A Simulation Framework for Modeling the Within-Patient Evolutionary Dynamics of SARS-CoV-2

John W. Terbot II [1,2,*], Brandon S. Cooper[2], Jeffrey M. Good[2], and Jeffrey D. Jensen [1,*]

[1]School of Life Sciences, Center for Evolution & Medicine, Arizona State University, Tempe, Arizona, USA

[2]Division of Biological Sciences, University of Montana, Missoula, Montana, USA

*Corresponding authors: E-mails: jterbot@asu.edu; jeffrey.d.jensen@asu.edu.

Accepted: November 07, 2023

Abstract

The global impact of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has led to considerable interest in detecting novel beneficial mutations and other genomic changes that may signal the development of variants of concern (VOCs). The ability to accurately detect these changes within individual patient samples is important in enabling early detection of VOCs. Such genomic scans for rarely acting positive selection are best performed via comparison of empirical data with simulated data wherein commonly acting evolutionary factors, including mutation and recombination, reproductive and infection dynamics, and purifying and background selection, can be carefully accounted for and parameterized. Although there has been work to quantify these factors in SARS-CoV-2, they have yet to be integrated into a baseline model describing intrahost evolutionary dynamics. To construct such a baseline model, we develop a simulation framework that enables one to establish expectations for underlying levels and patterns of patient-level variation. By varying eight key parameters, we evaluated 12,096 different model–parameter combinations and compared them with existing empirical data. Of these, 592 models (~5%) were plausible based on the resulting mean expected number of segregating variants. These plausible models shared several commonalities shedding light on intrahost SARS-CoV-2 evolutionary dynamics: severe infection bottlenecks, low levels of reproductive skew, and a distribution of fitness effects skewed toward strongly deleterious mutations. We also describe important areas of model uncertainty and highlight additional sequence data that may help to further refine a baseline model. This study lays the groundwork for the improved analysis of existing and future SARS-CoV-2 within-patient data.

Key words: evolutionary genomics, population genetics, SARS-CoV-2, viral evolution.

Significance

Despite its tremendous impact on human health, a comprehensive evolutionary baseline model has yet to be developed for studying the within-host population genomics of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Importantly, such modeling would enable improved analysis and provide insights into the key evolutionary dynamics governing SARS-CoV-2 evolution. Given this need, we have here quantified a set of plausible baseline models via large-scale simulation. The commonly shared features of these relevant models—including severe infection bottlenecks, low levels of progeny skew, and a high rate of strongly deleterious mutations—lay the foundation for sophisticated analyses of SARS-CoV-2 evolution within patients using these baseline models.

Introduction

The emergence of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in late 2019 is the most impactful human pathogen to arise thus far in the 21st century. Since its emergence, SARS-CoV-2 has been directly responsible for nearly 8 million deaths as of December 2022 (Institute for Health Metrics and Evaluation 2022, Hay and Murray 2023). However, the true impact of SARS-CoV-2—considering underreporting, late reporting, and indirect deaths (e.g., via strained healthcare systems)— is likely much greater, with worldwide excess mortality estimated to exceed 14 million as of December 2021 (Wang et al. 2022, Msemburi et al. 2023). Moreover, SARS-CoV-2 continues to persist worldwide and appears likely to become an endemic virus going forward, as observed with other human coronaviruses (HCoV-229E, -NL63, -OC43, and -HKU1; Corman et al. 2018).

Understanding the evolutionary dynamics of SARS-CoV-2 and predicting new variants of concern (VOCs) that may result in waves of increased infection and mortality remains vital. Although the tools for studying these questions as framed via interhost spread of SARS-CoV-2 have been well-developed (Rambaut et al. 2020, O'Toole et al. 2021), the ability to study intrahost evolution of SARS-CoV-2 is considerably less established. Transmission between hosts is an obviously important stage in viral spread; however, interhost spread is a brief portion of the viral life cycle, with the entirety of viral reproductive activity occurring within a host. Thus, dissecting the intrahost evolutionary dynamics are of key importance in monitoring contemporary and future SARS-CoV-2 spread. In particular, mutations that influence evasion of the host immune system, increase success in cell invasion, and otherwise improve the successful completion of metabolic and reproductive tasks within a host cell could all be of clinical consequence. Given complete information, such mutations would first be detectable within a single host. Although strongly beneficial mutations can eventually become identifiable when observing interhost data through their increased prevalence within the metapopulation (should they escape stochastic loss), the evolutionary dynamics that ultimately dictate their spread will be determined at the intrahost level.

For the viral population within a single patient, these episodic, beneficial mutations may be expected to modify patterns of genomic variation in a manner that would deviate from background patterns produced under constantly operating evolutionary processes (e.g., via a selective sweep of the beneficial mutation; see review of Charlesworth and Jensen 2021). As such, as with any natural population, the study of viral intrahost evolution requires the construction of an evolutionary "null" model to quantify expected baseline levels and patterns of genomic variation (Johri et al. 2020, Jensen 2021, Terbot et al. 2023). At a minimum, a viral baseline model should include mutation, recombination, reproductive dynamics, purifying and background selection, and the history of bottlenecks and growth characterizing patient infection (Irwin et al. 2016). Without comparison with such a baseline model, it is not possible to determine if observed within-patient allele frequencies are attributable to these common evolutionary processes or to the comparatively rare action of positive selection (Barton 1998, Johri, Stephan, et al. 2022).

In this study, we present a first attempt to construct such a model and narrow the range of key parameter values governing the intrahost evolutionary dynamics of SARS-CoV-2. Using forward simulations and comparison with existing patient data summarizing intrahost variation, we identify more and less likely areas of parameter space. We find that transmission and infection bottlenecks appear to be severe (i.e., on the order of <5 virions) in plausible models, consistent with other recent results for SARS-CoV-2 (Lythgoe et al. 2021, Martin and Koelle 2021, Bendall et al. 2023) and the transmission of other airborne viruses like seasonal influenza (McCrone et al. 2018, Valesano et al. 2020). We also describe important areas of uncertainty and correlations between parameter values. For example, if the distribution of new mutational effects is heavily skewed toward strongly deleterious mutations, the range of uncertainty in other parameter values is inflated. Furthermore, we highlight additional data that may help to further narrow this parameter space. For example, the commonly employed 2% minor allele frequency threshold cutoff greatly limits model resolution and may be improved by higher-coverage sequencing of individual patient samples that increases the confidence in individual single nucleotide polymorphism (SNP) calls. Thus, the presented exploration of the SARS-CoV-2 evolutionary parameter space will be valuable in informing future modeling studies, in interpreting newly emerging patient data, and in guiding future data collection.

Results

A total of 12,096 model–parameter combinations were simulated (tables 1 and 2 and fig. 1) in five replicates each; for each replicate, the number of SNPs segregating with an allele frequency of 2% or greater was tallied. Average SNP counts for each model–parameter combination were then compared with a threshold range—(0, 5]—based on existing empirical data (table 3). The great majority was rejected for generating too few or too many segregating SNPs, leaving 592 models remaining (supplementary table S1, Supplementary Material online). Due to the extended nature of the models, the feasible model–parameter sets can be readily categorized according to the 108 possible combinations of the required parameters (fig. 1). Of these,

Table 1

Outline of Model Complexity

Model	+*k*	Parameters
Bottleneck	4	Bottleneck, mutation rate (μ), infection duration, carrying capacity (K)
+ Recombination	1	Recombination rate (R)
+ Progeny skew	2	Probability of multiple coalescent event (Ξ), burst size
+ DFE	1	Ratio of neutral to strongly deleterious mutations

Each row represents an extended model which includes all parameters in that row and from rows above it. The number of parameters added by each stage is detailed in column "+*k*", and the specific parameters added are described in the rightmost column.

Table 2

Parameter Values Used in Models

	Burn-In	Lowest	Low	Midpoint	High	Units	Citations
Initial bottleneck	N/A	N/A	1	5	100	Virions	Popa et al. 2020, Braun et al. 2021, Lythgoe et al. 2021, Martin and Koelle 2021, Bendall et al. 2023
Infection duration	2.50e4	168 (7)	336 (14)	672 (28)	1008 (42)	Hours (days)	Bar-On et al. 2020, Lauer et al. 2020, Li et al. 2020, Du et al. 2022, Wu et al. 2022, table 3
Mutation rate	2.14e-6	N/A	2.14e-7	2.14e-6	2.14e-5	Mut/nt/cycle	Terbot et al. 2023
Carrying capacity	1.00e5	N/A	5e3	5e4	1e5	Virions	Bar-On et al. 2020, Sender et al. 2021
Recombination rate	5.50e-5	N/A	1e-5	5.5e-5	1e-4	Events/nt/ cycle	Terbot et al. 2023
Progeny skew probability	0.003	N/A	0.0001	0.003	0.1	N/A	Bar-On et al. 2020, Sender et al. 2021
Progeny skew amount	100	N/A	20	100	200	Virions	Based on computational limits
DFE (neutral: deleterious)	1:1	N/A	4:1	1:1	1:4	N/A	Terbot et al. 2023

Each parameter, other than infection duration, had three possible levels: a low point estimate, a high point estimate, and a midpoint estimate. The column labeled "Burn-In" details the parameter levels used during the common burn-in; note that aside from infection duration and carrying capacity, all burn-in parameter levels used were the midpoint value. The highest carrying capacity and an extended infection duration were used to allow mutations to accumulate and begin to reach equilibrium as may be expected across the entire metapopulation of a pathogen during a prolonged pandemic.

18 specific parameter combinations produced viable models, highlighting which parameters are likely to have the strongest influence on intrahost evolution of SARS-CoV-2. We found that all plausible models included a severe infection bottleneck (i.e., a bottleneck under five virions appears most consistent with the data, represented here by a bottleneck of one; figs. 2 and 3). Importantly, this number pertains only to successful infecting virions; viruses that have low initial infection rates in new hosts (per virion) may require the physical transmission of many virions in order to achieve infection while still being considered to have a severe infection bottleneck within a population genetic context. Most retained models were also characterized by low or midpoint mutation rates; only three models with the higher mutation rate among the ranges considered were retained. All three of these models were full models using a distribution of fitness effects (DFE) with the largest proportion of strongly deleterious mutations, the highest carrying capacity, and briefest infection duration; the parameter values for recombination and progeny skew varied between these models. In contrast, all 12 required parameter sets using the lowest mutation rate and a severe bottleneck of 1 cleared the SNP threshold, and 5 of the 12 sets with the midpoint mutation rate resulted in plausible models. Generally, larger carrying capacities resulted in better fitting models (4/12 parameter sets had plausible models with the lowest value of carrying capacity and a bottleneck of 1, 6/12 with the midpoint value for carrying capacity, and 8/12 for the highest carrying capacity).

Within these required parameter sets, there are a total of 112 extended models (1 bottleneck model, 3 with the addition of recombination, 27 with the additions of recombination and progeny skew, and 81 with the additions of recombination, progeny skew, and a DFE); the number of models that cleared the filtering requirement varied considerably between the required parameter sets. Briefly, 9 of the 18 required parameter sets had bottleneck models consistent with the empirical data (of which 8 had examples of all extended models clearing the threshold as well), 11 of the 18 had consistent recombination-only models, 10 out of 18 had consistent recombination + progeny skew models, and all 18 had consistent full models. Generally, if a

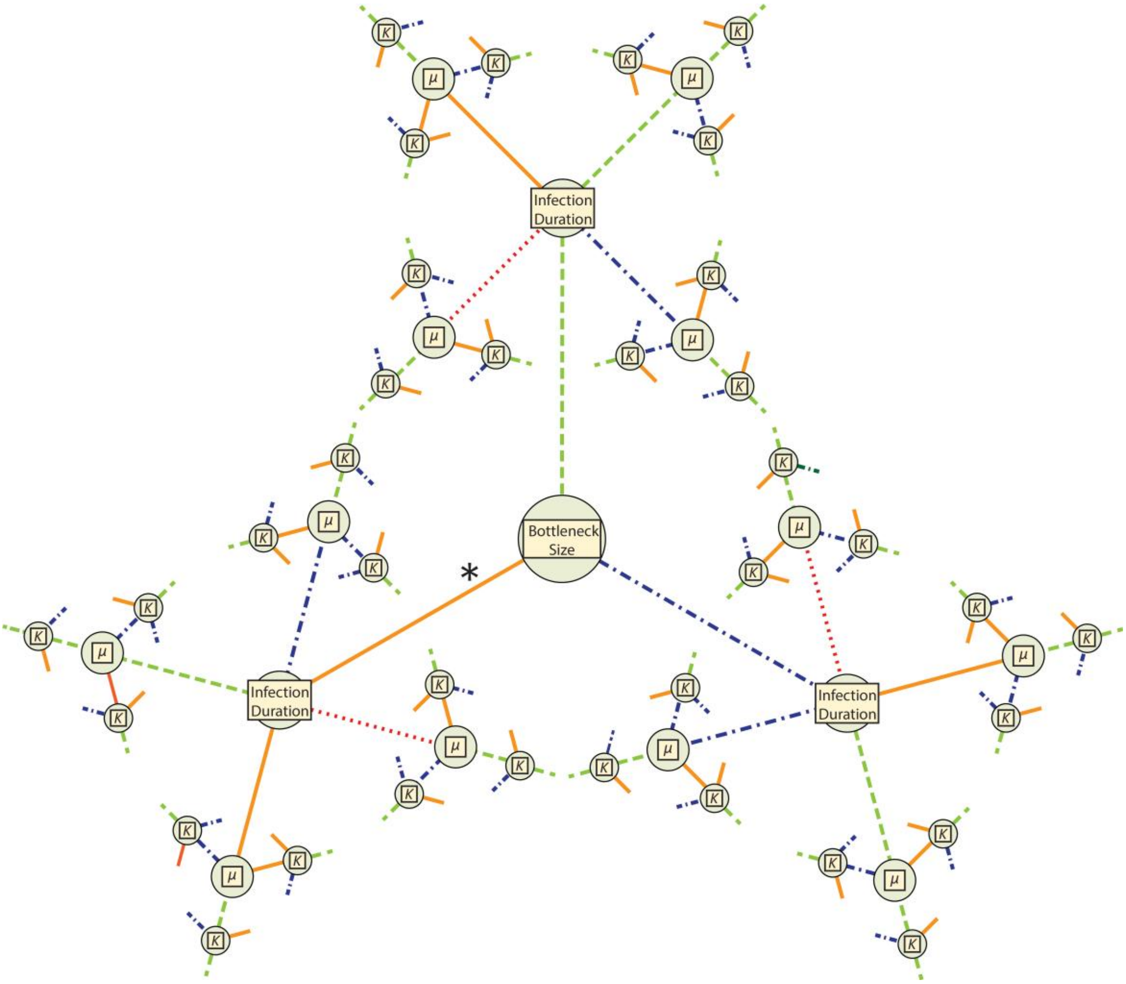

Fig. 1—Representations of the total parameter space for the examined models. Each required parameter (i.e., bottleneck size, infection duration, mutation rate [μ], and carrying capacity [K]) is represented by a node. Line color and shape correspond to parameter value (see table 2): dotted, red is lowest; solid, orange is low (the lowest value for bottleneck size, μ, and K); dashed, green is the midpoint value; and dotted–dashed, blue is the high value. The asterisk (*) denotes the only area of parameter space that had plausible models: those with bottlenecks of 1 (see fig. 2).

Table 3
Summary of Empirical Studies Referenced

Paper	Time to Sample	MAF Filter (%)	Intrahost SNPs
Lythgoe et al. 2021	"Symptomatic individuals on admission to the hospital"	3	1.4 (mean)
Valesano et al. 2021	−7 to 20 days post symptom onset	2	1 (median); 0–2 (IQR)
Wang et al. 2021	10 to 37 days post symptom onset (18.09 days [mean])	5	7.33 (mean); 1–23 (range)
Bendall et al. 2023	Symptoms and positive test within 7 days	2[a]	~0.82 (mean); 0–5 (range)
Gu et al. 2023	Within 5 days of symptom onset (2.3 days [mean])	2.50	10.05 (mean); 5 (median)

Sampling schema and results from empirical studies that reported the number of intrahost SNPs identified in patient samples. Note that all studies apart from Lythgoe et al. (2021) utilized serial sampling within at least one patient during their data collection; this did not impact their reported number of intrahost SNPs but did allow for confirmation of SNPs as true positives.

[a] Intrahost SNP had to clear MAF filter in two technical replicates.

simpler model could be accepted, then more complicated models extending that model also tended to produce plausible results. The distribution of these nonrejected models is more fully detailed in table 4 and fig. 2. In terms of general patterns, models with severe infection bottlenecks, lower progeny skews, and DFEs containing a greater proportion of strongly deleterious mutations were more likely to be accepted (fig. 3, and see Discussion).

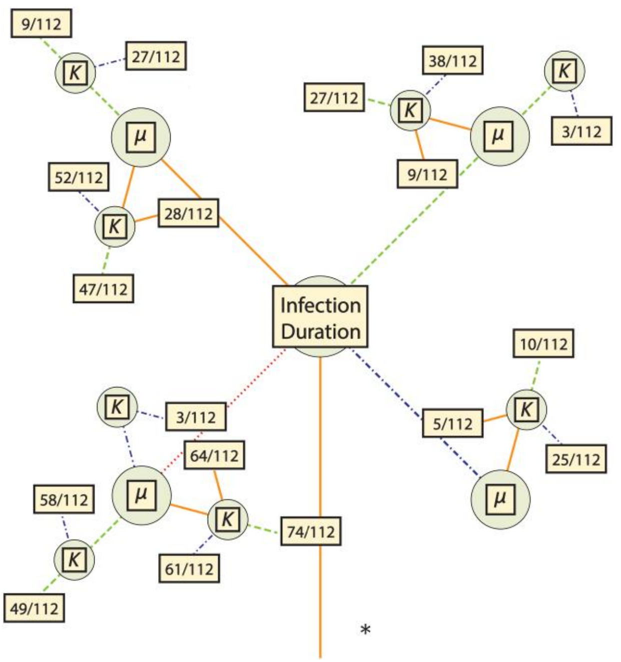

Fig. 2—Parameter space for all plausible models. Line color and shape correspond to the parameter values as in fig. 1. For each set of required parameters, a box is included that contains the fraction of plausible models out of the total number of models using that set. Further information on the specific models which were found to be plausible is detailed in table 4.

Discussion

The evaluation of intrahost population genomic data from SARS-CoV-2 patient samples has the potential to detect signatures of positive selection and allow for the early identification of VOCs. However, it is important to recognize that levels and patterns of genetic variation in a population are the result of a variety of factors including mutation, recombination, reproductive dynamics, purifying and background selection, and the history of bottlenecks and growth characterizing patient infection. One of the key methods of differentiating between these alternative explanations is through comparisons between empirical data and plausible simulated data (Irwin et al. 2016, Johri et al. 2020, Jensen 2021). However, given that the possible model and parameter space is essentially infinite, a key first step in developing an evolutionary baseline model is determining the relevant parameter bounds.

Our study supports three main conclusions that help to define a feasible parameter space governing intrahost SARS-CoV-2 evolutionary dynamics. First, our results support recent conclusions that SARS-CoV-2 transmission is generally associated with severe infection bottlenecks of one or a few virions (Bendall et al. 2023). Specifically, we found a bottleneck of one to be the only bottleneck size tested that produced plausible models. However, the unexamined space between one and five (the midpoint value for bottlenecks) may also produce plausible models as they were not tested in this study. Regardless, a bottleneck of one to four virions conforms with the recent empirical findings based on patterns of intrahost SARS-CoV-2 variation between likely transmissions pairs (Bendall et al. 2023).

Secondly, our results highlight the importance of considering the pervasive effects of purifying and background selection, via a consideration of the DFE, in constraining levels of variation segregating above the 2% frequency threshold. The widespread production of strongly deleterious mutations appears necessary to explain the apparent contradiction between the high mutation rates of RNA viruses (Drake and Holland 1999, Elena and Sanjuán 2005, Jensen et al. 2020) and the generally low number of SNPs identified in patients (Lythgoe et al. 2021, Valesano et al. 2021, Wang et al. 2021, Bendall et al. 2023, Gu et al. 2023). Given that coding regions comprise most of the SARS-CoV-2 genome, this result is not particularly surprising. This importance may be observed in the simulated parameter sets: among the models that accumulated only neutral mutations, 11.8% were plausible; among the models with primarily neutral mutations, 23.9% were plausible; among the models with equivalent rates of neutral and strongly deleterious mutations, 28.4% were plausible; and among the models with primarily strongly deleterious mutations, 42.8% were plausible. Current best estimates of the true DFE underlying SARS-CoV-2 intrahost evolution suggest a bimodal DFE (Flynn et al. 2022, Terbot et al. 2023) with peaks centered around strongly deleterious and neutral fitness effects. At a minimum, between 15% and 20% of sites in the SARS-CoV-2 genome seem to be entirely invariant (Neher 2022), providing a minimum estimate for the proportion of strongly deleterious mutations in the DFE. However, other studies indicate that this value is more likely in the range of 40–50% (Flynn et al. 2022, Terbot et al. 2023). Therefore, although the DFE most severely skewed toward strongly deleterious mutations produced the most plausible models, it is unlikely to reflect the true proportion of strongly deleterious mutations. However, it is notable that the two other DFE distributions reflecting current best understandings of the true DFE both produced around twice as many viable models as models including only neutral mutations, emphasizing the central importance of purifying selection in the evolution of SARS-CoV-2 within patients.

Finally, lower progeny skew values produce more plausible models. Specifically, models with progeny skew and containing either lower probabilities of skewed offspring events and/or lower burst sizes were overrepresented amongst acceptances (397/555 plausible models, or 71.5%). Reproduction of SARS-CoV-2 within a host cell is generally nonlytic and instead involves release of new virions through continuous budding (Bar-On et al. 2020; Park et al. 2020). The prominence of lower levels of progeny skew in plausible models is consistent with a lack of

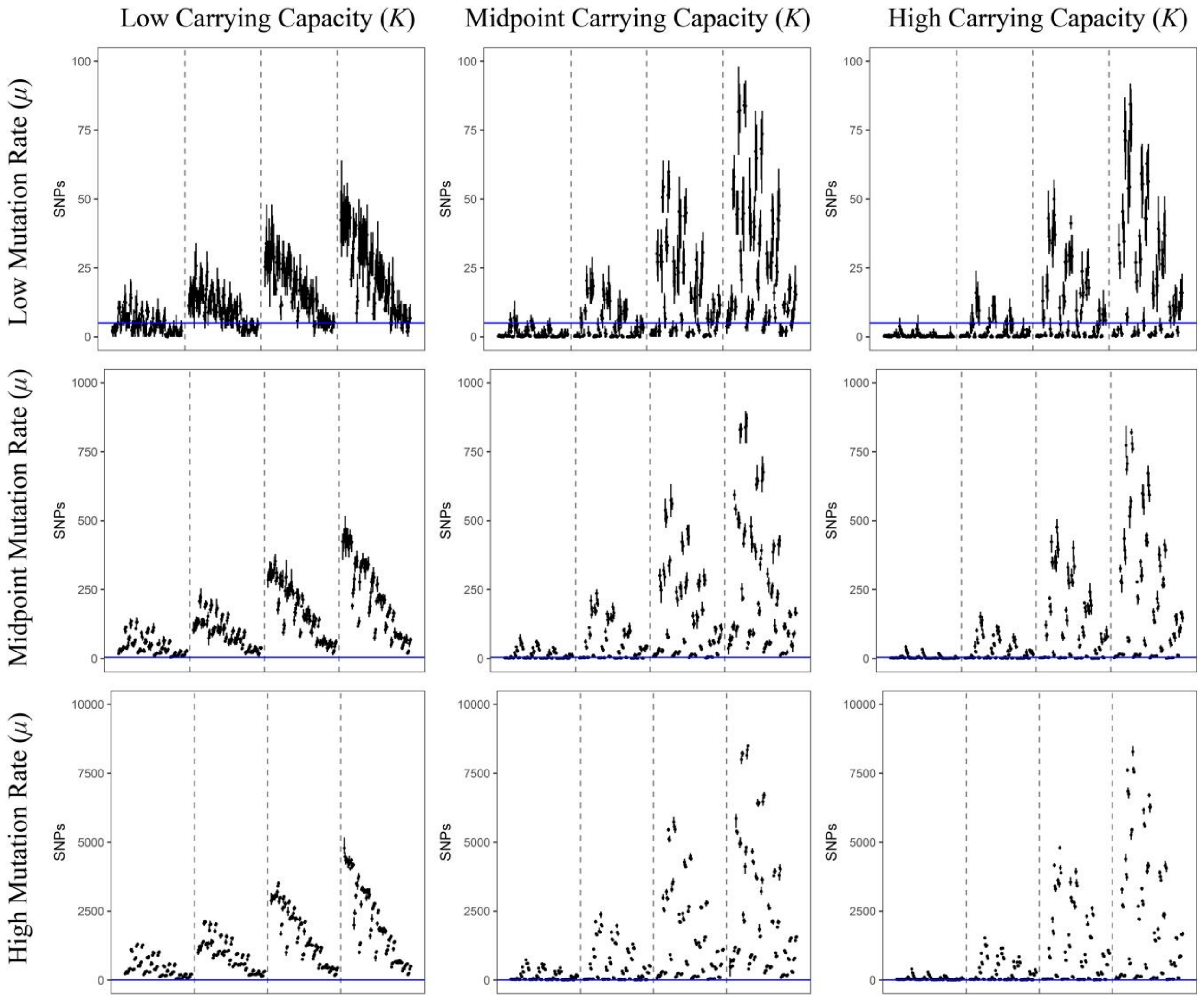

Fig. 3—Each line represents the range of filtered SNPs for a particular model using a sampling of 1,000 genomes, and each point is the mean of that model's replicates. All models in this figure used the lowest bottleneck size (i.e., 1); the carrying capacity used increases in panels from left to right, and the value of the mutation rate used increases in panels from top to bottom. Within each panel, dashed lines separate models into subpanels with different infection durations, increasing from left to right. Within each subpanel, the order of the models is the same and is detailed in supplementary table S3, Supplementary Material online. The blue, horizontal line represents the threshold of 5 SNPs used in this study to accept or reject a potential model (note that *y* axes differ by row). Models with a mean value lower than this line (but not zero) were accepted as plausible. Similar figures for models using larger bottlenecks are available as supplementary figs. 1 and 2, Supplementary Material online.

a single large burst of reproduction (i.e., lower values of burst size). Alternatively, it may be related to the production of subgenomic RNA which are not packaged directly into offspring virions, but through recombination, mutations present in their sequences can be incorporated into offspring virions (i.e., lower values of *Ξ*; Langsjoen et al. 2020, Gribble et al. 2021).

Further model differentiation is limited by current data available for intrahost variation of SARS-CoV-2. The 2% allele frequency cutoff used in this study mirrors the cutoff used in several previous studies (table 3). Although this criterion successfully narrowed the relevant evolutionary parameter space in our simulations, the scarce number of SNPs is insufficient to explore more sophisticated statistics related to the site frequency spectrum or linkage disequilibrium. However, new mutations arising during a typical patient infection will be rare and thus may be missed when applying a frequency-based filter. There is of course an inherent trade-off between including lower frequency SNPs and reducing the confidence in individual SNP calls (Jacot et al. 2021). This trade-off can be partially ameliorated via the generation of high-quality and higher-depth sequencing. A rule of thumb proposed by Lauring (Lauring 2020) suggests that coverage should be 10 times the inverse of

Table 4

Number of Plausible Models Given Required Parameters and Model Complexity

Infection Duration	Mutation Rate (μ)	Carrying Capacity (K)	Bottleneck Count	Bottleneck Percent (%)	Recomb. Count	Recomb. Percent (%)	Prog. Skew Count	Prog. Skew Percent	Full Count (%)	Full Percent (%)	All Count	All Percent (%)
Lowest	Low	Low	1	100.00	2	66.70	6	22.20	55	67.90	64	57.10
Lowest	Low	Midpoint	1	100.00	3	100.00	14	51.90	56	69.10	74	66.10
Lowest	Low	High	1	100.00	2	66.70	19	70.40	39	48.10	61	54.50
Lowest	Midpoint	Midpoint	1	100.00	3	100.00	8	29.60	37	45.70	49	43.80
Lowest	Midpoint	High	1	100.00	3	100.00	11	40.70	43	53.10	58	51.80
Lowest	High	High	0	0.00	0	0.00	0	0.00	3	3.70	3	2.70
Low	Low	Low	0	0.00	0	0.00	0	0.00	28	34.60	28	25.00
Low	Low	Midpoint	1	100.00	3	100.00	10	37.00	33	40.70	47	42.00
Low	Low	High	0	0.00	3	100.00	12	44.40	40	49.40	52	46.40
Low	Midpoint	Midpoint	0	0.00	0	0.00	0	0.00	9	11.10	9	8.00
Low	Midpoint	High	0	0.00	3	100.00	3	11.10	21	25.90	27	24.10
Midpoint	Low	Low	0	0.00	0	0.00	0	0.00	9	11.10	9	8.00
Midpoint	Low	Midpoint	1	100.00	2	66.70	1	3.70	23	28.40	27	24.10
Midpoint	Low	High	1	100.00	3	100.00	9	33.30	25	30.90	38	33.90
Midpoint	Midpoint	High	0	0.00	0	0.00	0	0.00	3	3.70	3	2.70
High	Low	Low	0	0.00	0	0.00	0	0.00	5	6.20	5	4.50
High	Low	Midpoint	0	0.00	0	0.00	0	0.00	10	12.30	10	8.90
High	Low	High	1	100.00	1	33.30	0	0.00	23	28.40	25	22.30

Counts and percentage of models within a given set of values for the required parameters of carrying capacity, mutation rate, and infection duration detailed in the first three columns (the initial bottleneck for all accepted models was the low value [i.e., a single virion]). Each set of required parameters had a single bottleneck model (columns 4 and 5), 3 recombination models (columns 6 and 7), 27 progeny skew models (columns 8 and 9), and 81 full models (columns 10 and 11). The final 2 columns represent the count and percentage of accepted models (out of a possible 112) given a set of required parameters. Counts and percentage of models within a given set of values for the required parameters of carrying capacity, mutation rate, and infection duration are detailed in the first three columns (the initial bottleneck for all accepted models was the low value [i.e., a single virion]). Each set of required parameters had a single bottleneck model (columns 4 and 5), 3 recombination models (columns 6 and 7), 27 progeny skew models (columns 8 and 9), and 81 full models (columns 10 and 11). The final two columns represent the count and percentage of accepted models (out of a possible 112) given a set of required parameters.

the variant's frequency. So, a coverage depth of 1,000 would be required to confidently detect variants at 1% of the population, a depth of 2,000 for a 0.5% variant, and so on.

These general bioinformatic recommendations assume that sequencing data will effectively sample intrahost variation without bias. However, most SARS-CoV-2 sequencing protocols rely on targeted PCR or probe-based enrichment of the SARS-CoV-2 genome to reduce background contamination of host nucleotides. Targeted enrichment approaches have been widely used in biology and are usually sensitive to any standing genetic variation that impacts the binding efficacy of enrichment probes or primers (Mamanova et al. 2010, Jones and Good 2016). Indeed, assay-dependent effects have been a constant concern during the development SARS-CoV-2 sequence protocols as demonstrated by the recurrent need to redesign enrichment primers (e.g., the ARTIC protocol) to fully sequence the genomes of newly circulating VOCs (Ulhuq et al. 2023). SARS-CoV-2-targeted enrichment is also sensitive to low viral loads (Lam et al. 2021), which could further limit understanding of changes in intrahost diversity through the time course of an infection. Given these concerns, technical replication of individual patient samples may be required to reliably detect and quantify the frequency of rare variants.

In addition to limitations relating to the detection of rare alleles, our study was also unable to consider the impacts of host compartmentalization (i.e., localized subpopulations within different organs and areas of organ systems) on intrahost genetic diversity. Compartmentalization may be an important factor for SARS-CoV-2 evolution in regard to the production of viral reservoirs in prolonged infections or infection of immune-compromised patients (Gonzalez-Reiche et al. 2023, Normandin et al. 2023). However, there is currently insufficient information regarding the number and connectivity of compartments used by SARS-CoV-2 to allow for this aspect of model complexity to be reasonably parameterized. As more empirical evidence regarding compartmentalization becomes available, future simulation studies may be able to incorporate compartmentalization to determine the impacts—specifically the influence of gene flow and recombination among compartments—on the intrahost population genetics of SARS-CoV-2.

With that said, based on the currently available data, our study has successfully quantified areas of the SARS-CoV-2 evolutionary parameter space that are most plausible. In addition, the simulation framework presented here may be utilized to compare against future sequencing studies, which will likely enable a further narrowing of likely models. However, even the relatively broad parameter space here identified may be utilized to more effectively screen for newly emerging positively selected mutations (e.g., those contributing to the rapid spread of VOCs) and in so doing reduce the traditionally high false-positive rates traditionally associated with such selection scans (Johri et al 2020, Johri, Aquadro, et al. 2022). This work will also likely be useful in providing a baseline simulation for studies looking at intrahost population genetics of SARS-CoV-2 over time within a single host. Such studies may also allow for greater discrimination between baseline models while retaining the use of a robust, minimum minor allele frequency by providing information on how genetic variation accumulates (or persists) over the course of an infection within a single patient.

Materials and Methods

Simulations

We used the SLiM software package (v4.0.1, Haller and Messer 2023) to conduct forward-in-time simulations. We performed all simulations using SLiM's non–Wright-Fisher tick cycle. To represent the single-stranded, haploid genome of one metabolically active virion, we simulated genomes consisting of 30 kb. Each tick of the simulation represented ~1 h, during which all virions in the population produce one "child virion" excluding cases of progeny skew (described further below). We ran four different, extended models to gain insights into the importance of various population genetic factors. The simplest model, hereafter referred to as the bottleneck model, consisted of four parameters: the mutation rate, the initial number of virions drawn from the burn-in period (or bottleneck size), the carrying capacity of the host, and the number of ticks over which the simulation runs, that is, the time between infection and sampling (infection duration). The other three models added key factors: recombination, recombination plus progeny skew, and a full model adding recombination, progeny skew, and a DFE. The latter describes the proportion new mutations that are strongly deleterious or neutral. The model and parameter space are summarized in tables 1 and 2. We based the tested parameter ranges upon the current literature, as recently described in Terbot et al. (2023) unless otherwise stated.

Progeny genomes added mutations according to a single, genome-wide mutation rate. Contrary to SLiM's default behavior for mutations arising at the same site, the simulation retained only the most recent mutation occurring at a given site. The initial bottleneck was performed by drawing virions from a common burn-in simulation, and the size of this bottleneck reflected the range reported in the literature (Popa et al. 2020, Braun et al. 2021, Lythgoe et al. 2021, Martin and Koelle 2021, Bendall et al. 2023). We based the levels used for carrying capacity on the number of virions estimated at peak infection of 10^9–10^{11} (Bar-On et al. 2020, Sender et al. 2021). Due to computational constraints, and because this figure represents a census population size as opposed to an effective population size, we scaled the carrying capacity downward to 5×10^3 to 10^5 virions (a 10^{-4} to 5×10^{-8} scaling). Owing to this scaling of population size, we scaled up the range of mutation and recombination rates accordingly to maintain a constant population-scaled product with the effective population size. The duration of infection was chosen to represent a range of potential times from initial infection to sampling, that is, including both time from initial infection to symptom onset and time from symptom onset to sampling. Given that most empirical studies report sampling in terms of days from symptom onset and variation therein (table 3) and there is considerable uncertainty regarding the incubation time of SARS-CoV-2 from initial infection to symptom presentation (Bar-On et al. 2020, Lauer et al. 2020, Li et al. 2020, Du et al. 2022, Wu et al. 2022), the range of infection durations modeled extends from brief (7 days) to extended (42 days).

To model recombination, we used a single parameter for the genome-wide recombination rate. For each virion, the simulation chose another random virion to serve as the recombination partner, and the simulation used this recombination partner for all progeny produced by the focal virion in that tick cycle, including those experiencing progeny skew. We used a multiple-merger coalescent model of progeny skew (Irwin et al. 2016) which required two parameters: Ξ corresponding to the probability of a virion having multiple reproduction events in a single tick and the size of this reproductive burst. Values of Ξ represent the eclipse time of a virion (10 h, i.e., a value of $\Xi = 0.1$) as the max rate (Bar-On et al. 2020) or the product of this eclipse time and the number of virions present in a cell that are actively budding virions (i.e., 1,000 reproductively active virions per cell multiplied by 10 h, or a value of $\Xi = 0.0001$) as the minimum rate (Sender et al. 2021). The geometric mean of the minimum and maximum value (0.003) was used as the midpoint value. Further empirical and simulated work on the eclipse period and number of metabolically active virions per infected cell could help define these values more precisely. We selected the highest level of burst size (200) by determining the largest possible burst that could still be completed with the maximum levels of Ξ and carrying capacity and the requested computing resources; we selected lower levels as proportions of that maximum value (20 and 100). Finally, we parameterized the DFE as the ratio of strongly deleterious relative to neutral mutations (4:1, 1:1, and 1:4).

Simulations used a common, burn-in source population which was created using a full-model simulation run for 25,000 ticks at the highest value of carrying capacity and the mid-level value for all other parameters. We replicated each parameter combination five times. Therefore, we ran a total of 12,096 model–parameter combinations (represented visually in fig. 1), resulting in 60,480 replicates. To simulate variation in sequencing depth of empirical studies, we independently sampled 100 and 1,000 genomes and stored them as .ms files. A total of 26 replicates failed to complete their initial runs after exceeding their allotted computational resources (supplementary table S2, Supplementary Material online). These were resubmitted with additional computational resources and included alongside replicates that successfully completed their initial submission.

Simulation Assessment

Using the scikit-allel package (Miles et al. 2021) and custom Python script (Van Rossum and Drake 2009), we calculated a set of summary statistics for each .ms file. The primary statistic we used to identify models resulting in levels of variation similar to that observed in empirical patient samples was the total number of SNPs (often referred to as intrahost single nucleotide variations or iSNVs in the SARS-CoV-2 literature). We compared the simulated data with multiple recent studies that have sought to quantify the amount of intrahost variation in SARS-CoV-2 (Lythgoe et al. 2021, Valesano et al. 2021, Wang et al. 2021, Bendall et al. 2023, Gu et al. 2023). To make this comparison, we applied filtering criteria to the simulated data reflecting that implemented in the empirical data (i.e., removing variants segregating below 2% frequency). From the existing literature, the number of SNPs found within an intrahost population of SARS-CoV-2 clearing this frequency threshold tends to be five or fewer, but not zero (table 3). Additionally, we required each parameter combination to clear this threshold for both the 100 and 1,000 genome sampling schemes. We performed model filtering, and figure creation was performed using a custom R script (R Core Team 2018) and the packages ggplot2 (Wickham 2016) and gridExtra (Auguie 2017). All eidos, bash, python, and R scripts have been deposited on GitHub (https://github.com/jwterbot2/SARS-CoV-2_InitialBaselineModel).

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

Research reported in this manuscript was supported by the National Institute of General Medical Sciences of the National Institutes of Health (NIH) under award number P30GM140963 (to J.M.G. and B.S.C.) and by U54GM104940, which funds the Louisiana Clinical and Translational Science Center. 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. We are grateful to Parul Johri and Vivak Soni for their advice regarding simulation design. We would also like to thank the computing resources used during this study: the common burn-in and rerun replicates were run on the Agave computing cluster at ASU; all other simulation replicates were run using the Open Science Pool resources (OSG 2006, Pordes et al. 2007, Sfiligoi et al. 2009).

Data Availability

Scripts and code used in this study along with a copy of data used to generate figures are available online via GitHub (https://github.com/jwterbot2/SARS-CoV-2_InitialBaselineModel).

Literature Cited

Auguie B. 2017. gridExtra: miscellaneous functions for 'grid' graphics. https://cran.r-project.org/package=gridExtra.

Bar-On YM, Flamholz A, Phillips R, Milo R. 2020. SARS-CoV-2 (COVID-19) by the numbers. eLife 9:e57309.

Barton NH. 1998. The effect of hitch-hiking on neutral genealogies. Genet Res. 72:123–133.

Bendall EE, et al. 2023. Rapid transmission and tight bottlenecks constrain the evolution of highly transmissible SARS-CoV-2 variants. Nat Commun. 14:272.

Braun KM, et al. 2021. Acute SARS-CoV-2 infections harbor limited within-host diversity and transmit via tight transmission bottlenecks. PLoS Pathog. 17:e1009849.

Charlesworth B, Jensen JD. 2021. Effects of selection at linked sites on patterns of genetic variability. Annu Rev of Ecol Evol Syst. 52: 177–197.

Corman VM, Muth D, Niemeyer D, Drosten C. 2018. Hosts and sources of endemic human coronaviruses. Adv Virus Res. 100:163–188.

Drake JW, Holland JJ. 1999. Mutation rates among RNA viruses. Proc Natl Acad Sci U S A. 96:13910–13913.

Du Z, et al. 2022. Shorter serial intervals and incubation periods in SARS-CoV-2 variants than the SARS-CoV-2 ancestral strain. J Travel Med. 29:taac052.

Elena SF, Sanjuán R. 2005. Adaptive value of high mutation rates of RNA viruses: separating causes from consequences. J Virol. 79: 11555–11558.

Flynn JM, et al. 2022. Comprehensive fitness landscape of SARS-CoV-2 M^{pro} reveals insights into viral resistance mechanisms. eLife 11: e77433.

Gonzalez-Reiche AS, et al. 2023. Sequential intrahost evolution and onward transmission of SARS-CoV-2 variants. Nat Commun. 14: 3235.

Gribble J, et al. 2021. The coronavirus proofreading exoribonuclease mediates extensive viral recombination. PLoS Pathog. 17: e1009226.

Gu H, et al. 2023. Within-host genetic diversity of SARS-CoV-2 lineages in unvaccinated and vaccinated individuals. Nat Commun. 14:1793.

Haller BC, Messer PW. 2023. SLiM 4: multispecies eco-evolutionary modeling. Am Nat. 201:E127–E139.

Hay SI, Murray CJL. 2023. Conflicting COVID-19 excess mortality estimates—authors' reply. Lancet. 401:433–434.

Institute for Health Metrics and Evaluation. 2022. Covid-19 projections. https://covid19.healthdata.org/global? view=cumulative-deaths&tab=trend. (July 13, 2023)

Irwin KK, et al. 2016. On the importance of skewed offspring distributions and background selection in virus population genetics. Heredity (Edinb). 117:393–399.

Jacot D, Pillonel T, Greub G, Bertelli C. 2021. Assessment of SARS-CoV-2 genome sequencing: quality criteria and low-frequency variants. J Clin Microbiol. 59:e0094421.

Jensen JD. 2021. Encyclopedia of virology. Amsterdam: Elsevier. p. 227–232.

Jensen JD, Stikeleather RA, Kowalik TF, Lynch M. 2020. Imposed mutational meltdown as an antiviral strategy. Evolution 74: 2549–2559.

Johri P, Aquadro CF, et al. 2022. Recommendations for improving statistical inference in population genomics. PLoS Biol. 20:e3001669.

Johri P, Charlesworth B, Jensen JD. 2020. Toward an evolutionarily appropriate null model: jointly inferring demography and purifying selection. Genetics 215:173–192.

Johri P, Stephan W, Jensen JD. 2022. Soft selective sweeps: addressing new definitions, evaluating competing models, and interpreting empirical outliers. PLoS Genet. 18:e1010022.

Jones MR, Good JM. 2016. Targeted capture in evolutionary and ecological genomics. Mol Ecol. 25:185–202.

Lam C, et al. 2021. SARS-CoV-2 genome sequencing methods differ in their abilities to detect variants from low-viral-load samples. J Clin Microbiol. 59:e0104621.

Langsjoen RM, Muruato AE, Kunkel SR, Jaworski E, Routh A. 2020. Differential alphavirus defective RNA diversity between intracellular and extracellular compartments is driven by subgenomic recombination events. mBio 11:e00731-20.

Lauer SA, et al. 2020. The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application. Ann Intern Med. 172:577–582.

Lauring AS. 2020. Within-host viral diversity: a window into viral evolution. Annu Rev Virol. 7:63–81.

Li Q, et al. 2020. Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. N Engl J Med. 382: 1199–1207.

Lythgoe KA, et al. 2021. SARS-CoV-2 within-host diversity and transmission. Science 372:eabg0821.

Mamanova L, et al. 2010. Target-enrichment strategies for next-generation sequencing. Nat Meth. 7:111–118.

Martin MA, Koelle K. 2021. Comment on "genomic epidemiology of superspreading events in Austria reveals mutational dynamics and transmission properties of SARS-CoV-2". Sci Transl Med. 13: eabh1803.

McCrone JT, et al. 2018. Stochastic processes constrain the within and between host evolution of influenza virus. eLife 7:e35962.

Miles A, et al. 2021. cggh/scikit-allel: v1.3.3. doi: 10.5281/zenodo.4759368.

Msemburi W, et al. 2023. The WHO estimates of excess mortality associated with the COVID-19 pandemic. Nature 613:130–137.

Neher RA. 2022. Contributions of adaptation and purifying selection to SARS-CoV-2 evolution. Virus Evol. 8:veac113.

Normandin E, et al. 2023. High-depth sequencing characterization of viral dynamics across tissues in fatal COVID-19 reveals compartmentalized infection. Nat Commun. 14:574.

OSG. 2006. OSPool. doi: 10.21231/906P-4D78.

O'Toole Á, et al. 2021. Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. Virus Evol. 7: veab064.

Park WB, et al. 2020. Virus isolation from the first patient with sars-cov-2 in Korea. J Korean Med Sci. 35:e84.

Popa A, et al. 2020. Genomic epidemiology of superspreading events in Austria reveals mutational dynamics and transmission properties of SARS-CoV-2. Sci Transl Med. 12:eabe2555.

Pordes R, et al. 2007. The open science grid. J Phys Conf Ser. 78: 012057.

Rambaut A, et al. 2020. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol. 5:1403–1407.

R Core Team. 2018. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Sender R, et al. 2021. The total number and mass of SARS-CoV-2 virions. Proc Natl Acad Sci U S A. 118:e2024815118.

Sfiligoi I et al. 2009. The pilot way to grid resources using glideinWMS. In: 2009 WRI world congress on computer science and information engineering. IEEE pp. 428–432. doi: 10.1109/CSIE.2009.950.

Terbot JW, et al. 2023. Developing an appropriate evolutionary baseline model for the study of SARS-CoV-2 patient samples. PLoS Pathog. 19:e1011265.

Ulhuq FR, et al. 2023. Analysis of the ARTIC V4 and V4.1 SARS-CoV-2 primers and their impact on the detection of Omicron BA.1 and BA.2 lineage-defining mutations. Microb Genom. 9: mgen000991.

Valesano AL, et al. 2020. Influenza B viruses exhibit lower within-host diversity than influenza A viruses in human hosts. J Virol. 94: e01710–19.

Valesano AL, et al. 2021. Temporal dynamics of SARS-CoV-2 mutation accumulation within and across infected hosts. PLoS Pathog. 17: e1009499.

Van Rossum G, Drake FL. 2009. Python 3 reference manual. doi: 10.5555/1593511.

Wang Y, et al. 2021. Intra-host variation and evolutionary dynamics of SARS-CoV-2 populations in COVID-19 patients. Genome Med. 13: 30.

Wang H, et al. 2022. Estimating excess mortality due to the COVID-19 pandemic: a systematic analysis of COVID-19-related mortality, 2020–21. Lancet. 399:1513–1536.

Wickham H. 2016. ggplot2: elegant graphics for data analysis. doi: 10.1007/978-3-319-24277-4.

Wu Y, et al. 2022. Incubation period of COVID-19 caused by unique SARS-CoV-2 strains. JAMA Netw Open. 5:e2228008.

Associate editor: Barbara Holland