Chronic infections can generate SARS-CoV-2-like bursts of viral evolution without epistasis

Edwin Rodríguez-Horta[1,2], **John Strahan**[3], **Aaron R. Dinner**[3,†], **and John P. Barton**[1,†]

[1]Department of Computational and Systems Biology, University of Pittsburgh School of Medicine, USA. [2]Group of Complex Systems and Statistical Physics, Department of Theoretical Physics, Physics Faculty, University of Havana, Cuba. [3]Department of Chemistry and James Franck Institute, University of Chicago, Chicago, Illinois 60637, USA.
†Address correspondence to: dinner@uchicago.edu, jpbarton@pitt.edu.

Multiple SARS-CoV-2 variants have arisen during the first years of the pandemic, often bearing many new mutations. Several explanations have been offered for the surprisingly sudden emergence of multiple mutations that enhance viral fitness, including cryptic transmission, spillover from animal reservoirs, epistasis between mutations, and chronic infections. Here, we simulated pathogen evolution combining within-host replication and between-host transmission. We found that, under certain conditions, chronic infections can lead to SARS-CoV-2-like bursts of mutations even without epistasis. Chronic infections can also increase the global evolutionary rate of a pathogen even in the absence of clear mutational bursts. Overall, our study supports chronic infections as a plausible origin for highly mutated SARS-CoV-2 variants. More generally, we also describe how chronic infections can influence pathogen evolution under different scenarios.

Introduction

During the SARS-CoV-2 pandemic, multiple variants of concern (VOC) have arisen and spread widely throughout the human population, driving waves of infections and mortality[1–4]. The spread of new VOCs has been facilitated by their ability to evade adaptive immunity developed by previous infections or vaccines[5,6]. VOC mutations can also increase virus transmissibility in other ways, such as by improving the receptor binding ability of the viral Spike protein or increasing viral load[5,6].

A singular and unexpected characteristic of early VOCs has been their abrupt emergence. New variants such as Alpha, Delta, and Omicron appeared bearing many mutations that had not been previously observed, seemingly making a large evolutionary leap compared to co-circulating variants. This phenomenon is surprising given the tight transmission bottlenecks inferred for SARS-CoV-2 (refs. [7,8]). During acute infections, few mutations are produced and even fewer are expected to be passed on in new infections[7,8]. In principle, one would then expect viral evolution to proceed through the gradual accumulation of advantageous (i.e., transmission-increasing) mutations.

Multiple hypotheses have been put forward to explain the sudden appearance of a new, highly transmissible variant with a large number of novel mutations[9]. One possibility is cryptic transmission, where undetected circulation in humans allows for long-term viral evolution[10–12]. However, given the number of novel mutations observed in VOCs, this scenario would require that variants remain undetected for long periods of time. Circulation in animal reservoirs, followed by subsequent spillover to humans, could also explain the sudden appearance of VOCs with many mutations[13–16]. As an alternative to hidden circulation in unobserved human or animal populations, epistasis (i.e., non-additive effects of mutations on viral transmissibility) has been cited as a possible factor underlying VOC emergence[17–20]. If multiple mutations are needed to confer a significant fitness advantage to the virus, then mutants with a small number of mutations may not be observed at high frequencies in humans.

Based on clinical data, chronic SARS-CoV-2 infections have emerged as a plausible source of highly divergent variants. Typical, acute SARS-CoV-2 infections resolve within days to weeks. However, in some individuals, chronic infections can persist for months. Chronically infected individuals often have compromised immune systems that are unable to fully clear infections[21–24]. During chronic infections, there is sufficient time for SARS-CoV-2 to generate multiple mutations, which can rise in frequency and ultimately fix in the viral population within that individual. Genomic analyses have shown that the rate of accumulation of mutations within chronically infected individuals is higher than the rate of SARS-CoV-2 evolution between individuals[25]. In addition, VOC mutations have been observed in chronically-infected individuals[25,26]. Accelerated selection of antibody evasion mutations has also been observed in long-term infections treated with monoclonal antibodies or convalescent plasma[27].

Given the potential importance of chronic infections in the evolution of SARS-CoV-2, mathematical modelers have begun exploring its anticipated epidemiological effects in theory and simulations[17,18,28,29]. Recent work has incorporated immunocompromised hosts into susceptible/infected/recovered (SIR) epidemiological models that also include some component of within-host viral evolution[17,28]. Smith and Ashby predicted that large jumps in the proportion of novel variants should only be observed when there is a significant amount of epistasis between immune escape mutations and a sufficient proportion of the population is immunocompromised[17]. In other words, the role of immunocompromised hosts in this model is to allow the virus

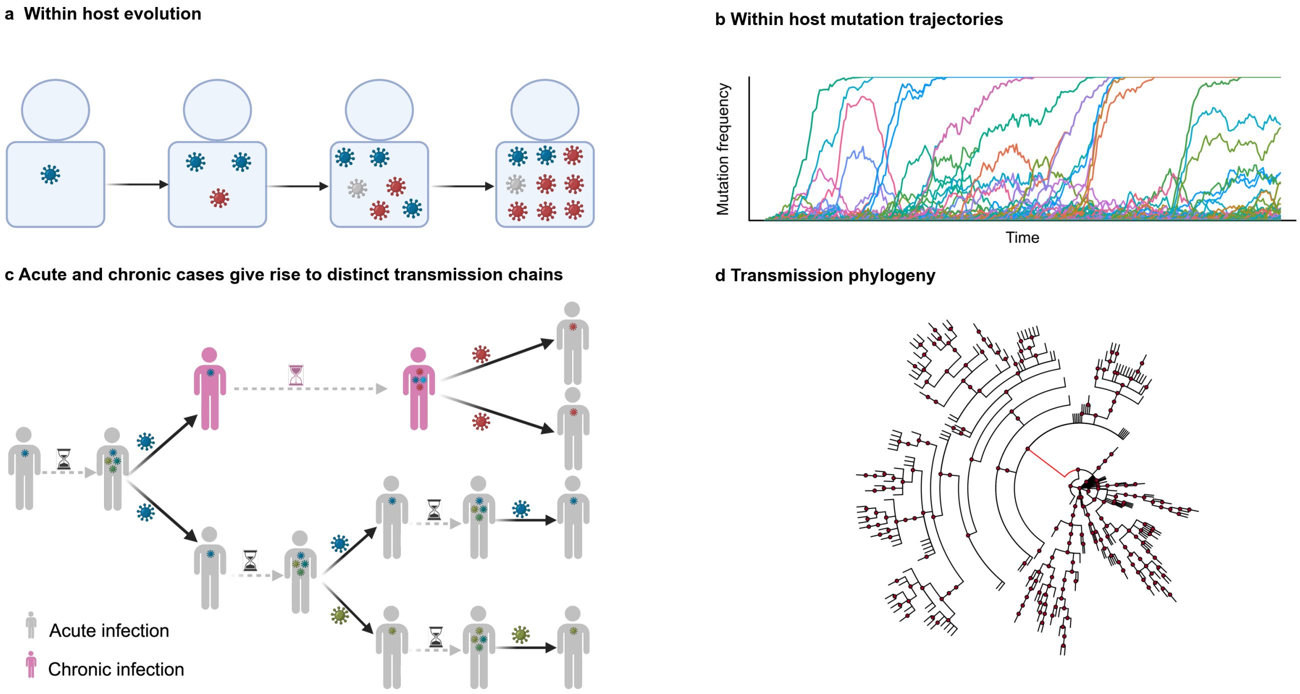

a Within host evolution
b Within host mutation trajectories
Mutation frequency
Time
c Acute and chronic cases give rise to distinct transmission chains
d Transmission phylogeny
Acute infection
Chronic infection

Fig. 1. Global evolutionary model for intra-host and between-host levels of evolution. **a**, The viral population, begins with a single starting genotype and undergoes discrete and non-overlapping generations of Wright-Fisher evolution subject to mutation, selection, and genetic drift. **b**, Example frequency of mutations over time during intrahost evolution. **c**, The population of individuals comprises patients with acute infection and patients with chronic infections. Viral diversity is generated during intrahost evolution. During transmission between acutely infected hosts, most mutations are lost due to tight transmission bottlenecks. Chronic infection can allow for the evolution and transmission of a highly divergent variant. **d**, Example phylogeny for between-host transmission, including transmission from one chronically infected individual (long branch in red). Figures **a** and **c** were created in BioRender.com.

to cross epistatic fitness valleys. Additional work has also considered fitness valley crossing for infections of different durations, but without modeling effects on transmission[29].

In an extensive study, Ghafari et al. considered the effects of chronic infection on the emergence of highly transmissible VOCs[18]. In their model, VOC mutations fix at a constant rate within chronically infected hosts. They consider multiple fitness landscapes for transmission between individuals, including models where VOC mutations make equal, additive contributions to transmission and "plateau-crossing" models where individual mutations have small effects until a critical number are accumulated. They concluded that chronic infection could facilitate the emergence of VOCs, defined as variants with specific transmission-increasing mutations, especially with plateau-crossing fitness landscapes.

Here, we develop a generic model of pathogen evolution, coupling evolution within hosts and transmission between individuals. The primary goal of our model is to understand how chronic infections can affect pathogen evolutionary dynamics over long times. Using transmission effects of mutations inferred from SARS-CoV-2 data[30], we show that bursts of mutations like those observed during the pandemic can occur even with a simple, additive fitness landscape. In particular, we explore how the within-host mutation rate, typical duration of infection, and fraction of infections that are chronic affect the likelihood of mutational bursts. We find that bursty evolution is especially likely when the acute infection time is short compared to the duration of chronic infections. Our results highlight scenarios in which chronic infections produce evolutionary dynamics that are qualitatively different from those that are observed in most simple evolutionary models.

Results

Model of pathogen evolution within and between hosts

The global evolutionary dynamics of pathogens such as SARS-CoV-2 are a consequence of processes that occur within and between infected individuals. Evolution within individuals generates a genetically diverse cloud or "quasispecies" of variants[31–33]. Differential transmission of variants between hosts ultimately results in pathogen evolution across individuals. We include both levels of evolution in our model (**Fig. 1**).

To model the emergence and accumulation of mutations within each host, we use a standard, stochastic Wright-Fisher model[34]. We assume that the pathogen population begins with a single starting genotype – consistent with tight transmission bottlenecks – and evolves in discrete generations subject to selection, mutation, and genetic drift. In each replication cycle, neutral and positively selected mutations are randomly introduced with rates μ_N and μ_B, respectively. These mutation rates represent combinations of the basic probability per replication cycle that a new mutation is introduced and the probability that that mutation is beneficial or neutral. We assume that significantly deleterious mutations are rare enough to be efficiently eliminated by selection, and do not model them explicitly.

The distribution of fitness effects of beneficial and neutral mutations was derived from selection coefficients learned from SARS-CoV-2 temporal genomic data[30] (see **Supplementary Fig. 1**). In our model, the fitness effects of mutations are additive, so that the net increase or decrease in fitness s_a for a virus a with n mutations is

$$s_a = \sum_{i=1}^{n} s_i \,. \tag{1}$$

Here, the s_i are the selection coefficients that quantify the fitness effect of each mutation i. A positive selection coefficient indicates a beneficial mutation that increases fitness, while a negative coefficient indicates a deleterious mutation.

We assume that mutations that improve viral replication within a host also improve transmission between individuals. In principle, the effects of mutations on replicative fitness and transmission fitness can be different[35]. As an example, some immune escape mutations generated during HIV-1 infection can be deleterious in other contexts, causing them to revert when the virus is transmitted to a new host[36–40]. However, within-host mutations that produce fitness gains for replication and increase viral load can contribute to increased transmission, as has been shown for Spike mutations in SARS-CoV-2 (refs.[41–43]). Furthermore, VOC mutations have been observed within individuals, including adaptive mutations concentrated in the Spike protein's receptor binding domain and N-terminal domains[25,26]. Despite these complications, we have aligned selection pressures within and between hosts for simplicity. Even in this simple case, complex evolutionary dynamics can occur.

We model transmission between individual donors and receptors of infection using a branching process that considers superspreading. In our model, the number of secondary infections is drawn from a negative binomial distribution $P_{NB}(k, k/(k+R^i))$, with k the dispersion parameter and $R^i = \bar{R}\left(1+\langle f \rangle^i\right)$ the effective reproductive number associated to the donor i. The negative binomial distribution has been used to model superspreading in past studies of viruses such as SARS and SARS-CoV-2 (refs.[44–48]). Here, $\bar{R}$ and $\langle f \rangle^i$ are the average baseline (reference) reproductive number and average fitness of the virus population from the donor host, respectively. As soon as infection is transmitted, donors are removed from the population. New infections are established by a single, randomly selected pathogen from the donor. Thus, most of the variant diversity previously generated is lost. This mimics the characteristic narrow transmission bottleneck observed in some pathogens, including SARS-CoV-2 (refs.[7,8]).

The time between when an individual is first infected and when the infection is transmitted to a new host, which we refer to as the generation time, constrains the level of viral genetic diversity that can accumulate and be transmitted. The generation time varies based on the nature of the infection. Most infections are acute and cleared by the immune system in a short period, parameterized by t_a. For SARS-CoV-2, we assume that two rounds of viral replication occur over each of the t_a days of infection[49]. In immunocompromised hosts, infections can last far longer, with a generation time $t_c \gg t_a$. A longer generation time allows for more rounds of viral replication, facilitating the accumulation of genetic diversity. Each time an infection is transmitted, we take the probability that the new host develops a chronic infection to be $p_c \ll 1$. The probability that a new infection is of short duration (acute) is then $1 - p_c$.

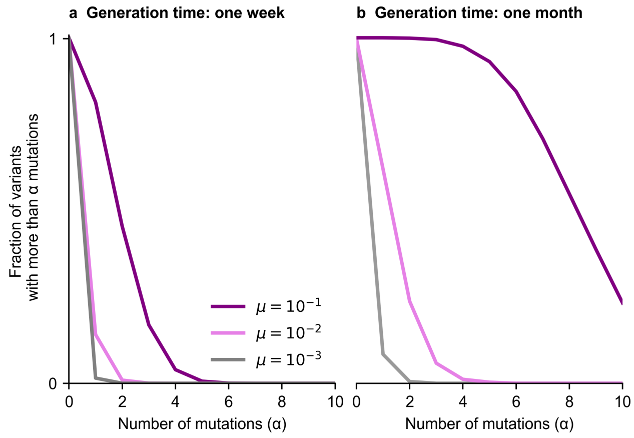


Fig. 2. Genetic diversity after typical generation times of acute and chronic infections. **a**, Fraction of variants in the intra-host viral population that acquires more than α mutations over one week of infection. **b**, distribution of accumulated mutations after one month of infection. Higher mutation rates lead to the accumulation of more mutations. We consider the same rate, μ, for beneficial or neutral mutations.

Simulating pathogen evolution

We simulated multiple realizations of the evolutionary model over 1000 days. At each simulation time, we recorded the average number of mutations in transmitted variants (variants randomly sampled from within-host populations that are transmitted in new infections) and the number of chronically infected individuals across individual populations. Generation times for acute cases were set between 2 and 9 days, covering the values estimated from known infector-infectee transmission pair data or household data across different continents[50,51]. For chronic cases, where generation times are less well determined, we sampled them from a log-normal distribution with mean $\mu_L = 150$ days and standard deviation $\sigma_L = 80$ days. We maintained a fixed neutral mutation supply rate of $\mu_N = 10^{-4}$ mutations/cycle, based on the underlying mutation rate of SARS-CoV-2 (ref.[49]) times the fraction of nonsynonymous mutations found neutral in the selection coefficient estimate from SARS-CoV-2 time series data[30] (Methods). We varied the beneficial mutation supply rate to explore its effect on pathogen evolutionary dynamics.

Patterns of mutation accumulation

Within infected individuals, mutations accumulate progressively in viral populations over time (**Fig. 2**). Higher mutation rates naturally lead to a more rapid accumulation of mutations. Longer generation times (i.e., more generations of within-host evolution) also allow for more genetic diversity to accumulate within individuals, which can then potentially be transmitted to new hosts.

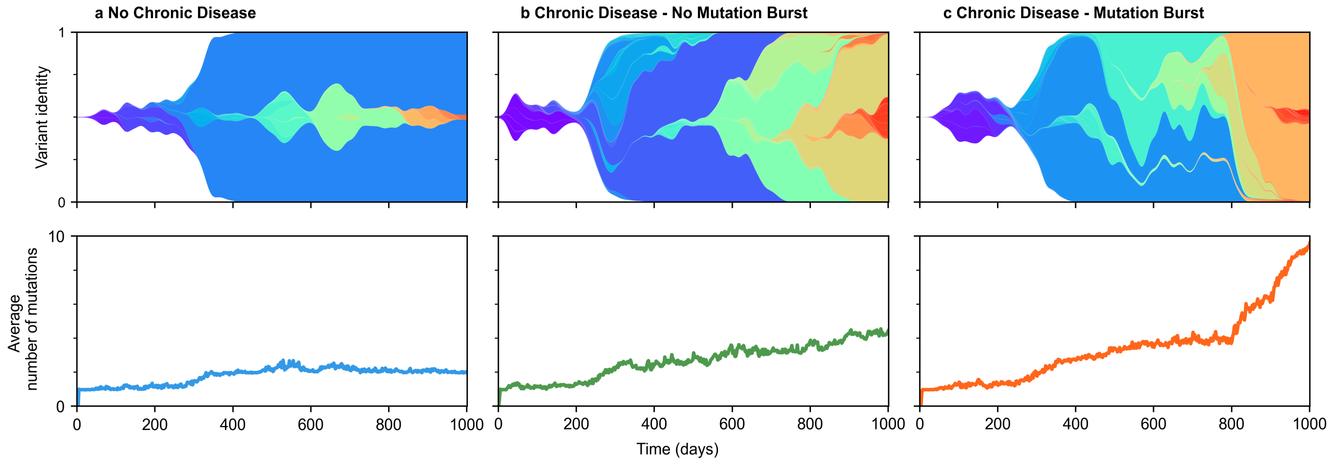


Fig. 3. Evolution of viral variants under different scenarios. a, Dynamic evolution of variants within a population exclusively composed of acute cases. **b**, Population consisting of both acute and chronic cases but without mutation burst. **c**, Population consisting of both acute and chronic cases with one mutation burst at t=800 days. For all different simulations we consider beneficial and neutral mutation rates $\mu_B = 10^{-3}$ mutations/cycle and $\mu_N = 10^{-4}$ mutations/cycle respectively. For **b** and **c**, chronic cases are included in the percentage per transmission event $p_c = 10^{-3}$ and generation times t_c are drawn from a log-normal distribution with mean $\mu_L = 150$ days and standard deviation $\sigma_L = 80$ days.

Across infected individuals, we found that the evolution of viral populations fell into roughly three patterns (**Fig. 3**). In cases where there are few or no chronic infections, we observe few viral mutations (**Fig. 3a**). Single viral lineages tend to dominate the viral population with little competition between them.

When a significant number of chronic infections occur, two distinct outcomes are possible. In one case, the accumulation of mutations in viral populations accelerates and there is significant competition between viral lineages, but the increase in mutations over time remains roughly linear (**Fig. 3b**). In other simulations, we observe sudden "bursts" of mutations in viral populations, reminiscent of the emergence of SARS-CoV-2 VOCs (**Fig. 3c**).

Phase diagram for mutational bursts

To explore the relationship between parameter space and the emergence of mutational bursts, we generated a "phase diagram" of the number of mutational bursts per chronic disease case as a function of the model parameters (**Fig. 4**). To classify bursts versus linear accumulation of mutations, we first determined the distribution of maximum mutation accumulation rates across individuals in simulations without any chronic infections (Methods). We then identified bursts as events in which the rate of mutation accumulation was 3.5 or more standard deviations greater than the average maximum mutation accumulation rate in the simulations with only acute infections. We investigated a wide range of parameters, varying the fraction of chronic infections p_c, acute generation times t_a, and rate of beneficial mutations μ_B (see Methods). For each choice of parameters, we computed the number of mutational bursts per chronic disease case over 1000 simulations.

We found several factors that facilitated the emergence of mutational bursts (**Fig. 4**). Intuitively, bursts occurred more frequently when beneficial mutation rates were higher (see analogous heatmap in **Supplementary Fig. 3** for a lower mutation rate). We also found that bursts occurred more frequently when the acute generation time t_a was shorter. Longer acute generation times lead to greater similarity in the viral populations in acute and chronically infected individuals, homogenizing the accumulation of mutations and decreasing the likelihood of abrupt increases in mutations.

Interestingly, we found that the likelihood of mutational bursts depends nonlinearly on the fraction of chronic infections. As the fraction of chronic infections increases, new adaptive mutations are generated more frequently and spread throughout the population, making isolated bursts unlikely. At very high frequencies of chronic infections, several pathogen variants with many mutations can be produced simultaneously. These variants then compete among hosts, reducing several potential bursts to a single one (see **Supplementary Fig. 4**).

Effects of chronic infections on evolutionary rate

A recent study found that the evolutionary rate of SARS-CoV-2 within a chronically infected individual was higher than the estimated global evolutionary rate of the virus, measured by the rate of substitutions over time[25]. This can be attributed, in principle, to the absence of stringent bottlenecks imposed by transmission events. In our simulations, we observed that mutational bursts can occur due to the spread of new pathogen variants that evolved for long times within chronically infected individuals. Do chronic infections affect the overall evolutionary rate even in the absence of bursts?

To answer this question, we quantified the rate of accumulation of mutations across individuals over time in different scenarios (**Fig. 5**). Specifically, we measured the evolutionary rate within chronically infected individuals and the evolutionary rate between individuals in three different cases: in simulations with no chronic infections ($p_c = 0$), with chronic infections ($p_c > 0$) but without any observed mutational burst, and with chronic infections and at least one observed burst. Each measurement was averaged over 1000

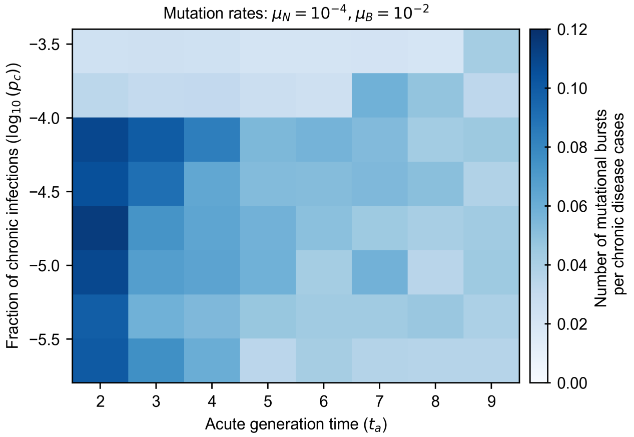


Fig. 4. Number of mutational bursts per chronic disease case for beneficial mutation rate of 10^{-2} mutations/cycle. The frequency of mutational bursts decreases as acute generation times increase because chronic and acute subpopulations exhibit greater similarity, leading to homogenization within individual populations and reducing the likelihood of abrupt mutational events. The dependence on the fraction of chronic infections is nonlinear: while moderate levels of chronic infections lead to more frequent bursts, very high levels cause competition among multiple pathogen variants, each with numerous mutations, reducing the occurrence of isolated bursts. Each value represents an average over 1000 simulations.

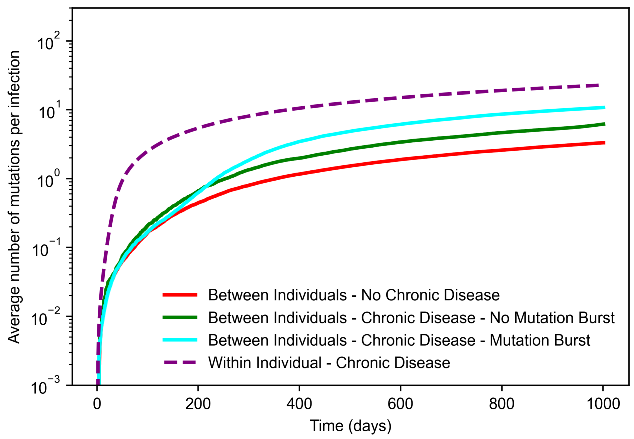


Fig. 5. Average number of accumulated mutations per infection in simulations of within-host evolution and between-host evolution with and without chronic infections. For each observation time, the reported value represents the average number of mutations within intra-host viral populations of actively infected individuals, normalized by the total number of infected individuals at that time. The higher evolutionary rate is obtained within a chronically infected individual due to the absence of stringent bottlenecks imposed by transmission events. Between-host evolution with chronic infections, even in the absence of a burst, leads to an increased rate of mutation accumulation compared to populations with only acute infections. The simulations were conducted with a beneficial mutation rate of $\mu_B = 10^{-3}$ mutations/cycle, an acute generation time of $t_a = 2$ days, and a probability of new chronic infection of $p_c = 4 \times 10^{-4}$. Each curve represents an average of over 1000 simulations.

simulations. Our results align with clinical data. Namely, the evolutionary rate within a single infected individual was higher than across the population of infected individuals in all cases.

As expected, we found that the evolutionary rate between individuals was highest in populations with chronic infections and where at least one mutational burst was observed. However, even in the absence of a burst, the presence of chronic infections still leads to an increased rate of mutation accumulation compared to populations with only acute infections. Thus, chronic infections could still accelerate pathogen evolution through the generation and transmission of adaptive mutations, even without the production of SARS-CoV-2 VOC-like variants.

Discussion

In this work, we modeled pathogen evolution within and between hosts including different types of infections: acute, short-term infections and rare chronic ones. The goal of our study was to understand how chronic infections can influence pathogen evolution over long times. Even with a simple, additive fitness landscape, we found that chronic infections can lead to SARS-CoV-2-like evolutionary dynamics, as mutants with multiple novel mutations arise and spread through the population. Such "bursts" of mutations were especially likely when acute generation times were short.

We found that the frequency of chronic infections had a strong and nonlinear effect on the frequency of mutational bursts. When chronic infections were rare, the number of observed mutational bursts scaled roughly linearly with the frequency of chronic infections. However, frequent chronic infections result in the generation and transmission of more adaptive mutations. Mutants with different beneficial mutations compete for hosts, making it more difficult for a single, dominant variant to quickly emerge.

Our model employs several simplifying assumptions that could be revisited in future work. First, we assumed that the fitness effects of pathogen mutations within and between hosts were the same. While certain mutations, such as those that increase viral load, are highly likely to improve both within-host replication and transmission between individuals, others may only be advantageous in particular scenarios. The distribution of fitness effects of mutations is also challenging to determine. Here, we used data from a recent study of SARS-CoV-2 evolution to parameterize our model[30]. While the fitness effects of mutations in this study have extensive experimental support, they are subject to noise, and they were determined solely from between-host transmission rather than within-host replication. We have also assumed that the fitness effects of mutations are the same across hosts. Experiments[52,53] and computational analyses[54] have found many similarities between the fitness effects of mutations for genetically similar viruses, but some differences between hosts would be expected in real scenarios.

The model that we have developed is a type of "metapopulation" model[55], considering evolution both within and between hosts. Past studies have used such models to infer epidemiological dynamics[56] and explain phylogenetic structure[57,58], among other applications[59]. Our work contributes to this area by exploring how different types of infections (i.e., acute versus chronic infection) contribute to pathogen evolutionary rates.

ACKNOWLEDGEMENTS

The work reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Numbers R35GM138233 and R35GM136381. A.R.D. acknowledges support from the National Science Foundation through the Physics Frontier Center for Living Systems (PHY-2317138).

AUTHOR CONTRIBUTIONS

All authors contributed to methods development, data analysis, interpretation of results, and writing the paper. E.R.-H. and J.S. led computational analyses. A.R.D. and J.P.B. supervised the project.

References

1. Mascola, J. R., Graham, B. S. & Fauci, A. S. SARS-CoV-2 Viral Variants—Tackling a Moving Target. *JAMA* **325**, 1261–1262 (2021). URL https://doi.org/10.1001/jama.2021.2088. https://jamanetwork.com/journals/jama/articlepdf/2776542/jama_mascola_2021_ed_210014_1617663063.80169.pdf.
2. Peacock, T. P., Penrice-Randal, R., Hiscox, J. A. & Barclay, W. S. Sars-cov-2 one year on: evidence for ongoing viral adaptation. *J. Gen. Virol.* **102** (2021). URL https://doi.org/10.1099/jgv.0.001584.
3. Harvey, W. T. *et al.* Sars-cov-2 variants, spike mutations and immune escape. *Nature Reviews Microbiology* **19**, 409–424 (2021). URL https://doi.org/10.1038/s41579-021-00573-0.
4. Tabatabai, M. *et al.* An analysis of covid-19 mortality during the dominancy of alpha, delta, and omicron in the usa. *Journal of Primary Care & Community Health* **14**, 21501319231170164 (2023). URL https://doi.org/10.1177/21501319231170164. PMID: 37083205, https://doi.org/10.1177/21501319231170164.
5. Salehi-Vaziri, M. *et al.* The ins and outs of sars-cov-2 variants of concern (vocs). *Archives of Virology* **167**, 327–344 (2022). URL https://doi.org/10.1007/s00705-022-05365-2.
6. Carabelli, A. M. *et al.* Sars-cov-2 variant biology: immune escape, transmission and fitness. *Nature Reviews Microbiology* **21**, 162–177 (2023). URL https://doi.org/10.1038/s41579-022-00841-7.
7. Braun, K. M. *et al.* Acute sars-cov-2 infections harbor limited within-host diversity and transmit via tight transmission bottlenecks. *PLOS Pathogens* **17**, 1–26 (2021). URL https://doi.org/10.1371/journal.ppat.1009849.
8. Bendall, E. E. *et al.* Rapid transmission and tight bottlenecks constrain the evolution of highly transmissible sars-cov-2 variants. *Nature Communications* **14**, 272 (2023). URL https://doi.org/10.1038/s41467-023-36001-5.
9. Markov, P. V. *et al.* The evolution of SARS-CoV-2. *Nature Reviews Microbiology* **21**, 361–379 (2023). URL https://doi.org/10.1038%2Fs41579-023-00878-2.
10. Davis, J. T. *et al.* Cryptic transmission of sars-cov-2 and the first covid-19 wave. *Nature* **600**, 127–132 (2021). URL https://doi.org/10.1038/s41586-021-04130-w.
11. Wilkinson, E. *et al.* A year of genomic surveillance reveals how the sars-cov-2 pandemic unfolded in africa. *Science* **374**, 423–431 (2021). URL https://www.science.org/doi/abs/10.1126/science.abj4336. https://www.science.org/doi/pdf/10.1126/science.abj4336.
12. Adepoju, P. Challenges of sars-cov-2 genomic surveillance in africa. *The Lancet Microbe* **2**, e139 (2021). URL https://www.sciencedirect.com/science/article/pii/S2666524721000653.
13. Lu, L. *et al.* Adaptation, spread and transmission of sars-cov-2 in farmed minks and associated humans in the netherlands. *Nature Communications* **12**, 6802 (2021). URL https://doi.org/10.1038/s41467-021-27096-9.
14. Bashor, L. *et al.* Sars-cov-2 evolution in animals suggests mechanisms for rapid variant selection. *Proceedings of the National Academy of Sciences* **118**, e2105253118 (2021). URL https://www.pnas.org/doi/abs/10.1073/pnas.2105253118. https://www.pnas.org/doi/pdf/10.1073/pnas.2105253118.
15. Hale, V. L. *et al.* Sars-cov-2 infection in free-ranging white-tailed deer. *Nature* **602**, 481–486 (2022). URL https://doi.org/10.1038/s41586-021-04353-x.
16. Pickering, B. *et al.* Divergent sars-cov-2 variant emerges in white-tailed deer with deer-to-human transmission. *Nature Microbiology* **7**, 2011–2024 (2022). URL https://doi.org/10.1038/s41564-022-01268-9.
17. Smith, C. A. & Ashby, B. Antigenic evolution of SARS-CoV-2 in immunocompromised hosts. *Evolution, Medicine, and Public Health* **11**, 90–100 (2022). URL https://doi.org/10.1093/emph/eoac037. https://academic.oup.com/emph/article-pdf/11/1/90/49701865/eoac037.pdf.
18. Ghafari, M., Liu, Q., Dhillon, A., Katzourakis, A. & Weissman, D. B. Investigating the evolutionary origins of the first three sars-cov-2 variants of concern. *Frontiers in Virology* **2** (2022).
19. Martin, D. P. *et al.* Selection Analysis Identifies Clusters of Unusual Mutational Changes in Omicron Lineage BA.1 That Likely Impact Spike Function. *Molecular Biology and Evolution* **39**, msac061 (2022). URL https://doi.org/10.1093/molbev/msac061. https://academic.oup.com/mbe/article-pdf/39/4/msac061/43374245/msac061.pdf.
20. Zahradník, J. *et al.* Sars-cov-2 variant prediction and antiviral drug design are enabled by rbd in vitro evolution. *Nature Microbiology* **6**, 1188–1198 (2021). URL https://doi.org/10.1038/s41564-021-00954-4.
21. Choi, B. *et al.* Persistence and evolution of sars-cov-2 in an immunocompromised host. *New England Journal of Medicine* **383**, 2291–2293 (2020). URL https://doi.org/10.1056/NEJMc2031364. PMID: 33176080, https://doi.org/10.1056/NEJMc2031364.
22. Avanzato, V. A. *et al.* Case study: Prolonged infectious sars-cov-2 shedding from an asymptomatic immunocompromised individual with cancer. *Cell* **183**, 1901–1912.e9 (2020). URL https://www.sciencedirect.com/science/article/pii/S0092867420314562.
23. Clark, S. A. *et al.* Sars-cov-2 evolution in an immunocompromised host reveals shared neutralization escape mechanisms. *Cell* **184**, 2605–2617.e18 (2021). URL https://doi.org/10.1016/j.cell.2021.03.027.
24. Lee, C. Y. *et al.* Prolonged SARS-CoV-2 Infection in Patients with Lymphoid Malignancies. *Cancer Discovery* **12**, 62–73 (2022). URL https://doi.org/10.1158/2159-8290.CD-21-1033. https://aacrjournals.org/cancerdiscovery/article-pdf/12/1/62/3194898/62.pdf.
25. Chaguza, C. *et al.* Accelerated SARS-CoV-2 intrahost evolution leading to distinct genotypes during chronic infection. *Cell Rep. Med.* **4**, 100943 (2023).
26. Wilkinson, S. A. J. *et al.* Recurrent SARS-CoV-2 mutations in immunodeficient patients. *Virus Evolution* **8**, veac050 (2022). URL https://doi.org/10.1093/ve/veac050. https://academic.oup.com/ve/article-pdf/8/2/veac050/48422645/veac050.pdf.
27. Kemp, S. A. *et al.* SARS-CoV-2 evolution during treatment of chronic infection. *Nature* **592**, 277–282 (2021). URL https://doi.org/10.1038/s41586-021-03291-y.
28. R, K. & A, S. Antigenic escape is accelerated by the presence of immunocompromised hosts. *Proceedings of the Royal Society B: Biological Sciences* **289** (2022).
29. Van Egeren, D. *et al.* Controlling long-term sars-cov-2 infections can slow viral evolution and reduce the risk of treatment failure. *Scientific Reports* **11**, 22630 (2021). URL https://doi.org/10.1038/s41598-021-02148-8.
30. Lee, B. *et al.* Inferring effects of mutations on sars-cov-2 transmission from genomic surveillance data. *medRxiv* (2022). URL https://www.medrxiv.org/content/early/2022/01/14/2021.12.31.21268591. https://www.medrxiv.org/content/early/2022/01/14/2021.12.31.21268591.full.pdf.
31. Eigen, M., McCaskill, J. & Schuster, P. Molecular quasi-species. *The Journal of Physical Chemistry* **92**, 6881–6891 (1988).
32. Coffin, J. M. Hiv population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy. *Science* **267**, 483–489 (1995).
33. Domingo, E., Sheldon, J. & Perales, C. Viral quasispecies evolution. *Microbiology and Molecular Biology Reviews* **76**, 159–216 (2012).
34. Ewens, W. *Mathematical Population Genetics 1: Theoretical Introduction* (Springer Science and Business Media, 2012).
35. Wargo, A. R. & Kurath, G. Viral fitness: definitions, measurement, and current insights. *Current Opinion in Virology* **2**, 538–545 (2012). URL https://www.sciencedirect.com/science/article/pii/S1879625712001290. Virus evolution / Antivirals and resistance.
36. Allen, T. M. *et al.* Selection, transmission, and reversion of an antigen-processing cytotoxic t-lymphocyte escape mutation in human immunodeficiency virus type 1 infection. *Journal of virology* **78**, 7069–7078 (2004).
37. Leslie, A. *et al.* Hiv evolution: Ctl escape mutation and reversion after transmission. *Nature medicine* **10**, 282–289 (2004).
38. Friedrich, T. C. *et al.* Reversion of ctl escape–variant immunodeficiency viruses in vivo. *Nature medicine* **10**, 275–281 (2004).
39. Zanini, F. *et al.* Population genomics of intrapatient hiv-1 evolution. *Elife* **4**, e11282 (2015).
40. Gao, Y. & Barton, J. P. A binary trait model reveals the fitness effects of hiv-1 escape from t cell responses. *bioRxiv* (2024).
41. Korber, B. *et al.* Tracking changes in sars-cov-2 spike: Evidence that d614g increases infectivity of the covid-19 virus. *Cell* **182**, 812–827.e19 (2020). URL https://www.sciencedirect.com/science/article/pii/S0092867420308205.
42. Yurkovetskiy, L. *et al.* Structural and functional analysis of the d614g sars-cov-2 spike protein variant. *Cell* **183**, 739–751.e8 (2020). URL https://www.sciencedirect.com/science/article/pii/S0092867420312290.
43. Liu, Y. *et al.* The n501y spike substitution enhances sars-cov-2 infection and transmission. *Nature* **602**, 294–299 (2022). URL https://doi.org/10.1038/s41586-021-04245-0.
44. Irwin, J. A distribution arising in the study of infectious diseases. *Biometrika* **41**, 266–268 (1954).
45. Griffiths, D. Maximum likelihood estimation for the beta-binomial distribution and an application to the household distribution of the total number of cases of a disease. *Biometrics* **29**, 637–648 (1973).
46. Lipsitch, M. *et al.* Transmission dynamics and control of severe acute respiratory syndrome. *Science* **300**, 1966–1970 (2003).
47. Lloyd-Smith, J. O., Schreiber, S. J., Kopp, P. E. & Getz, W. M. Superspreading and the effect of individual variation on disease emergence. *Nature* **438**, 355–359 (2005).
48. Althouse, B. M. *et al.* Superspreading events in the transmission dynamics of SARS-CoV-2: Opportunities for interventions and control. *PLOS Biology* **18**, 1–13 (2020).
49. Bar-On, Y. M., Flamholz, A., Phillips, R. & Milo, R. Science forum: Sars-cov-2 (covid-19) by the numbers. *eLife* **9**, e57309 (2020). URL https://doi.org/10.7554/eLife.57309.
50. Chen, D. *et al.* Inferring time-varying generation time, serial interval, and incubation period distributions for covid-19. *Nature Communications* **13**, 7727 (2022). URL https://doi.org/10.1038/s41467-022-35496-8.
51. Hart, W. S. *et al.* Inference of the sars-cov-2 generation time using uk house-

hold data. *eLife* **11**, e70767 (2022). URL https://doi.org/10.7554/eLife.70767.

52. Doud, M. B., Ashenberg, O. & Bloom, J. D. Site-specific amino acid preferences are mostly conserved in two closely related protein homologs. *Molecular biology and evolution* **32**, 2944–2960 (2015).
53. Haddox, H. K., Dingens, A. S., Hilton, S. K., Overbaugh, J. & Bloom, J. D. Mapping mutational effects along the evolutionary landscape of hiv envelope. *Elife* **7**, e34420 (2018).
54. Shimagaki, K. S., Lynch, R. M. & Barton, J. P. Parallel hiv-1 evolutionary dynamics in humans and rhesus macaques who develop broadly neutralizing antibodies. *bioRxiv* (2024).
55. Ball, F. *et al.* Seven challenges for metapopulation models of epidemics, including households models. *Epidemics* **10**, 63–67 (2015).
56. Stadler, T. *et al.* Estimating the basic reproductive number from viral sequence data. *Molecular biology and evolution* **29**, 347–357 (2012).
57. Volz, E. M., Koopman, J. S., Ward, M. J., Brown, A. L. & Frost, S. D. Simple epidemiological dynamics explain phylogenetic clustering of hiv from patients with recent infection. *PLoS computational biology* **8**, e1002552 (2012).
58. Volz, E. M., Koelle, K. & Bedford, T. Viral phylodynamics. *PLoS computational biology* **9**, e1002947 (2013).
59. Frost, S. D. *et al.* Eight challenges in phylodynamic inference. *Epidemics* **10**, 88–92 (2015).

Methods

Global evolutionary model

Within-host virus evolution model

The viral population, consisting of N infected cells, begins with a single starting genotype and undergoes discrete and non-overlapping Wright-Fisher generations subject to selection, mutation, and genetic drift. For each replication cycle, neutral and positive selected mutations are randomly introduced from a binomial distribution with rates $\{\mu_N, \mu_B\}$ respectively. Once a mutation is generated for the genotype a, its selective effect s_a is drawn from distributions derived from experimental data. The genotype's fitness is then updated as $f_a = f^{wt} + s_a$, where f^{wt} represents the fitness of the wild-type genotype. Selected species in the subsequent generation are given by binomial sampling with success probability defined by their genotype fitness as $p_a = f_a / \sum_b f_b$.

Between-host virus evolution model

In our model, new infections are established by one virion from the infection donor, and most of the variant diversity previously generated is lost. This mimics the characteristic narrow transmission bottleneck observed in respiratory viruses like SARS-CoV-2. The variant passed from the infection donor to an infected individual is randomly selected from the intrahost virus population. Subsequently, variants with higher fitness can persist through the transmission bottleneck only if there is sufficient intrahost evolution time to elevate their frequency. However, the most common scenario (with acute infections) is that variants without strong selective advantage overcome the transmission bottleneck by chance, a phenomenon of genetic drift.

Another relevant feature that our model incorporates is superspreading, where a small fraction of infectious hosts are responsible for most transmissions. For this, the number of secondary infections caused by an infected individual, at its generation period, is drawn from a negative binomial distribution $P_{NB}(k, k/(k+R^i))$, where k is the dispersion parameter and $R^i = \bar{R}\left(1 + \langle f \rangle^i\right)$ represents the effective reproductive number associated to the host i. This is dependent on the average reproductive number ($\bar{R}$) and the average fitness of the virus population from the donor host $\langle f \rangle^i = \sum_{a=1}^{N} p_a f_a^i$, where p_a is the variant frequency. As soon as the infection occurs, donors are removed from the population, either due to death or immunity.

Dynamic simulations

We implemented the global evolutionary model in Julia. For the intra-host viral population size $N = 1000$ (ref.[60]), we ran multiple independent realizations of the evolutionary model over 1000 days. The distribution of both neutral and beneficial mutation effects used during within-host replication cycles was fitted from the distribution of selection coefficients learned from SARS-CoV-2 temporal genomic data[61], as shown in **Supplementary Fig. 1**. To model generation times for chronic cases, we assume a log-normal distribution with mean $\mu_L = 150$ days and standard deviation $\sigma_L = 80$ days. We select $k = 1.0$ as the dispersion parameter in the negative binomial distribution, which is within the estimated range for SARS-CoV-2[62]. We choose this moderate value to reflect that, in our model, transmission heterogeneity is not solely driven by k, as we consider an effective reproductive number R^i dependent on viral fitness in the donor, which also contributes to the non-homogeneous spread of secondary infections.

Burst detection method

To identify bursts of mutations along the trajectories of accumulated mutations, we follow a step-by-step process. Initially, we calculate the slope at each time point for M trajectories obtained from simulations involving only acute cases. Subsequently, we apply a Savitzky–Golay filter[63] with a time window length (w) and polynomial order (p) to smooth the slope time series for each simulation. We extract the maximum values from the smoothed slope time series and use them to build a Gaussian null distribution (see **Supplementary Fig. 2a**).

We use the same smoothing process for the slope time series of simulations involving chronic disease cases. We then calculate the z-score for each time point using the mean and standard deviation obtained from the previously established null distribution. Mutation bursts are identified as outliers in this null distribution, defined by instances where the z-score exceeds 3.5. Given that multiple time points near the jump meet this criterion, we identify change points in the z-score time series. These change points delineate the start and end times of each jump, with the midpoint representing the burst time. The entire procedure is summarized in **Supplementary Fig. 2b**.

Data and code

The data and code used in our analysis can be accessed from the GitHub repository at https://github.com/bartonlab/paper-SARS-CoV-2-evolution. This repository also contains the code used to analyze data and generate the figures presented in this paper.

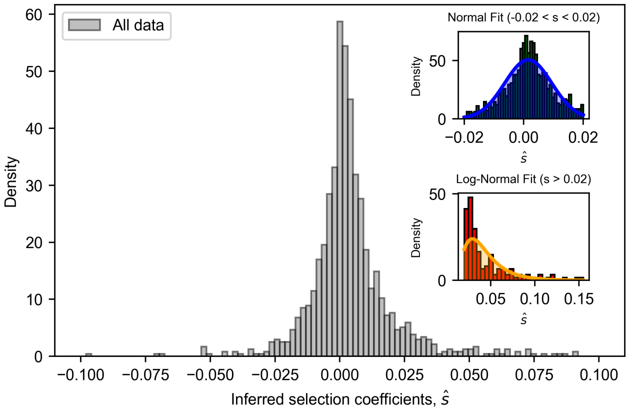
References

60. Bar-On, Y. M., Flamholz, A., Phillips, R. & Milo, R. Science forum: Sars-cov-2 (covid-19) by the numbers. *eLife* **9**, e57309 (2020). URL https://doi.org/10.7554/eLife.57309.

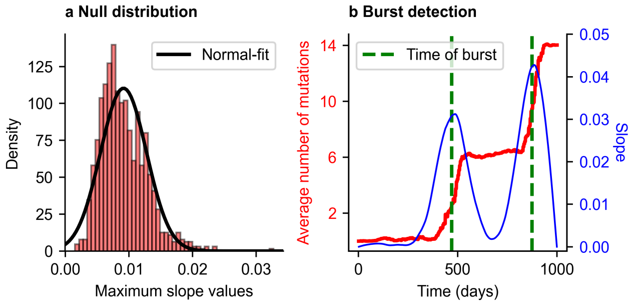
61. Lee, B. *et al.* Inferring effects of mutations on sars-cov-2 transmission from genomic surveillance data. *medRxiv* (2022). URL https://www.medrxiv.org/content/early/2022/01/14/2021.12.31.21268591. https://www.medrxiv.org/content/early/2022/01/14/2021.12.31.21268591.full.pdf.

62. Wegehaupt, O., Endo, A. & Vassall, A. Superspreading, overdispersion and their implications in the sars-cov-2 (covid-19) pandemic: a systematic review and meta-analysis of the literature. *BMC Public Health* **23**, 1003 (2023). URL https://doi.org/10.1186/s12889-023-15915-1.

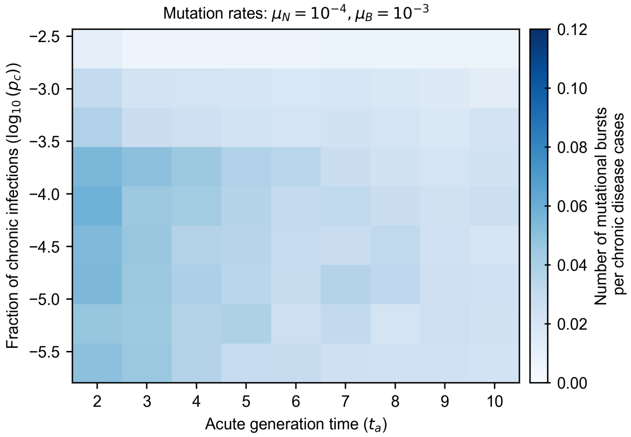
63. Savitzky, A. & Golay, M. J. E. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry* **36**, 1627–1639 (1964). URL https://doi.org/10.1021/ac60214a047.



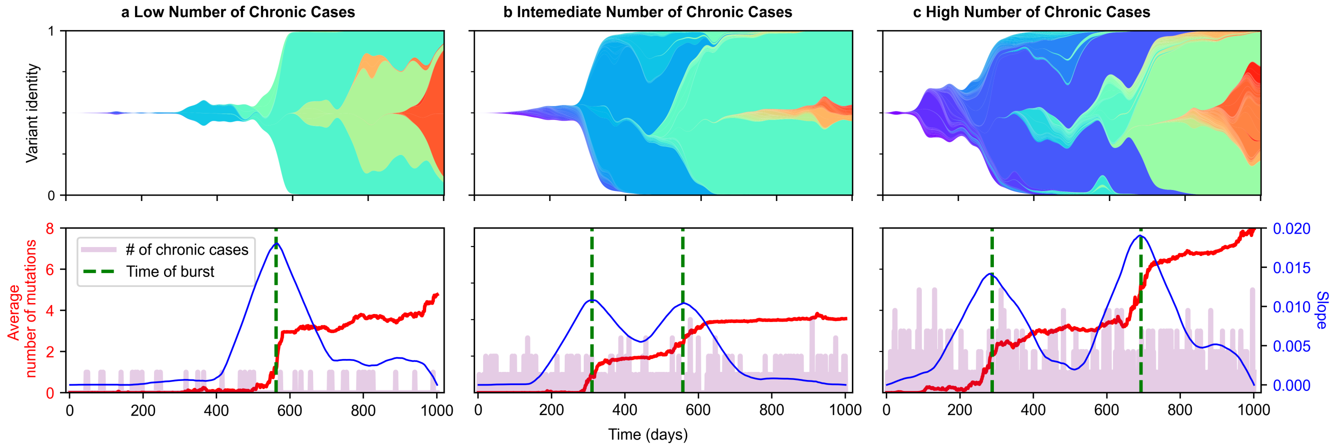
Supplementary Fig. 1. Inferred transmission effects of SARS-CoV-2 mutations. The main plot displays a histogram of selection coefficient values inferred from SARS-CoV-2 temporal genomic data[61]. The top-right inset plot shows the normal distribution fit for coefficient values considered neutral $(-0.02 < \hat{s} < 0.02)$; from this distribution, neutral mutation effects during within-host evolution were sampled. The bottom-right inset plot shows the log-normal distribution fit for values greater than 0.02, representing significantly beneficial mutations; from this distribution, beneficial mutation effects during within-host evolution were sampled.



Supplementary Fig. 2. Detection of mutation bursts. a, Distribution of maximum slopes of accumulated mutation trajectories without chronic infection for acute generation time of $t_a = 4.0$ days and mutation rates: $\mu_B = 10^{-3}$ beneficial mutations per cycle and $\mu_N = 10^{-4}$ neutral mutations per cycle. **b,** For a simulation with chronic infection fraction $p_c = 10^{-4}$, the number of accumulated mutations averaged over an individual's population is shown in red. The blue curve indicates the smoothed slope time series with two peaks, detected by z-score time series change points and represented by the vertical green dashed lines. For smoothing using the Savitzky–Golay filter, we use parameters $w = 150$ and $p = 1$.



Supplementary Fig. 3. Number of mutational bursts per chronic disease case for beneficial mutation rate of 10^{-3} mutations/cycle. This figure is analogous to **Fig. 4** in the main text, but with a lower beneficial mutation rate.



Supplementary Fig. 4. Dynamic evolution of the viral population under varying chronic infection probabilities. **a**, Low number of chronic cases, corresponding to a probability per transmission event of $p_c = 4 \times 10^{-4}$. **b** Intermediate number of chronic cases, with a probability per transmission event of $p_c = 3.7 \times 10^{-3}$. **c**, High number of chronic cases, resulting from a probability per transmission event of $p_c = 7.0 \times 10^{-3}$. For all simulations, we consider beneficial and neutral mutation rates $\mu_B = 10^{-4}$ mutations/cycle and $\mu_N = 10^{-4}$ mutations/cycle, respectively. Generation times are set at $t_a = 2$ for acute cases, while for chronic cases, they follow a log-normal distribution with a mean of $\mu_L = 150$ days and a standard deviation of $\sigma_L = 80$ days.