD. Studying population genetic processes in viruses: From drug-resistance evolution to patient infection dynamics. Encyclopedia of Virology. Elsevier; 2021. p. 227–232. https://doi.org/10.1016/B978-0-12-814515-9.00113–2
14. Bar-On YM, Flamholz A, Phillips R, Milo R. SARS-CoV-2 (COVID-19) by the numbers. Elife. 2020: 9. https://doi.org/10.7554/eLife.57309 PMID: 32228860
15. Eckerle LD, Lu X, Sperry SM, Choi L, Denison MR. High fidelity of murine hepatitis virus replication is decreased in nsp14 exoribonuclease mutants. J Virol. 2007; 81:12135–12144. https://doi.org/10.1128/JVI.01296-07 PMID: 17804504
16. Sanjuán R, Nebot MR, Chirico N, Mansky LM, Belshaw R. Viral mutation rates. J Virol. 2010; 84:9733–9748. https://doi.org/10.1128/JVI.00694-10 PMID: 20660197
17. Zhao Z, Li H, Wu X, Zhong Y, Zhang K, Zhang Y-P, et al. Moderate mutation rate in the SARS coronavirus genome and its implications. BMC Evol Biol. 2004; 4:21. https://doi.org/10.1186/1471-2148-4-21 PMID: 15222897
18. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data—from vision to reality. Euro Surveill. 2017: 22. https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494 PMID: 28382917
19. Zhao W-M, Song S-H, Chen M-L, Zou D, Ma L-N, Ma Y-K, et al. The 2019 novel coronavirus resource. Yi Chuan. 2020; 42:212–221. https://doi.org/10.16288/j.yczz.20-030 PMID: 32102777
20. Vasilarou M, Alachiotis N, Garefalaki J, Beloukas A, Pavlidis P. Population genomics insights into the first wave of COVID-19. Life (Basel, Switzerland). 2021; 11:1–19. https://doi.org/10.3390/life11020129 PMID: 33562321
21. Li X, Giorgi EE, Marichannegowda MH, Foley B, Xiao C, Kong X-P, et al. Emergence of SARS-CoV-2 through recombination and strong purifying selection. Sci Adv. 2020: 6. https://doi.org/10.1126/sciadv.abb9153 PMID: 32937441
22. Chaw S-M, Tai J-H, Chen S-L, Hsieh C-H, Chang S-Y, Yeh S-H, et al. The origin and underlying driving forces of the SARS-CoV-2 outbreak. J Biomed Sci. 2020; 27:73. https://doi.org/10.1186/s12929-020-00665-8 PMID: 32507105
23. Díez-Fuertes F, Iglesias-Caballero M, García-Pérez J, Monzón S, Jiménez P, Varona S, et al. A founder effect led early SARS-CoV-2 transmission in Spain. J Virol. 2021: 95. https://doi.org/10.1128/JVI.01583-20 PMID: 33127745
24. Amicone M, Borges V, Alves MJ, Isidro J, Zé-Zé L, Duarte S, et al. Mutation rate of SARS-CoV-2 and emergence of mutators during experimental evolution. Evol Med Public Health. 2022; 10:142–155. https://doi.org/10.1093/emph/eoac010 PMID: 35419205
25. Morales AC, Rice AM, Ho AT, Mordstein C, Mühlhausen S, Watson S, et al. Causes and consequences of purifying selection on SARS-CoV-2. Genome Biol Evol. 2021: 13. https://doi.org/10.1093/gbe/evab196 PMID: 34427640
26. Neher RA. Contributions of adaptation and purifying selection to SARS-CoV-2 evolution. BioRxiv. 2022.08.22. 50473 [Preprint]. 2022 [posted 2022 August 22; cited 2022 October 13]. Available from: https://www.biorxiv.org/content/10.1101/2022.08.22.504731
27. Drake JW, Holland JJ. Mutation rates among RNA viruses. Proc Natl Acad Sci. 1999; 96:13910–13913. https://doi.org/10.1073/pnas.96.24.13910 PMID: 10570172
28. Elena SF, Sanjuán R. Adaptive value of high mutation rates of RNA viruses: separating causes from consequences. J Virol. 2005; 79:11555–11558. https://doi.org/10.1128/JVI.79.18.11555-11558.2005 PMID: 16140732
29. Denison MR, Graham RL, Donaldson EF, Eckerle LD, Baric RS. Coronaviruses: An RNA proofreading machine regulates replication fidelity and diversity. RNA Biol. 2011; 8:270–279. https://doi.org/10.4161/rna.8.2.15013 PMID: 21593585

30. Lynch M. Evolutionary layering and the limits to cellular perfection. Proc Natl Acad Sci. 2012; 109:18851–18856. https://doi.org/10.1073/pnas.1216130109 PMID: 23115338

31. Jensen JD, Lynch M. Considering mutational meltdown as a potential SARS-CoV-2 treatment strategy. Heredity. 2020; 124:619–620. https://doi.org/10.1038/s41437-020-0314-z PMID: 32251365

32. Bank C, Renzette N, Liu P, Matuszewski S, Shim H, Foll M, et al. An experimental evaluation of drug-induced mutational meltdown as an antiviral treatment strategy. Evolution. 2016; 70:2470–2484. https://doi.org/10.1111/evo.13041 PMID: 27566611

33. Hill WG, Robertson A. The effect of linkage on limits to artificial selection. Genet Res. 1966; 8:269–294. Available from: http://www.ncbi.nlm.nih.gov/pubmed/5980116. PMID: 5980116

34. Felsenstein J. The evolutionary advantage of recombination. Genetics. 1974; 78:737–756. https://doi.org/10.1093/genetics/78.2.737 PMID: 4448362

35. Lai MM. Coronavirus: organization, replication and expression of genome. Annu Rev Microbiol. 1990; 44:303–333. https://doi.org/10.1146/annurev.mi.44.100190.001511 PMID: 2252386

36. Lai MM, Cavanagh D. The molecular biology of coronaviruses. Adv Virus Res. 1997; 48:1–100. Available from: http://www.ncbi.nlm.nih.gov/pubmed/9233431. https://doi.org/10.1016/S0065-3527(08)60286-9 PMID: 9233431

37. Graham RL, Baric RS. Recombination, reservoirs, and the modular spike: mechanisms of coronavirus cross-species transmission. J Virol. 2010; 84:3134–3146. https://doi.org/10.1128/JVI.01394-09 PMID: 19906932

38. Goldstein SA, Brown J, Pedersen BS, Quinlan AR, Elde NC. Extensive recombination-driven coronavirus diversification expands the pool of potential pandemic pathogens. BioRxiv: 2021.02.03. 429646 [Preprint]. [posted 2021 February 4; revised 2021 June 28; cited 2022 October 13]. Available from: https://www.biorxiv.org/content/10.1101/2021.02.03.429646. PMID: 33564759

39. Rehman SU, Shafique L, Ihsan A, Liu Q. Evolutionary trajectory for the emergence of novel coronavirus SARS-CoV-2. Pathogens. 2020; 9:240. https://doi.org/10.3390/pathogens9030240 PMID: 32210130

40. Rahimi A, Mirzazadeh A, Tavakolpour S. Genetics and genomics of SARS-CoV-2: A review of the literature with the special focus on genetic diversity and SARS-CoV-2 genome detection. Genomics. 2021; 113:1221–1232. https://doi.org/10.1016/j.ygeno.2020.09.059 PMID: 33007398

41. Li X, Zai J, Zhao Q, Nie Q, Li Y, Foley BT, et al. Evolutionary history, potential intermediate animal host, and cross-species analyses of SARS-CoV-2. J Med Virol. 2020; 92:602–611. https://doi.org/10.1002/jmv.25731 PMID: 32104911

42. Boni MF, Lemey P, Jiang X, Lam TT-Y, Perry BW, Castoe TA, et al. Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. Nat Microbiol. 2020; 5:1408–1417. https://doi.org/10.1038/s41564-020-0771-4 PMID: 32724171

43. Xiao K, Zhai J, Feng Y, Zhou N, Zhang X, Zou J-J, et al. Isolation of SARS-CoV-2-related coronavirus from Malayan pangolins. Nature. 2020; 583:286–289. https://doi.org/10.1038/s41586-020-2313-x PMID: 32380510

44. MacLean OA, Lytras S, Weaver S, Singer JB, Boni MF, Lemey P, et al. Natural selection in the evolution of SARS-CoV-2 in bats created a generalist virus and highly capable human pathogen. PLoS Biol. 2021; 19:e3001115. https://doi.org/10.1371/journal.pbio.3001115 PMID: 33711012

45. Turakhia Y, Thornlow B, Hinrichs A, McBroome J, Ayala N, Ye C, et al. Pandemic-scale phylogenomics reveals the SARS-CoV-2 recombination landscape. Nature. 2022; 609:994–997. https://doi.org/10.1038/s41586-022-05189-9 PMID: 35952714

46. Sabin S, Morales-Arce AY, Pfeifer SP, Jensen JD. The impact of frequently neglected model violations on bacterial recombination rate estimation: a case study in *Mycobacterium canettii* and *Mycobacterium tuberculosis*. G3 (Bethesda). 2022: 12. https://doi.org/10.1093/g3journal/jkac055 PMID: 35253851

47. Jackson B, Boni MF, Bull MJ, Colleran A, Colquhoun RM, Darby AC, et al. Generation and transmission of interlineage recombinants in the SARS-CoV-2 pandemic. Cell. 2021; 184:5179–5188.e8. https://doi.org/10.1016/j.cell.2021.08.014 PMID: 34499854

48. Focosi D, Maggi F. Recombination in coronaviruses, with a focus on SARS-CoV-2. Viruses. 2022; 14:1239. https://doi.org/10.3390/v14061239 PMID: 35746710

49. Dezordi FZ, Resende PC, Naveca FG, do Nascimento VA, de Souza VC, Dias Paixão AC, et al. Unusual SARS-CoV-2 intrahost diversity reveals lineage superinfection. Microb Genom. 2022: 8. https://doi.org/10.1099/mgen.0.000751 PMID: 35297757

50. Zhou H-Y, Cheng Y-X, Xu L, Li J-Y, Tao C-Y, Ji C-Y, et al. Genomic evidence for divergent co-infections of co-circulating SARS-CoV-2 lineages. Comput Struct Biotechnol J. 2022; 20:4015–4024. https://doi.org/10.1016/j.csbj.2022.07.042 PMID: 35915661

51. Amoutzias GD, Nikolaidis M, Tryfonopoulou E, Chlichlia K, Markoulatos P, Oliver SG. The remarkable evolutionary plasticity of coronaviruses by mutation and recombination: Insights for the COVID-19 pandemic and the future evolutionary paths of SARS-CoV-2. Viruses. MDPI; 2022. https://doi.org/10.3390/v14010078 PMID: 35062282

52. Gribble J, Stevens LJ, Agostini ML, Anderson-Daniels J, Chappell JD, Lu X, et al. The coronavirus proofreading exoribonuclease mediates extensive viral recombination. PLoS Pathog. 2021; 17: e1009226. https://doi.org/10.1371/journal.ppat.1009226 PMID: 33465137

53. Baric RS, Fu K, Chen W, Yount B. High recombination and mutation rates in mouse hepatitis virus suggest that coronaviruses may be potentially important emerging viruses. Adv Exp Med Biol. 1995; 380:571–576. https://doi.org/10.1007/978-1-4615-1899-0_91 PMID: 8830544

54. Ignatieva A, Lyngsø RB, Jenkins PA, Hein J. KwARG: Parsimonious reconstruction of ancestral recombination graphs with recurrent mutation. Bioinformatics. 2020; 37:3277–3284. https://doi.org/10.1093/bioinformatics/btab351 PMID: 33970217

55. Ignatieva A, Hein J, Jenkins PA. Ongoing Recombination in SARS-CoV-2 revealed through genealogical reconstruction. Mol Biol Evol. 2022: 39. https://doi.org/10.1093/molbev/msac028 PMID: 35106601

56. Müller NF, Kistler KE, Bedford T. Recombination patterns in coronaviruses. BioRxiv. 2021.04.28.44180 [Preprint]. 2022 [posted 2021 April 28; revised 2022 February 8; cited 2022 October 13]. Available from: https://www.biorxiv.org/content/10.1101/2021.04.28.441806. https://doi.org/10.1101/2021.04.28.441806 PMID: 33948594

57. Wu F, Zhao S, Yu B, Chen Y-M, Wang W, Song Z-G, et al. A new coronavirus associated with human respiratory disease in China. Nature. 2020; 579:265–269. https://doi.org/10.1038/s41586-020-2008-3 PMID: 32015508

58. Zhang Y-Z, Holmes EC. A genomic perspective on the origin and emergence of SARS-CoV-2. Cell. 2020; 181:223–227. https://doi.org/10.1016/j.cell.2020.03.035 PMID: 32220310

59. Beichman AC, Huerta-Sanchez E, Lohmueller KE. Using genomic data to infer historic population dynamics of nonmodel organisms. Annu Rev Ecol Evol Syst. 2018; 49:433–456. https://doi.org/10.1146/annurev-ecolsys-110617-062431

60. Maynard Smith J, Haigh J. The hitch-hiking effect of a favourable gene. Genet Res. 1974; 23:23–35. https://doi.org/10.1017/S0016672308009579 PMID: 4407212

61. Charlesworth B, Morgan MT, Charlesworth D. The effect of deleterious mutations on neutral molecular variation. Genetics. 1993; 134:1289–1303. https://doi.org/10.1093/genetics/134.4.1289 PMID: 8375663

62. Johri P, Charlesworth B, Jensen JD. Toward an evolutionarily appropriate null model: Jointly inferring demography and purifying selection. Genetics. 2020; 215:173–192. https://doi.org/10.1534/genetics.119.303002 PMID: 32152045

63. Johri P, Charlesworth B, Howell EK, Lynch M, Jensen JD. Revisiting the notion of deleterious sweeps. Genetics. 2021: 219. https://doi.org/10.1093/genetics/iyab094 PMID: 34125884

64. Morales-Arce AY, Johri P, Jensen JD. Inferring the distribution of fitness effects in patient-sampled and experimental virus populations: two case studies. Heredity. 2022; 128:79–87. https://doi.org/10.1038/s41437-021-00493-y PMID: 34987185

65. Kepler L, Hamins-Puertolas M, Rasmussen DA. Decomposing the sources of SARS-CoV-2 fitness variation in the United States. Virus Evol. 2021: 7. https://doi.org/10.1093/ve/veab073 PMID: 34642604

66. Lythgoe KA, Hall M, Ferretti L, de Cesare M, MacIntyre-Cockett G, Trebes A, et al. SARS-CoV-2 within-host diversity and transmission. Science. 2021: 372. https://doi.org/10.1126/science.abg0821 PMID: 33688063

67. Tonkin-Hill G, Martincorena I, Amato R, Lawson ARJ, Gerstung M, Johnston I, et al. Patterns of within-host genetic diversity in SARS-CoV-2. Elife. 2021: 10. https://doi.org/10.7554/eLife.66857 PMID: 34387545

68. Ghafari M, du Plessis L, Raghwani J, Bhatt S, Xu B, Pybus OG, et al. Purifying selection determines the short-term time dependency of evolutionary rates in SARS-CoV-2 and pH1N1 influenza. Mol Biol Evol. 2022: 39. https://doi.org/10.1093/molbev/msac009 PMID: 35038728

69. Sanjuán R, Moya A, Elena SF. The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. Proc Natl Acad Sci. 2004; 101:8396–8401. https://doi.org/10.1073/pnas.0400146101 PMID: 15159545

70. Sanjuán R. Mutational fitness effects in RNA and single-stranded DNA viruses: Common patterns revealed by site-directed mutagenesis studies. Philos Trans R Soc Lond B Biol Sci. 2010:1975–1982. https://doi.org/10.1098/rstb.2010.0063 PMID: 20478892

71. Johri P, Eyre-Walker A, Gutenkunst RN, Lohmueller KE, Jensen JD. On the prospect of achieving accurate joint estimation of selection with population history. Genome Biol Evol. 2022: 14. https://doi.org/10.1093/gbe/evac088 PMID: 35675379

72. Johri P, Riall K, Becher H, Excoffier L, Charlesworth B, Jensen JD. The impact of purifying and background selection on the inference of population history: Problems and prospects. Mol Biol Evol. 2021; 38:2986–3003. https://doi.org/10.1093/molbev/msab050 PMID: 33591322

73. Zwart MP, Elena SF. Matters of size: Genetic bottlenecks in virus infection and their potential impact on evolution. Annu Rev Virol. 2015; 2:161–179. https://doi.org/10.1146/annurev-virology-100114-055135 PMID: 26958911

74. Popa A, Genger J-W, Nicholson MD, Penz T, Schmid D, Aberle SW, et al. Genomic epidemiology of superspreading events in Austria reveals mutational dynamics and transmission properties of SARS-CoV-2. Sci Transl Med. 2020; 12:2555. https://doi.org/10.1126/scitranslmed.abe2555 PMID: 33229462

75. Sobel Leonard A, Weissman DB, Greenbaum B, Ghedin E, Koelle K. Transmission bottleneck size estimation from pathogen deep-sequencing data, with an application to human influenza A virus. J Virol. 2017; 91:171–188. https://doi.org/10.1128/JVI.00171-17 PMID: 28468874

76. Shen Z, Xiao Y, Kang L, Ma W, Shi L, Zhang L, et al. Genomic diversity of severe acute respiratory syndrome–coronavirus 2 in patients with coronavirus disease 2019. Clin Infect Dis. 2020; 71: 713–720. https://doi.org/10.1093/cid/ciaa203 Erratum in: Clinical Infectious Diseases. 2021;73:2374. PMID: 32129843

77. Braun KM, Moreno GK, Wagner C, Accola MA, Rehrauer WM, Baker DA, et al. Acute SARS-CoV-2 infections harbor limited within-host diversity and transmit via tight transmission bottlenecks. PLoS Pathog. 2021; 17:e1009849. https://doi.org/10.1371/journal.ppat.1009849 PMID: 34424945

78. Martin MA, Koelle K. Comment on "Genomic epidemiology of superspreading events in Austria reveals mutational dynamics and transmission properties of SARS-CoV-2.". Sci Transl Med. 2021; 13:1803. https://doi.org/10.1126/scitranslmed.abh1803 PMID: 34705523

79. San JE, Ngcapu S, Kanzi AM, Tegally H, Fonseca V, Giandhari J, et al. Transmission dynamics of SARS-CoV-2 within-host diversity in two major hospital outbreaks in South Africa Virus Evol. 2021: 7. https://doi.org/10.1093/ve/veab041 PMID: 34035952

80. Valesano AL, Rumfelt KE, Dimcheff DE, Blair CN, Fitzsimmons WJ, Petrie JG, et al. Temporal dynamics of SARS-CoV-2 mutation accumulation within and across infected hosts. PLoS Pathog. 2021; 17: e1009499. https://doi.org/10.1371/journal.ppat.1009499 PMID: 33826681

81. Du Z, Wang S, Bai Y, Gao C, Lau EHY, Cowling BJ. Within-host dynamics of SARS-CoV-2 infection: A systematic review and meta-analysis. Transbound Emerg Dis. 2022. https://doi.org/10.1111/tbed.14673 PMID: 35907777

82. Wang Y, Wang D, Zhang L, Sun W, Zhang Z, Chen W, et al. Intra-host variation and evolutionary dynamics of SARS-CoV-2 populations in COVID-19 patients. Genome Med. 2021; 13:30. https://doi.org/10.1186/s13073-021-00847-5 PMID: 33618765

83. Ewing GB, Jensen JD. The consequences of not accounting for background selection in demographic inference. Mol Ecol. 2016; 25:135–141. https://doi.org/10.1111/mec.13390 PMID: 26394805

84. Pouyet F, Aeschbacher S, Thiéry A, Excoffier L. Background selection and biased gene conversion affect more than 95% of the human genome and bias demographic inferences. Elife. 2018: 7. https://doi.org/10.7554/eLife.36317 PMID: 30125248

85. Rousselle M, Mollion M, Nabholz B, Bataillon T, Galtier N. Overestimation of the adaptive substitution rate in fluctuating populations. Biol Lett. 2018; 14:20180055. https://doi.org/10.1098/rsbl.2018.0055 PMID: 29743267

86. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 2009; 5: e1000695. https://doi.org/10.1371/journal.pgen.1000695 PMID: 19851460

87. Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M. Robust demographic inference from genomic and SNP data. PLoS Genet. 2013: 9. https://doi.org/10.1371/journal.pgen.1003905 PMID: 24204310

88. Kelleher J, Wong Y, Wohns AW, Fadil C, Albers PK, McVean G. Inferring whole-genome histories in large population datasets. Nat Genet. 2019; 51:1330–1338. https://doi.org/10.1038/s41588-019-0483-y PMID: 31477934

89. Steinrücken M, Kamm J, Spence JP, Song YS. Inference of complex population histories using whole-genome sequences from multiple populations. Proc Natl Acad Sci. 2019; 116:17115–17120. https://doi.org/10.1073/pnas.1905060116 PMID: 31387977

90. Renzette N, Gibson L, Bhattacharjee B, Fisher D, Schleiss MR, Jensen JD, et al. Rapid intrahost evolution of human cytomegalovirus is shaped by demography and positive selection. PLoS Genet. 2013; 9:e1003735. https://doi.org/10.1371/journal.pgen.1003735 PMID: 24086142

91. Sackman A, Pfeifer S, Kowalik T, Jensen J. On the demographic and selective forces shaping patterns of human cytomegalovirus variation within hosts. Pathogens. 2018; 7:16. https://doi.org/10.3390/pathogens7010016 PMID: 29382090

92. Jensen JD, Kowalik TF. A consideration of within-host human cytomegalovirus genetic variation. Proc Natl Acad Sci. 2020; 117:816–817. https://doi.org/10.1073/pnas.1915295117 PMID: 31874930

93. Dolan A, Cunningham C, Hector RD, Hassan-Walker AF, Lee L, Addison C, et al. Genetic content of wild-type human cytomegalovirus. J Gen Virol. 2004; 85:1301–1312. https://doi.org/10.1099/vir.0.79888-0 PMID: 15105547

94. Sijmons S, Thys K, Mbong Ngwese M, Van Damme E, Dvorak J, Van Loock M, et al. High-throughput analysis of human cytomegalovirus genome diversity highlights the widespread occurrence of gene-disrupting mutations and pervasive recombination. J Virol. 2015; 89:7673–7695. https://doi.org/10.1128/JVI.00578-15 PMID: 25972543

95. Khailany RA, Safdar M, Ozaslan M. Genomic characterization of a novel SARS-CoV-2. Gene Rep. 2020; 19:100682. https://doi.org/10.1016/j.genrep.2020.100682 PMID: 32300673

96. Naqvi AAT, Fatima K, Mohammad T, Fatima U, Singh IK, Singh A, et al. Insights into SARS-CoV-2 genome, structure, evolution, pathogenesis and therapies: Structural genomics approach. Biochim Biophys Acta Mol Basis Dis. 2020; 1866:165878. https://doi.org/10.1016/j.bbadis.2020.165878 PMID: 32544429

97. Tellier A, Lemaire C. Coalescence 2.0: A multiple branching of recent theoretical developments and their applications. Mol Ecol. 2014; 23:2637–2652. https://doi.org/10.1111/mec.12755 PMID: 24750385

98. Matuszewski S, Hildebrandt ME, Achaz G, Jensen JD. Coalescent processes with skewed offspring distributions and non-equilibrium demography. Genetics. 2018; 208:323–338. https://doi.org/10.1534/genetics.117.300499 PMID: 29127263

99. Sackman AM, Harris RB, Jensen JD. Inferring demography and selection in organisms characterized by skewed offspring distributions. Genetics. 2019; 211:1019–1028. https://doi.org/10.1534/genetics.118.301684 PMID: 30651284

100. Morales-Arce AY, Harris RB, Stone AC, Jensen JD. Evaluating the contributions of purifying selection and progeny-skew in dictating within-host *Mycobacterium tuberculosis* evolution. Evolution. 2020; 74:992–1001. https://doi.org/10.1111/evo.13954 PMID: 32233086

101. Zárate S, Pond SLK, Shapshak P, Frost SDW. Comparative study of methods for detecting sequence compartmentalization in human immunodeficiency virus type 1. J Virol. 2007; 81:6643–6651. https://doi.org/10.1128/JVI.02268-06 PMID: 17428864

102. Renzette N, Bhattacharjee B, Jensen JD, Gibson L, Kowalik TF. Extensive genome-wide variability of human cytomegalovirus in congenitally infected infants. PLoS Pathog. 2011; 7:e1001344. https://doi.org/10.1371/journal.ppat.1001344 PMID: 21625576

103. Renzette N, Pokalyuk C, Gibson L, Bhattacharjee B, Schleiss MR, Hamprecht K, et al. Limits and patterns of cytomegalovirus genomic diversity in humans. Proc Natl Acad Sci. 2015; 112:E4120–E4128. https://doi.org/10.1073/pnas.1501880112 PMID: 26150505

104. Renzette N, Pfeifer SP, Matuszewski S, Kowalik TF, Jensen JD. On the analysis of intrahost and interhost viral populations: Human cytomegalovirus as a case study of pitfalls and expectations. J Virol. 2017: 91. https://doi.org/10.1128/JVI.01976-16 PMID: 27974561

105. Pokalyuk C, Renzette N, Irwin KK, Pfeifer SP, Gibson L, Britt WJ, et al. Characterizing human cytomegalovirus reinfection in congenitally infected infants: an evolutionary perspective. Mol Ecol. 2017; 26:1980–1990. https://doi.org/10.1111/mec.13953 PMID: 27988973

106. Jary A, Leducq V, Malet I, Marot S, Klement-Frutos E, Teyssou E, et al. Evolution of viral quasispecies during SARS-CoV-2 infection. Clin Microbiol Infect. 2020; 26:1560.e1–1560.e4. https://doi.org/10.1016/j.cmi.2020.07.032 PMID: 32717416

107. Rueca M, Bartolini B, Gruber CEM, Piralla A, Baldanti F, Giombini E, et al. Compartmentalized replication of SARS-Cov-2 in upper vs. lower respiratory tract assessed by whole genome quasispecies analysis. Microorganisms. 2020; 8:1302. https://doi.org/10.3390/microorganisms8091302 PMID: 32858978

108. Xu D, Zhang Z, Wang F-S. SARS-associated coronavirus quasispecies in individual patients. N Engl J Med. 2004; 350:1366–1367. https://doi.org/10.1056/NEJMc032421 PMID: 15044654

109. Park D, Huh HJ, Kim YJ, Son DS, Jeon HJ, Im EH, et al. Analysis of intrapatient heterogeneity uncovers the microevolution of Middle East respiratory syndrome coronavirus. Cold Spring Harb Mol Case Stud. 2016; 2:a001214. https://doi.org/10.1101/mcs.a001214 PMID: 27900364

110. Farjo M, Koelle K, Martin MA, Gibson LL, Walden KK, Rendon G, et al. Within-host evolutionary dynamics and tissue compartmentalization during acute SARS-CoV-2 infection. BioRxiv: 2022.06.21.497047 [Preprint]. 2022 [posted 2022 June 22; revised 2022 June 23; cited 2022 October 13]. Available from: https://www.biorxiv.org/content/10.1101/2022.06.21.497047. https://doi.org/10.1101/2022.06.21.497047

111. Truffot A, Andréani J, Le Maréchal M, Caporossi A, Epaulard O, Germi R, et al. SARS-CoV-2 variants in immunocompromised patient given antibody monotherapy. Emerg Infect Dis. 2021; 27:2725–2728. https://doi.org/10.3201/eid2710.211509 PMID: 34352197

112. Van Cleemput J, van Snippenberg W, Lambrechts L, Dendooven A, D'Onofrio V, Couck L, et al. Organ-specific genome diversity of replication-competent SARS-CoV-2. Nat Commun. 2021; 12:6612. https://doi.org/10.1038/s41467-021-26884-7 PMID: 34785663

113. Amato KA, Haddock LA, Braun KM, Meliopoulos V, Livingston B, Honce R, et al. Influenza A virus undergoes compartmentalized replication in vivo dominated by stochastic bottlenecks. Nat Commun. 2022; 13:3416. https://doi.org/10.1038/s41467-022-31147-0 PMID: 35701424

114. McMahan K, Giffin V, Tostanoski LH, Chung B, Siamatu M, Suthar MS, et al. Reduced pathogenicity of the SARS-CoV-2 omicron variant in hamsters. Med (N Y). 2022; 3:262–268.e4. https://doi.org/10.1016/j.medj.2022.03.004 PMID: 35313451

115. Peacock TP, Brown JC, Zhou J, Thakur N, Sukhova K, Kugathasan R, et al. The altered entry pathway and antigenic distance of the SARS-CoV-2 Omicron variant map to separate 2 domains of spike protein. BioRxiv: 2021.12.31. 474653 [Preprint]. 2022 [posted 2022 January 3; revised 2022 May 13; cited 2022 October 13]. Available from: https://www.biorxiv.org/content/10.1101/2021.12.31.474653. https://doi.org/10.1101/2021.12.31.474653

116. Khare S, Gurry C, Freitas L, Schultz MB, Bach G, Diallo A, et al. GISAID's role in pandemic response. China CDC Wkly. 2021; 3:1049–1051. https://doi.org/10.46234/ccdcw2021.255 PMID: 34934514

117. Chen C, Jiang D, Ni M, Li J, Chen Z, Liu J, et al. Phylogenomic analysis unravels evolution of yellow fever virus within hosts. PLoS Negl Trop Dis. 2018; 12:e0006738. https://doi.org/10.1371/journal.pntd.0006738 PMID: 30188905

118. McCrone JT, Woods RJ, Martin ET, Malosh RE, Monto AS, Lauring AS. Stochastic processes constrain the within and between host evolution of influenza virus. Elife. 2018: 7. https://doi.org/10.7554/eLife.35962 PMID: 29683424

119. Bull RA, Eden J-S, Luciani F, McElroy K, Rawlinson WD, White PA. Contribution of intra- and interhost dynamics to norovirus evolution. J Virol. 2012; 86:3219–3229. https://doi.org/10.1128/JVI.06712-11 PMID: 22205753

120. Gire SK, Goba A, Andersen KG, Sealfon RSG, Park DJ, Kanneh L, et al. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. Science. 2014; 345:1369–1372. https://doi.org/10.1126/science.1259657 PMID: 25214632

121. Andersen KG, Shapiro BJ, Matranga CB, Sealfon R, Lin AE, Moses LM, et al. Clinical sequencing uncovers origins and evolution of lassa virus. Cell. 2015; 162:738–750. https://doi.org/10.1016/j.cell.2015.07.020 PMID: 26276630

122. Grubaugh ND, Weger-Lucarelli J, Murrieta RA, Fauver JR, Garcia-Luna SM, Prasad AN, et al. Genetic drift during systemic arbovirus infection of mosquito vectors leads to decreased relative fitness during host switching. Cell Host Microbe. 2016; 19:481–492. https://doi.org/10.1016/j.chom.2016.03.002 PMID: 27049584

123. Sapoval N, Mahmoud M, Jochum MD, Liu Y, Elworth RAL, Wang Q, et al. SARS-CoV-2 genomic diversity and the implications for qRT-PCR diagnostics and transmission. Genome Res. 2021; 31:635–644. https://doi.org/10.1101/gr.268961.120 PMID: 33602693

124. Sijmons S, Van Ranst M, Maes P. Genomic and functional characteristics of human cytomegalovirus revealed by next-generation sequencing. Viruses. 2014; 6:1049–1072. https://doi.org/10.3390/v6031049 PMID: 24603756

125. Renzette N, Kowalik TF, Jensen JD. On the relative roles of background selection and genetic hitchhiking in shaping human cytomegalovirus genetic diversity. Mol Ecol. 2016; 25:403–413. https://doi.org/10.1111/mec.13331 PMID: 26211679

126. Holmes EC. Patterns of intra- and interhost nonsynonymous variation reveal strong purifying selection in dengue virus. J Virol. 2003; 77:11296–11298. https://doi.org/10.1128/jvi.77.20.11296-11298.2003 PMID: 14512579

**127.** Lassalle F, Depledge DP, Reeves MB, Brown AC, Christiansen MT, Tutill HJ, et al. Islands of linkage in an ocean of pervasive recombination reveals two-speed evolution of human cytomegalovirus genomes. Virus Evol. 2016; 2:vew017. https://doi.org/10.1093/ve/vew017 PMID: 30288299

**128.** Zuckerman NS, Bucris E, Erster O, Mandelboim M, Adler A, Burstein S, et al. Prolonged detection of complete viral genomes demonstrated by SARS-CoV-2 sequencing of serial respiratory specimens. PLoS ONE. 2021; 16:e0255691. https://doi.org/10.1371/journal.pone.0255691 PMID: 34351998

**129.** Foll M, Poh Y-P, Renzette N, Ferrer-Admetlla A, Bank C, Shim H, et al. Influenza virus drug resistance: a time-sampled population genetics perspective. PLoS Genet. 2014; 10:e1004185. https://doi.org/10.1371/journal.pgen.1004185 PMID: 24586206

**130.** Foll M, Shim H, Jensen JD. WFABC: a Wright-Fisher ABC-based approach for inferring effective population sizes and selection coefficients from time-sampled data. Mol Ecol Resour. 2015; 15:87–98. https://doi.org/10.1111/1755-0998.12280 PMID: 24834845

**131.** 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. https://doi.org/10.7554/eLife.77433 PMID: 35723575

**132.** 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; 376:1327–1332. https://doi.org/10.1126/science.abm1208 PMID: 35608456

**133.** Jensen JD, Charlesworth B. Population genetic considerations regarding evidence for biased mutation rates in *Arabidopsis thaliana*. Mol Biol Evol. 2023; 40:msac275. https://doi.org/10.1093/molbev/msac275 PMID: 36572441

**134.** Goyal A, Reeves DB, Cardozo-Ojeda EF, Schiffer JT, Mayer BT. Viral load and contact heterogeneity predict SARS-CoV-2 transmission and super-spreading events. Elife. 2020; 10:e63537. https://doi.org/10.7554/eLife.63537 PMID: 33620317