Distribution Agreement

Signature:

______________________________ __________

Qihan Liu Date

The effect of population structure and the mode of selection on multi-locus adaptation

By

Qihan Liu
Doctor of Philosophy

Physics

Daniel B. Weissman, Ph.D.
Advisor

Stefan Boettcher, Ph.D.
Committee Member

Katia Koelle, Ph.D.
Committee Member

Minsu Kim, Ph.D.
Committee Member

Ilya Nemenman, Ph.D.
Committee Member

Accepted:

Kimberly Jacob Arriola, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

The effect of population structure and the mode of selection on multi-locus adaptation

By

Qihan Liu
B.S., University of Science and Technology of China, China, 2017

Advisor: Daniel B. Weissman, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Physics
2023

Abstract

The effect of population structure and the mode of selection on multi-locus adaptation
By Qihan Liu

Evolution is influenced by many factors, for example, epistasis, sex, and mutation. This thesis investigates key factors influencing evolution, including effect of population structure and mode of selection with multiple loci. The first part explores adaptation in structured populations with clonal interference by introducing drastic population dynamics and synchronized sexual reproduction. This approach effectively utilizes genetic diversity preserved by population structure, leading to enhanced adaptation. Surprisingly, the rate of adaptation in structured populations is comparable to that in well-mixed populations, even in a consistent environment. The second part focuses on the impact of modifying the fitness function to simulate effective population structure. By imposing limits on the fittest individuals using a logistic fitness function, moderately fit individuals are promoted, resulting in increased genetic diversity and accelerated adaptation in sexual populations. This finding highlights the importance of considering the fitness distribution in shaping the adaptive process. In the third part, a simulation analysis is applied to investigate the emergence of variants of concern (VOCs) in SARS-CoV-2. A quantitative framework captures evolutionary pathways, considering both between-host transmission and within-host chronic infections. Results suggest that VOCs are primarily driven by multiple mutations from individuals with acute or chronic infections. Addressing chronic infections becomes vital in reducing future VOC emergence. Our findings have implications for optimal evolution strategies with clonal interference, evolutionary experiments, epidemic disease analysis, and future pathogen prediction.

The effect of population structure and the mode of selection on multi-locus adaptation

By

Qihan Liu
B.S., University of Science and Technology of China, China, 2017

Advisor: Daniel B. Weissman, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Physics
2023

Acknowledgments

Now I am about to submit my graduation thesis, but looking back on my journey, everything seems so dreamlike and unreal. Growing up in a village with scarce resources, I experienced numerous critical moments along the way. Looking back, these seemingly fortunate moments were actually delicate tightrope walks with no any room for error. Amidst the celebration, there is also a deep sense of helplessness when confronted with the harsh realities of life. Among my classmates, I witnessed firsthand the challenges they faced. Some had to manage their family's finances for grandparents and younger siblings, while their parents worked far away and could only visit them once a year. Others had to assist with farming chores after school, and there were even those whose homes lacked basic amenities like access to water and electricity. While my family's conditions were modest, we at least had a stable income ensuring the basic necessities. Therefore, during times of overwhelmed academic pressure and self-doubt throughout my university journey, I would always remind myself of how hard-won this opportunity was. In light of this, I want to thank my parents for providing their support to the best of their abilities, despite my mother often regrets not being able to provide a better life during my childhood. Additionally, I am profoundly grateful to my classmates who generously shared their knowledge and support with me. They, too, deserved greater opportunities.

I would like to my heartfelt thanks to my advisor, Daniel Weissman, whose guidance and encouragement have been invaluable throughout my Ph.D. journey. His mentorship extends beyond research and has positively influenced my personal life as well. Some time ago, I was asked about my opinion of Daniel by a prospective applicant. I wholeheartedly stated that if I were to go back to when I first came to Emory, I would choose him as my advisor again, as he embodies my ideal image of an advisor. Additionally, I am thankful to the members of the Weissman Group, especially Linnea Bavik and Rohan Mehta, for their academic assistance and unwavering support in all aspects of life. Their contributions have been instrumental in shaping my research and personal growth.

I would also like to express my gratitude to the professors in the Emory physics department for their guidance and expertise. I extend my appreciation to the physics department staff, particularly Barbara Conner, for their invaluable assistance throughout my academic journey.

To my friends, particularly my four-year roommates Wei Li, Weijie Li, and Jian Wang, I am deeply grateful for your unwavering emotional support. Without your presence, I doubt I would have had the strength to persevere and complete my Ph.D.

Lastly, I want to express my heartfelt gratitude to my life partner, Ton Gia Truong. Your outgoing and lively nature has brought so many interesting people and experiences into my life. Your love and warmth have illuminated my journey, and without you, my life would be devoid of joy and light.

To all those who have played a role in shaping my life and supporting me along the way, I extend my sincerest thanks. Your contributions have been immeasurable, and I am forever grateful.

Contents

List of Figures

List of Tables

Introduction

The study of population genetics has a long history of theoretical analysis on evolution under different conditions, dating back to the pioneering work of Fisher, Wright, and Haldane [37, 69, 128]. In particular, the phenotypes can be viewed as combined impacts of many independent loci, each of which contributes a small additive benefit for selection [105]. When mutation supply is low, the population accumulates mutations through sequences of fixations [63, 122]. Each mutation fixes with a probability that depends only on its own selective coefficient. However, when the mutation supply is high, it becomes more complicated. In this case, new mutations will arise before the previous ones fix. This causes multiple lineages simultaneously compete for fixation and their sweeping process would overlap and interfere with each other. During this process, strongly fit individuals would out-compete other beneficial ones, resulting in the loss of favorable mutations. This process is called clonal interference, first articulated by Muller in 1932 and statistically described by Gerrish and Lenski in 1998 [41, 86]. There has been much research on how sexual reproduction can recombine mutations into a single individual and thereby break down clonal interference [27, 86, 88, 94, 108].

In Chapter 1, we extend this research to a model of a structured population with fluctuating rates of dispersal and sexual reproduction. This kind of pattern can be induced by occasional environmental stress, including, for example, starvation or exposure to toxins [38, 60, 85, 110]. The fact that similar environmental stimuli can trigger both dispersal and recombination produces two forms of synchronization: both processes within the same individual (individual level) and one process across individuals (population level). In microbial experiments, this process is practically feasible by introducing a stressful environment during occasional mixing process [65, 83]. It has been shown that population-level synchronization of conditional sexual reproduction can affect the rate of adaptation with additive mutation [7, 64, 100, 120]. However, these studies are for well-mixed populations. It remains unclear how this population and individual synchronization would interact with the population structure. Intuitively, the population structure would always hinder the adaptation when the effect of mutations is additive, since it would sacrifice the "exploitation" of existing mutations in exchange for "exploration" in the genotype space, which is, in this case, worthless when only a single peak exists [65, 87]. However, our simulations show that structured populations with individual- and/or population-synchronized recombination and dispersal can adapt faster than well-mixed ones. The population stricture can accumulate genetic diversity in different demes, and both synchronizations, especially population synchronization, can maximize the probability of recombining migrant individuals who are most likely to contain distinct mutations. Therefore, the rate of adaptation is increased as the population dynamics increase the probability for a mutation to survive from clonal interference and fix in the population by recombination.

In Chapter 1, it is particularly surprising that a structured population without neither population nor individual synchronization can adapt as fast, if not faster, than a well-mixed one. This implies that maintaining genetic diversity could be a universal strategy to maintain the variation-selection balance. Previous studies have demonstrated the importance of this balance [52, 53, 105, 107, 115, 133]. Specifically, the dependency of adaptation speed in selection strength is not always monotonic; in fact, a relaxed selection can be optimal for the long-term response of quantitative traits. Therefore, in Chapter 2, we consider the effect of a fitness function that limits the reproductive advantage of the fittest individuals. This can be seen as an effective description of the effects of spatial structure on reducing competition, or it could be a direct description of an artificial selection scheme. We find that such fitness functions can accelerate adaptation by promoting the reproduction of moderately fit individuals that provide the genetic diversity to fuel future adaptation. Our study highlights the significance of maintaining genetic diversity in the evolution of a sexual population, especially when clonal interference is strong.

Population genetics originally arose as a tool to solve practical biological problems. In Chapter 3, we use simulations to investigate the emergence of Variants of Concern (VOCs) of SARS-CoV-2 [1, 35, 113]. During 2020, three VOCs of SARS-CoV-2, Alpha, Beta, and Gamma, which share a large number of mutations, emerged independently and almost simultaneously, later followed by Delta and Omicron in 2021 which are genetically and phenotypically distinct [79, 103, 112, 112]. Here, we provide a quantitative framework that can generate the pathogenic dynamics based on different evolutionary pathways, including within- and between-host evolution and different fitness landscapes. We investigate the likelihood of those pathways to produce dynamics similar to those observed for the VOCs of SARS-CoV-2. The simulation results imply that the VOCs are most likely to have emerged from chronic infections by accumulating multiple key mutations. This suggests that a public health strategy of finding and treating chronic SARS-CoV-2 infections could lower the probability of future VOCs.

Chapter 1

Population structure can reduce clonal interference under synchronized recombination and dispersal

Abstract

In populations with limited recombination, clonal interference among beneficial mutations limits the maximum rate of adaptation. Spatial structure slows the spread of beneficial alleles; in purely asexual populations, this increases the amount of clonal interference. Beyond this extreme case, however, it is unclear how spatial structure and recombination interact to determine the amount of clonal interference. This interaction is particularly interesting because dispersal and recombination are often at least partially synchronized in natural populations, both at the individual and population level, as when plants switch from vegetative growth to sexual reproduction or stress responses increase both motility and recombination in microbes. We simulate island models of adapting populations and find that synchronized dispersal and re-

combination allow them to adapt *faster* than matched well-mixed populations. This is because the spatial structure preserves genetic diversity and the dispersal increases the chance that recombination events occur between diverged individuals from different demes, i.e., the pairings where negative linkage disequilibrium can be most effectively reduced.

1.1 Introduction

On smooth fitness landscapes, adaptation is driven by the fixation of beneficial mutations. When beneficial mutations are rare, they can fix independently from each other in sequential selective sweeps. But if the beneficial mutation supply is large, multiple beneficial mutations will be simultaneously polymorphic in the population, and may compete with each other for fixation. This "clonal interference" effectively puts an upper limit on the rate of adaptation, particularly when recombination is limited [32, 41, 86, 88, 120]. Spatial structure increases the time it takes for a selective sweep, and therefore increases the probability that multiple beneficial mutations will coexist and interfere. In asexual populations, strong spatial structure can drastically reduce the rate of adaptation [78]. On rugged fitness landscapes, in contrast, populations must find the best combinations of mutations to adapt, and these combinations may not involve the most individually beneficial mutations. It has long been suggested that spatial structure may facilitate adaptation, in this case, [8, 22, 23, 25, 65, 87, 116, 128], although it is controversial whether it actually does so in nature [24].

To a certain extent, these two opposing effects of space on adaptation on smooth and rugged landscapes are caused by the same mechanism: spacial structure impedes competition and thus maintains genetic diversity. When an asexual population climbs up a smooth fitness peak, it is always advantageous to maximize the reproduction of the fittest genotype currently present in the population; thus, spatial structure necessarily increases clonal interference and slows down adaptation [65, 78, 87]. However, the situation is less clear when recombination is introduced. At least in some situations, adaptation is primarily driven by recombination between genotypes that are not exceptionally fit. This can certainly be the case in rugged fitness landscapes [22, 25, 128], but it can also be true in smooth ones [100]. Thus while maximizing the reproduction of the fittest individuals maximizes the rate of adaptation in the current generation, it also reduces the genetic diversity that would fuel future adaptation [107, 115]. This suggests that by slowing selection, spatial structure may actually be able to increase the rate of adaptation, even in smooth fitness landscapes.

Natural environments constantly fluctuate, and these fluctuations can affect dispersal and recombination. In particular, both dispersal and recombination are often increased by stressful conditions. Mobile microbes can generally gain mobility from exposure to stressful environments such as starvation, temperature shift, and exposure to toxins [85, 110], in extreme cases manifesting as collective swarming behavior [5, 28]. Some microbes can also switch between asexual and sexual reproduction under stress. For example, bacteria can be induced to sexual reproduction by starvation [38, 60], or can become transformable in order to absorb and integrate exogenous DNA as a general response to stress [19, 20, 72, 104]. A similar response is found in yeast, which can perform meiosis and sporulate under starvation or high oxidative stress [6, 29, 40, 60, 77, 89].

The fact that dispersal and recombination are triggered by some similar environmental stimuli will tend to introduce some synchronization, both between the two processes (individual synchronization) and within processes across individuals (population synchronization, assuming that some environmental fluctuations affect many individuals in the population). In other systems, such as plants that can switch between local vegetative growth or sexual reproduction via dispersed pollen, the two processes may also be mechanistically synchronized. It has been shown that population-level synchronization of facultative recombination can affect the rate of adaptation on both smooth [7, 64, 100, 120] and rugged landscapes [25, 119]. But these results are largely from well-mixed populations, and it is still unclear how synchronization of recombination interacts with population structure and the possible additional individual-level synchronization with dispersal. Intuitively, it seems plausible that population structure could allow diversity to build up in the population in different demes, which could then be brought together to produce new genotypes in a burst of synchronized dispersal and recombination, allowing the population to balance exploration and exploitation of the fitness landscape (Fig. 1.1).

In this paper, we simulate evolution on smooth fitness landscapes and test the effects of population structure and fluctuating, possibly synchronized, recombination and dispersal on the rate of adaptation. We find that when the beneficial mutation supply is high, structured populations with individual- and/or population-synchronized recombination and dispersal can adapt faster than well-mixed ones. Population-level synchronization has the strongest effect, as it maximizes the probability that recombination events will occur between individuals with different beneficial mutations, the pairings that have the highest potential for generating extremely fit offspring.

1.2 Model

We consider a structured population distributed in D demes with N haploid individuals in each deme. We assume an island model, with bi-directional dispersal between all pairs of demes at a time-dependent rate $m(t)$ with average $\overline{m}$. As a baseline, we also simulate a well-mixed population with only one deme with the same total population size, $D \times N$ individuals. In the simulation results, $D = 10$ and $N = 10^6$ unless stated otherwise. Each individual has $L = 1000$ unlinked loci. Beneficial mutations

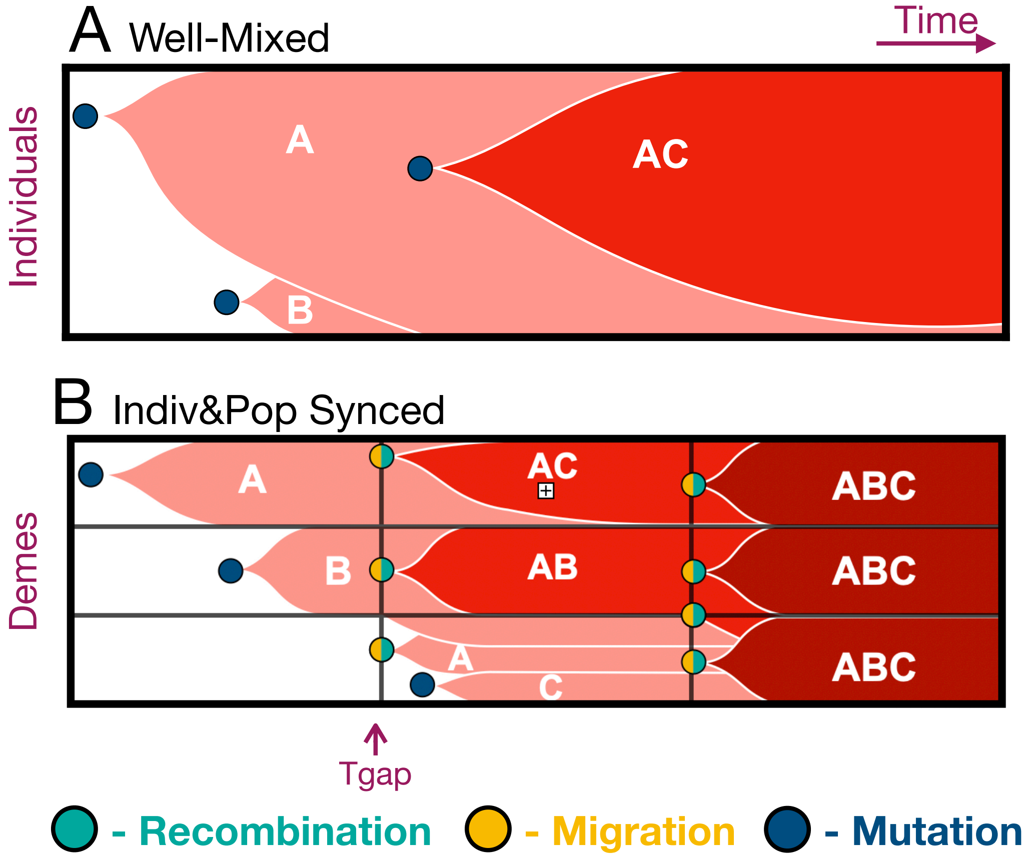


Figure 1.1: Schematic illustration of how population structure and synchronization can combine to accelerate adaptation. Darker red represents higher fitness, i.e., more mutations. **(A)** In a well-mixed population, mutations can rapidly spread to the whole population but drive all other diversity extinct in the process, with few opportunities for recombination between competing genotypes. **(B)** In a structured population, sweeps are slower but this preserves other beneficial mutations, which then can be brought together by synchronized dispersal and sexual reproduction, leading to faster long-term adaptation.

are uniformly distributed and occur at genomic rate U. We neglect back-mutations. Each mutation has the same log-fitness advantage $s = 0.05$ and all mutations combine multiplicatively, i.e., the fitness landscape is completely smooth. Individuals are facultatively sexual, with outcrossing occurring at a time-dependent rate $f(t)$ with average $\overline{f}$. Because most reproduction is asexual, linkage disequilibrium can be large even though all loci are unlinked.

We allow for two possible ways that dispersal and recombination can be synchronized, as illustrated in Fig. 1.2. The first, individual-level synchronization, is a synchronization between the two processes and can be observed within a single generation. If the expected fractions of migrant and sexually produced offspring in that generation are $m(t)$ and $f(t)$ respectively, then there is individual-level synchronization if the expected fraction of offspring that are both migrants and sexually produced is $> m(t)f(t)$. If, for example, $m(t) > f(t)$, then in the most extreme case of individual-level synchronization, *all* sexually produced offspring would also be migrants. For simplicity, we focus on this extreme case in our simulations.

The other form of synchronization, population-level, can only be observed by comparing across multiple generations, and involves synchronizing dispersal, sexual reproduction, or both across individuals. It is automatically induced by the time dependence of $m(t)$ and $f(t)$. In generations when m is high, many individuals disperse, and in generations when it is low, few do, and similarly for f and sexual reproduction. In populations with both forms of synchronization, both dispersal and sexual reproduction are concentrated in the same generations and in the same individuals within those generations.

We consider a limiting form of population-level synchronization, in which dispersal and/or sexual reproduction occur only every t_{gap} generations, with $t_{\text{gap}} = 1$ corresponding to a steady rate, i.e., the absence of population synchronization. For $t_{\text{gap}} > 1$, then if the average probability of, e.g., dispersal is $\overline{m}$, then there will be no dispersal for $t_{\text{gap}} - 1$ generations and then dispersal with probability $\overline{m}/t_{\text{gap}}$ for a single generation. Larger values of t_{gap} therefore correspond to increased population-level synchronization. Note that since the probabilities of dispersal and sexual reproduction must both be ≤ 1 in each generation, t_{gap} must be less than $\overline{m}$ and $\overline{f}$. While this model is largely chosen for simplicity, it is also inspired by the design of microbial evolution experiments that test the effect of spatial structure [65, 87] or facultative outcrossing [83] on the rate of adaptation. In these experiments, organisms have multiple generations of asexual reproduction within separate demes, interrupted by occasional generations of sexual reproduction and/or mixing among demes.

We use a form of Wright-Fisher reproduction, in which the entire population is replaced every generation. To produce an individual in deme d in generation $t+1$, we first determine whether it is a resident (with probability $1 - m(t)$) or a migrant (with probability $m(t)$). We then determine if it is the offspring of uniparental or biparental reproduction. In the absence of individual-level synchronization, these have probabilities $1 - f(t)$ and $f(t)$, respectively. With individual-level synchronization, all resident individuals are produced uniparentally, while migrants have a probability $f(t)/m(t)$ of being produced biparentally. (We always keep $f(t) < m(t)$ in simulations with indivual-level synchronization.) Resident individuals draw their parent or parents from the individuals living in deme d at time t with probability proportional to their fitnesses, Migrant individuals draw their parents from a migrant pool. All demes contribute equally to the migrant pool, but within each deme's contribution, individuals are weighted proportional to their fitness. In other words, selection only acts within demes. For migrants produced by biparental reproduction, the two parents' demes are chosen independently, i.e., mating is assumed to take place within the migrant pool.

The key outcome variable is the rate of adaptation v, defined as the rate of increase of mean log fitness. This is proportional to the probability of fixation of beneficial

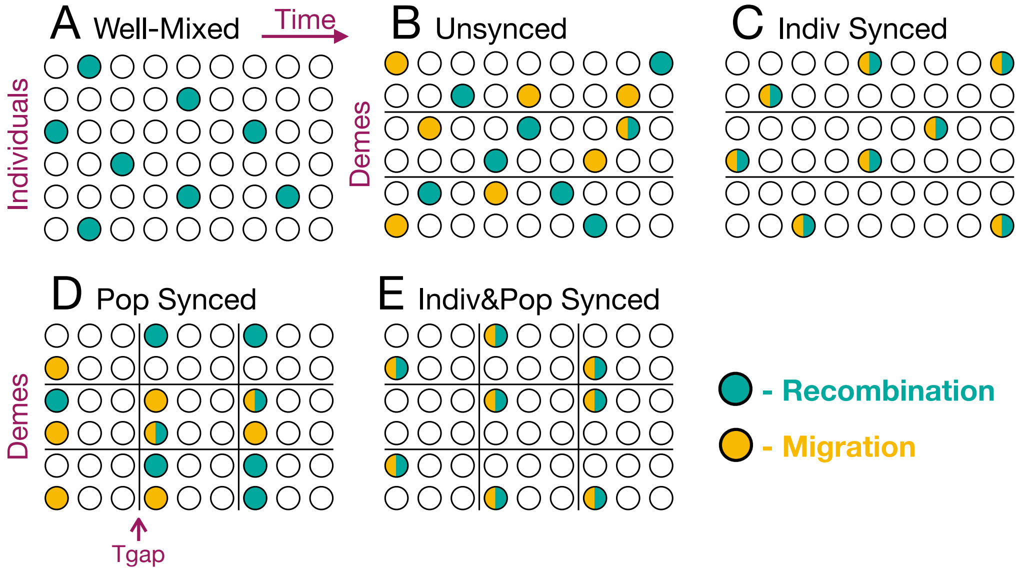


Figure 1.2: Illustrations of different possible forms of synchronization. **(A)** Well-mixed population with no synchronization of sexual reproduction. **(B)** Structured population with no synchronization of dispersal or sexual reproduction. **(C)** Structured population with individual-level synchronization: different dispersal events are independent from each other, and similarly for instances of sexual reproduction, but dispersal and sexual reproduction are correlated with each other. **(D)** Structured population with population-level synchronization: dispersal and sexual reproduction are concentrated in generations that occur every t_{gap} generations, but within one of these generations there is no association between the two processes. Population-level synchronization of sexual reproduction can also occur in well-mixed populations. **(E)** Structured population with both individual- and population-level synchronization: dispersal and sexual reproduction are correlated both between and within generations.

mutations, P_{fix}: $v = NDUP_{\text{fix}} \ln(1+s) \approx NDUP_{\text{fix}}s$. In the absence of clonal interference, $P_{\text{fix}} = 2s$ independent of the spatial structure [82], and the rate of adaptation is $v = v_0 \approx 2NDUs^2$ [120]. To understand the dynamics underlying the observed changes in v, we also track additional statistics of the populations. We measure total genetic diversity using the heterozygosity $H = \sum_i p_i(1-p_i)$, where p_i is the frequency of the mutant allele at locus i. According to Fisher's Fundamental Theorem [37], the speed of adaptation is equal to the (genetic) variance in fitness. In our simulations, the standard deviation of log fitness is always ≤ 0.15, which means that the variance in fitness is close to the variance in log fitness (see Text S1). In linkage equilibrium, this is simply $s^2 \sum_i p_i(1-p_i)$, proportional to the heterozygosity. The extent to which v/s^2 lags behind the heterozygosity therefore provides a measure of total multilocus negative linkage disequilibrium. To track the dynamics of the nose of the fitness distribution, we follow the frequency of "best genotypes", which we define as those within s of the maximum current log fitness in the population (see Text S2). Finally, to determine whether adaptation is being driven by mutations in the nose of the fitness distribution or recombinants leaping to the nose, we track the Hamming distance between the fittest individuals in the population in consecutive generations.

In our simulation, we keep $s = 0.05$ and $L = 1000$ for computational reasons. (Smaller values of s require longer run times, and larger values of L require more memory.) We choose the other parameters to probe the region in which both selection and clonal interference are strong, $1/N \ll P_{\text{fix}} \ll 2s$. This latter condition requires that mutation be frequent and recombination rare. We determined that we could achieve this with computationally feasible total population sizes with mutation rate $U = 5 \times 10^{-4}$ per generation and average frequency of sexual reproduction $\overline{f} = 1.25 \times 10^{-3}$ per generation. Unless otherwise stated, we set average dispersal probability to $\overline{m} = 2.5 \times 10^{-3}$ and the waiting time in population-synchronized simulations to

$t_{\text{gap}} = 100$.

All plots show data averaged over 100 independent simulation runs, with error bars smaller than the size of the plot markers. In plots showing the average dynamics over the course of a cycle of t_{gap} generations, the first and last data points are the beginning of two consecutive bursty cycles, i.e., $t_{\text{gap}} + 1$ generations are shown.

1.3 Results

1.3.1 Population structure can very slightly speed adaptation even without synchronization

In asexual populations, clonal interference is stronger in spatially structured populations than it is in well-mixed ones [65, 78]. Sexual reproduction reduces clonal interference in both spatially structured [78] and well-mixed populations, but how does it change their relative rates of adaptation, i.e., the effect of spatial structure? To investigate this, we first simulate populations with facultative sex and varying degrees of spatial structure but no synchronization (schematic: Fig. 1.2B; results: Fig. 1.2 and Fig. A.2B, orange lines).

At small total population sizes such that the mutation supply is low ($NU \lesssim 1$), well-mixed populations experience little clonal interference and adapt at close to the maximum rate v_0, even if they are asexual. In this regime, increasing spatial structure only impedes adaptation (Fig 1.2A, left side), because it slows down sweeps, increasing the probability that they overlap and interfere with each other. However, for large mutation supplies ($NU \gg 1$), well-mixed populations with limited recombination experience strong clonal interference. Surprisingly, in this regime, splitting the population into demes does not slow down adaptation and even very slightly accelerates it, with the population split into 100 demes evolving 6% faster than the well-mixed population (t-test, $p = 0.035$; see Fig. 1.2A).

To understand this result, note that under strong clonal interference the speed of adaptation is only weakly (logarithmically) dependent on population size [32, 88, 120, 121]. Thus, in a population large enough that each individual deme has a high mutation supply, each deme can adapt nearly as fast as a well-mixed population of the same size as the whole meta-population. So space only slightly slows down asexual evolution, and the increased genetic diversity between demes means that spatially structured populations can benefit more from recombination, pushing their overall rate of adaptation above well-mixed ones (Fig. A.1 and 1.2A).

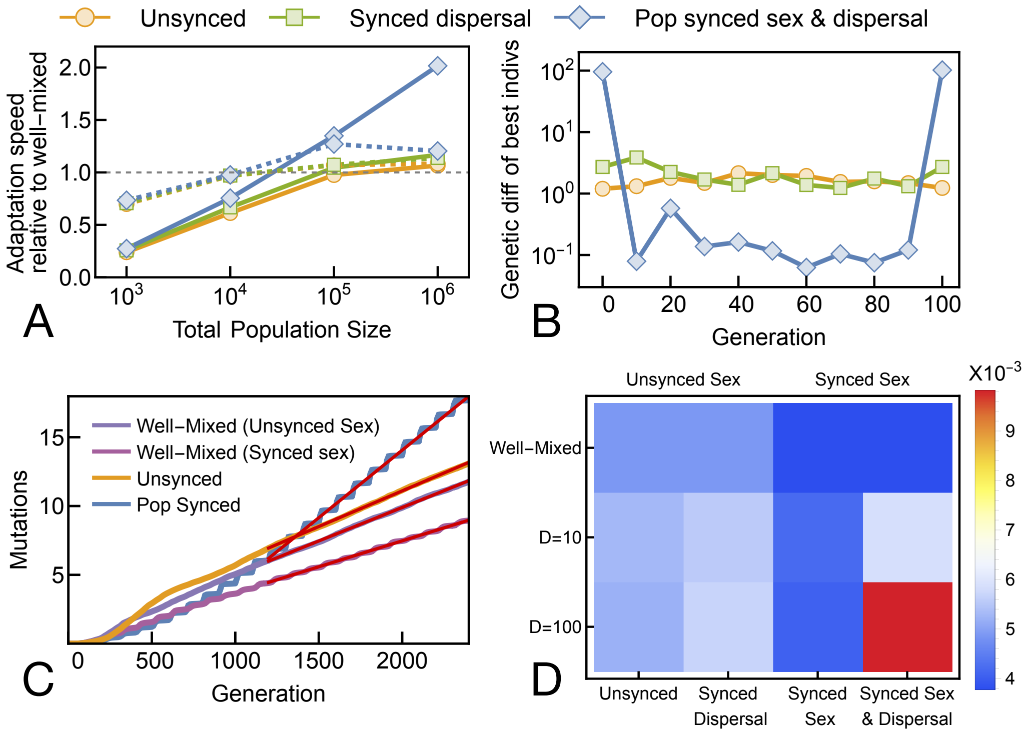

Unsynced
Synced dispersal
Pop synced sex & dispersal
Adaptation speed relative to well–mixed
2.0
1.5
1.0
0.5
0.0
10^3
10^4
10^5
10^6
Total Population Size
A
Genetic diff of best indivs
10^2
10^1
10^0
10^{-1}
0
20
40
60
80
100
Generation
B
Mutations
15
10
5
0
500
1000
1500
2000
Generation
Well–Mixed (Unsynced Sex)
Well–Mixed (Synced sex)
Unsynced
Pop Synced
C
Unsynced Sex
Synced Sex
Well–Mixed
D=10
D=100
Unsynced
Synced Dispersal
Synced Sex
Synced Sex & Dispersal
X10^{-3}
9
8
7
6
5
4
D

Figure 1.2: Population-level synchronization of dispersal and sexual reproduction accelerates adaptation by increasing the production of fit recombinants. **(A)** Ratios of adaptation speeds relative to a well-mixed population with unsynchronized sexual reproduction. Dashed lines show results from simulated populations comprising 10 demes, while solid lines show populations comprising 100 demes. The strongly structured (100-deme) population with synchronized sexual reproduction and dispersal adapts roughly twice as fast as other populations. Structured populations with no synchronization or synchronization of only dispersal have slight advantages over the well-mixed population (ratios=1.06 ($p = 0.035$) and 1.17 ($p < 0.001$), respectively, for populations of 100 demes). Details of the adaptation speeds are in Fig. A.2. In all other simulation result, $D = 100$ and $N = 10^6$ unless stated otherwise, since they produces the fastest adaptation. **(B)** Hamming distance between the fittest individuals in successive generations, shown over the course of the $t_{\text{gap}} = 100$ generations between rounds of synchronized dispersal/sexual reproduction. All curves are averages over 10 runs, each of which consists of 2400 generations (i.e., 24 t_{gap} periods). For populations without synchronization or synchronization of only dispersal, the small (≈ 1 mutation) distances indicate that adaptation is primarily driven by the accumulation of mutations on already-fit backgrounds. Populations with synchronized sexual reproduction and dispersal, on the other hand, leap ahead suddenly by ≈ 100 mutations due to recombination between divergent parents. **(C)** Mutation accumulation trajectories of different populations. The adaptation of the synchronized population is punctuated. Red lines indicate fits used for determining adaptation speeds. **(D)** Relative adaptation speeds (rates of increase of mean fitness) of different populations. The full combination of Strong structure and synchronized sexual reproduction and dispersal is needed to substantially accelerate adaptation. Individually, these factors only have very small effects or even slow down adaptation.

1.3.2 Population synchronization of sexual reproduction and dispersal can significantly accelerate adaptation

We first examine the effects of pure population synchronization, in which both sexual reproduction and dispersal are synchronized at the population level, with no additional individual-level synchronization (schematic: Fig. 1.2D). This represents a scenario in which, for example, an entire experimental population is forced to go through periodic rounds of dispersal and sexual reproduction [65, 83]. We see that when clonal interference is strong, this life history can approximately double the speed of adaptation relative to that of a well-mixed population with no synchronization (Fig. 1.2A). The increase in speed relative to a well-mixed population with the same synchronization of sexual reproduction is even larger (Fig. A.2A). Thus, spatial structure reduces clonal interference in this instance.

The population with synchronized sexual reproduction and dispersal achieves its higher adaptation speed via bursts of adaptation every t_{gap} generations. These bursts move both the nose and the mean of the fitness distribution. But the nose (Fig. 1.2B) shifts in genotype space by far more than the distribution moves toward the optimal genotype (Fig. 1.2C). This indicates that the leading genotype is replaced by a distantly related recombinant, unlike in the other populations. The combination of structure and synchronization apparently allows it to both maintain and exploit a large reservoir of beneficial mutations beyond those that are found in the current best genotype.

1.3.3 Sexual reproduction among migrants is crucial to the increase in adaptation speed

We next introduce individual-level synchronization, such as could be a response to individual-level environmental variation, as opposed to the global environmental fluctuations that lead to population synchronization. We find that it can also accelerate adaptation, although not by as much as population synchronization (Fig. 1.3A, yellow squares vs dark blue diamonds). Populations with both forms of synchronization adapt fastest of all (Fig. 1.3A, light blue triangles), although the increase in speed over pure population-level synchronization is modest at the large population sizes where synchronization accelerates adaptation the most.

There are two possible explanations for why individual-level synchronization provides only a modest acceleration in large populations that already have population-level synchronization: either it is not important that sexual reproduction be happening specifically among migrants, or population-level synchronization already creates a large enough pool of sexually reproducing migrants that there are only limited benefits to increasing it further. To test these possibilities, we modify the simulations with only population-level synchronization so that either only migrants or only residents can reproduce sexually. We do not increase the rate of sexual reproduction among the group where it is allowed, so this lowers the overall rate of sexual reproduction—drastically so when sexual reproduction is limited to migrants, who are typically a minority of the population. For example, at the value $t_{\text{gap}} = 100$ where synchronization provides the greatest benefit, 25% of individuals are migrants in the high-dispersal generations. But we see that limiting sexual reproduction to these migrants hardly slows down adaptation, while limiting it to the 75% of individuals who are residents reduces the rate of adaptation by more than a factor of two (Fig. 1.3B). We, therefore, see that sexual reproduction among migrants is essential to the advantage of synchronization, suggesting that the limited benefits of individual-level synchronization are simply because population-level synchronization already creates a strong association between the two processes.

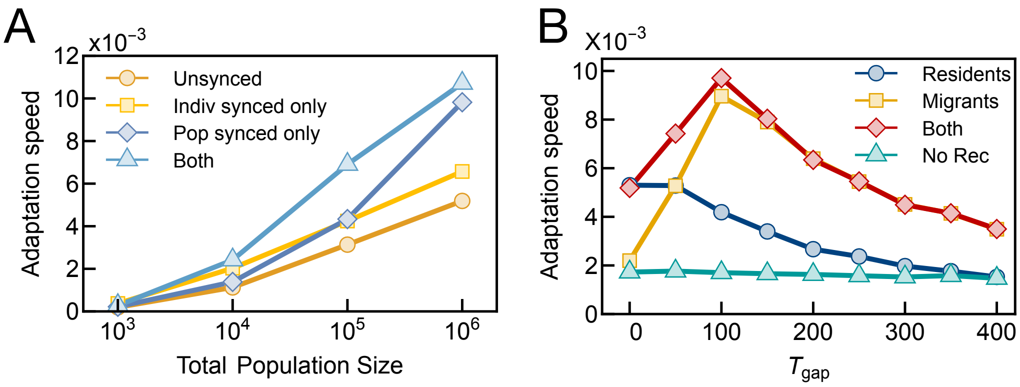


Figure 1.3: Sexual reproduction among migrants drives the increase in adaptation speed. **(A)** Adaptation speed of populations with differing synchrony between sexual reproduction and dispersal. Individual-level synchronization between sexual reproduction and dispersal accelerates adaptation, although not by as much as population-level synchronization. The two forms of synchronization act synergistically at moderate population sizes but have diminishing returns at a **(B)** effect on adaptation speed of limiting sexual reproduction to migrant or resident individuals in populations with population-synchronized sexual reproduction and dispersal. For low values of the synchronization parameter t_{gap}, almost all individuals are residents and sexual reproduction among migrants contributes negligibly to adaptation. At higher values of $t_{\mathrm{gap}} \approx 100$, even though migrants are still a minority, it is their offspring who are driving adaptation, with the majority residents making a negligible contribution. For even higher values $t_{\mathrm{gap}} > 200$, migrants make up a majority of the population in the generations in which sexual reproduction takes place.

1.3.4 The synchronization *between* sexual reproduction and dispersal is crucial to the increase in population speed

Population-level synchronization both synchronizes dispersal with sexual reproduction and synchronizes among multiple dispersal events and among multiple instances of sexual reproduction. It seems intuitive that the former effect should be the most important for adaptation, since it focuses sexual reproduction on pairings with the greatest genetic diversity. But individual-level synchronization also produces this effect, and does not accelerate adaptation nearly as much (Fig. 1.3A). This raises the question of whether perhaps the synchronization between sexual reproduction and dispersal is actually unimportant, with the acceleration being driven by just the separate synchronizations among dispersal events and among instances of sexual reproduction. To test this, we introduce an offset between generations of increased dispersal and generations of increased sexual reproduction. Experimentally, this would correspond to having separate passages in which cells were mixed across wells and in which some cells were forced to undergo sexual reproduction. In natural populations, it would correspond to different pathways triggering dispersal and sexual reproduction in response to different environmental cues.

We find that the rate of adaptation is maximized when sexual reproduction and dispersal occur in the same generation (Fig. 1.3A), confirming the initial intuitive expectation. The rate of adaptation decreases as the number of generations elapsing after dispersal and before sexual reproduction increases, as the genetic diversity introduced into demes by dispersal is lost. However, we find a new surprise, that the rate of adaptation partially recovers for very long delays, such that dispersal quickly *follows* sexual reproduction. Examining the statistics of the simulations further reveals that the initial decrease in the rate of adaptation can be explained by a decrease in the fitness of the fittest recombinants formed (Fig. 1.3B). This effect saturates for long delays, likely because within-deme genetic diversity reaches a steady state maintained by mutation. At this point, it is better for adaptation to delay sexual reproduction more so that fit recombinants can quickly be redistributed across demes in the *next* dispersal event. This is because the demes vary in fitness (Fig. 1.3C), so dispersal will tend to move very fit recombinants into less-fit demes, where they will compete with each other less. For these very long offsets, the absolute adaptation speed ($5.23 \pm 0.11 \times 10^{-3}$ for offset $= 99$) is similar to that for populations with synchronized dispersal but unsynchronized sexual reproduction ($5.68 \pm 0.15 \times 10^{-3}$, Fig. A.2C) or no synchronization ($5.19 \pm 0.12 \times 10^{-3}$, Fig. A.2B).

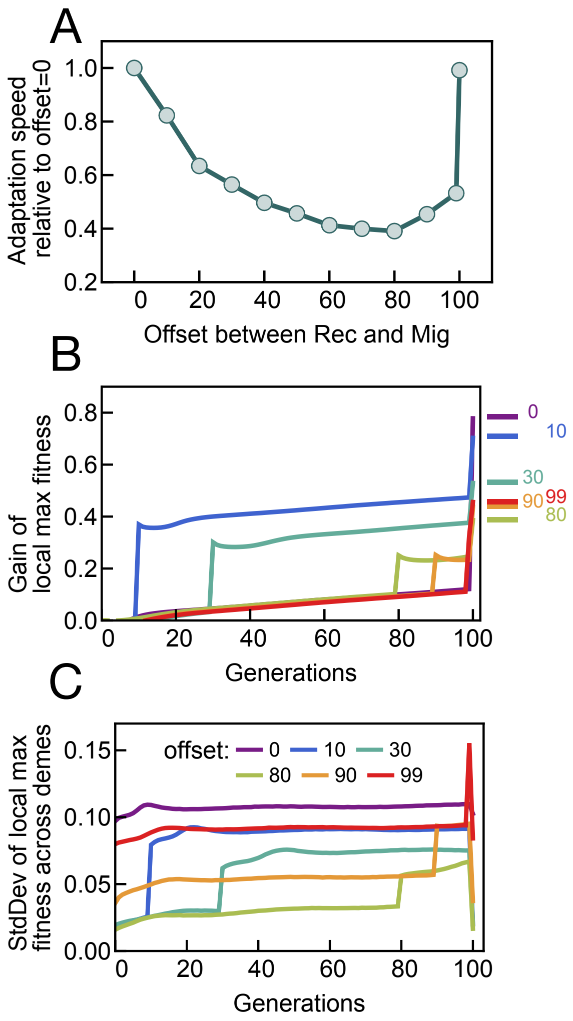

A
Adaptation speed
relative to offset=0
1.0
0.8
0.6
0.4
0.2
0
20
40
60
80
100
Offset between Rec and Mig
B
Gain of
local max fitness
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80
100
Generations
0
10
30
90
99
80
C
StdDev of local max
fitness across demes
0.15
0.10
0.05
0.00
offset:
0
10
30
80
90
99
0
20
40
60
80
100
Generations

Figure 1.3: Simultaneous sexual reproduction and dispersal maximizes the rate of adaptation. **(A)** Adaptation speed as a function of the delay between high-dispersal generations and high-sexual reproduction generations. Dispersal occurs at times 0 and 100 on the horizontal axis, so an offsets of 0 or 100 are identical. Increasing the delay before sexual reproduction lowers the rate of adaptation, as the increased within-deme genetic diversity introduced by dispersal is lost. Surprisingly, there is a slight uptick in the rate of adaptation for very long delays such that dispersal follows shortly *after* sexual reproduction. **(B)** Increase in the fitness of the fittest individual in a deme over the $t_{\text{gap}} = 100$ generations between dispersal events, for different values of the delay before sexual reproduction. This local maximum fitness jumps in the sexual reproduction generation, but by less as the delay increases and the genetic diversity introduced by dispersal decays. It appears to reach a steady state at ≈ 80 generations, partially explaining the uptick in panel A. **(C)** Standard deviation of local maximum fitness across demes shows the same qualitative pattern as overall adaptation speed (A), initially decreasing as the delay increases and then rebounding for very long delays. Overall, the standard deviations are substantial, suggesting that dispersal soon after recombination may help the fittest recombinants by moving them to less competitive demes, which could explain the uptick. Details are in Fig. A.5.

1.4 Discussion

In this paper, we demonstrate that population structure, by creating the possibility for synchronized dispersal and sexual reproduction, can increase the rate of adaptive evolution. To our knowledge, this is the first demonstration that a population structure that does not affect the fixation probability of isolated alleles can accelerate adaptation on smooth fitness landscapes without epistasis by reducing clonal interference among alleles. It does this both by alternately accumulating diversity in different demes and then bringing it together and forming fit recombinants, and by tending to concentrate recombination events in pairs of migrant individuals who are likely to be genetically distinct.

1.4.1 Synchronization of dispersal and sexual reproduction in experiments and natural populations

Our model is meant to match the most straightforward way of implementing population structure and sexual reproduction in experimental microbial population grown in batch culture. Since dispersal is naturally done at transfers, it is automatically synchronized [65, 87]. Synchronized sexual reproduction (or more generally, recombination) is also natural when working with organisms such as yeast or many bacteria where it must be induced by environmental conditions [19, 83].

We believe that our model also captures important features found in natural populations. Facultative sexual reproduction is common in nature [93]. Sexual reproduction often carries costs, including mate-finding, performing sexual behavior, and producing males [109]. Facultative sexual reproduction provides most of benefits in terms of creating genetic diversity with less cost than obligate sexual reproduction [12, 64, 93]. If sexual reproduction is rare, having it be synchronized helps reduce mate-finding costs [64].

Facultatively sexual reproduction is often a response to stressful conditions [93]. To the extent that these stressful conditions affect multiple individuals, this provides one mechanism for synchronization. As dispersal is also often a response to stress (e.g., [81]), it will naturally also be synchronized with sexual reproduction. While environmental fluctuations are implicitly the source of synchronization in our model, the fitness landscape is static. Presumably the primary reason that sexual reproduction and dispersal can be triggered by stress is that it is a signal that the organism is poorly adapted to its present habitat, and that it should try to improve the match by producing offspring with different haplotypes or moving to a new environment. In other words, the response most likely evolved as a way to track the fluctuating components of the fitness landscape. Our work shows that a side effect can be more rapid adaptation to the fixed components of the landscape as well.

Chapter 2

Promoting moderately fit individuals can increase adaptation speed under strong clonal interference

Abstract

When mutations are rare, strong selection can efficiently exploit the existing mutations to achieve the fastest adaptation, as the best genetic background can be built for future mutations. However, when mutations are common, different lineages carrying unique beneficial mutations can coexist and compete for fixation, leading to strong clonal interference. Under these conditions, strong selection eliminates distinct mutations on mediocre genetic backgrounds, ultimately decreasing genetic variation. Therefore, it is crucial to maintain a balance between variation and selection, especially in the presence of recombination. Recombination can reduce clonal interference and combine unique beneficial mutations from different competing lineages into a single individual. To achieve this, we propose a logistic fitness function an alternative

to the exponential function typically used in population genetics. The logistic fitness function promotes the evolutionary advantage of the moderately fit individuals and impedes the sweeping process of the fittest individuals, thereby preserving genetic diversity at the cost of short-term response. At the same time, the profound genetic diversity creates a faster long-term adaptation. We compare the performance of the logistic fitness function with two other fitness functions and find that the effect is not directly due to the limited selection of the fittest individuals, but rather due to the increased fitness of ordinarily fit individuals. Our findings suggest that balancing genetic diversity and selection for sexual reproduction could be a critical strategy for adaptation in a strong clonal interference environment.

2.1 Introduction

The strategy for achieving the fastest adaptation is contingent on the specific details of a population. In general, the rate of adaptation is primarily influenced by the mutation rate U, population size N, and the probability that new mutations become fixed P. When the mutation influx is small, the rate of mutation accumulation can be expressed as the product of the mutation supply and the fixation probability, $v = NUP$, where N is the population size, U is the mutation rate and P is the fixation probability. For a homogeneous population, the fixation probability of a beneficial allele is $P \approx 2s$ for $s \ll 1$, resulting in a baseline rate of mutation accumulation is $v \approx 2NUs$ [44, 63, 98]. In this case, stronger selection leads to faster adaptation, as the rate of adaptation is positively correlated with the selective coefficient s.

However, in cases where the mutation supply is abundant, clonal interference can occur, wherein multiple beneficial genotypes compete for fixation [41, 97]. In such circumstances, the balance between selection and mutation is crucial in the evolution process. Selection reduces the fitness distribution while mutation expands it [32].

Although strong and accurate selection can produce immediate responses, it can also reduce the effective population size by limiting the number of parents and lineages that contribute to the next generation [54, 105]. This would eventually lead to a decrease in the long-term response due to the lack of genetic variation.

The variation-selection balance is particularly important for sexual populations as recombination of mutations from competing lineages can lead to the emergence of novel traits [37, 86]. Previous research has demonstrated that the relationship between adaptation speed and selection strength is not always monotonic; in fact, a relaxed selection can optimize the long-term response of a quantitative trait [52, 53, 105, 107, 115, 133]. In addition, it has also been studied in the context of breeding problems, where the goal is to maximize response to artificial selection based on various factors such as effective population size and inbreeding rate [10, 13, 48, 106].

To maintain genetic variation, previous studies have employed either uniformly weaker selection or selecting a larger proportion of the leading nose of the population [52, 53, 105, 107, 115]. However, these methods still heavily rely on the fittest individuals in the population. In contrast, in the previous section, we demonstrated that population structure can have a similar effect in balancing selection and genetic variation. Specifically, our simulations showed that a structured population can exhibit similar, if not better, adaptation speed than a well-mixed population. This is because population structure extends the time for favorable mutations to fixation, allowing other genotypes to accumulate subsequent mutations and increase clonal interference [17, 39, 123]. These findings suggest an alternative artificial selection strategy, which is to limit the evolutionary advantage of the fittest individuals.

In this paper, we propose a general strategy for creating artificial selection by altering the fitness function to mimic the effects of population subdivision. By default, the population has an approximately exponential growth rate of its fitness, which is derived from the growth factor per generation $W = e^r \approx 1 + r$ [32]. We replace this exponential fitness function with a logistic fitness function to impose limitations on the best individuals. Our study shows that the logistic fitness function is effective only when clonal interference is strong. Under such conditions, the logistic fitness function can preserve genetic diversity by hindering the sweeping process of the fittest individuals, similar to the population subdivision. This approach fosters competing lineages to maintain genetic diversity at the cost of short-term adaptation, while yielding a better long-term response. We find that the rate of adaptation can increase by a factor of 2.79 compared to the original exponential fitness function, and that this effect is due to the increased fitness of moderately fit individuals. Our results suggest that a universal strategy for evolution in a strong clonal interference environment could involve limiting the fittest individuals and promoting mediocre individuals to maintain high genetic variation. Overall, our research provides insight into the delicate balance between selection and genetic diversity in population genetics and highlights the importance of maintaining this balance for long-term adaptation. The proposed logistic fitness function offers a potential strategy for achieving this balance, and future studies could further explore its efficacy in various evolutionary scenarios.

2.2 Model

Table 2.1: Symbol definitions

Symbols	Definition
W	Fitness
z	Breeding value, a trait in a population
N	Haploid population size
L	Genome size
r	Frequency of sexual reproduction
U	Genomic beneficial mutation rate
s	Selective coefficient of beneficial mutations
P	Fixation probability of a mutation
v	Adaptation speed in breeding values (mutation numbers)
v_0	v in the absence of interference

The focus of this paper is on the adaptation speeds of breeding values from different fitness functions. We define the breeding value $z = k$ as the number of additive beneficial alleles. This implies the adaptation speed in z is $v = NUP \approx 2NUs$[44, 63, 98].

By default, the fitness is exponentially correlated with the breeding value, $W(z) = e^{sz}$ [32, 120]. To limit the exponential advantage of preeminent individuals, we considered the shifted logistic fitness function:

$$W(z) = \frac{c}{1 + e^{-ks(z-z_0)}} \tag{2.1}$$

The three coefficients can be determined by requiring that the Taylor series expansion of $W(z)$ around $z = 0$ satisfies the following conditions:

$$\begin{cases} W(0) = & 1 \\ W'(0) = & 1 \\ W''(0) = & d \end{cases}$$

The first two requirements ensure the novel mutations have additive advantages in a homogeneous genetic background, which should be identical to the exponential fitness function. The second derivative is to tune the fitness function. The fitness function can be rewritten in the form of d:

$$W(z) = \frac{2(1-d)}{1 - 2d + e^{-2s(1-d)z}} \tag{2.2}$$

Notably, when $d = 1/2$, Eq.2.2 becomes the exponential function. In this paper, we set $d = -2.715$ based on simulation results, while maintaining the polynomial approximation for the infinitesimal model's monotonic increase.

The mean value of the fitness function should be 1 in a population with constant size, which is forced by the sampling method. Thus the effective fitness function is normalized and has a different coefficient c' where $z \sim \mathcal{N}(0, \sigma^2)$. According to Supplementary B.1, it should have the form as:

$$W_{\text{effective}}(z) = \frac{1 + \exp(ksz_0/\gamma)}{1 + \exp\left(-ks(z - z_0)\right)}; \quad \gamma = \sqrt{1 + \frac{k^2 s^2 \pi \sigma^2}{8}} \tag{2.3}$$

Overall, the logistic fitness function harbors mediocre individuals and limits the evolutionary fitness of the individuals on either side of the distribution. This is shown in Fig.2.1.

As the logistic fitness function is scaled based on the variance of the breeding value, the selective advantages of novel mutations differ under different fitness functions. Consequently, the condition $W'_{\text{effective}}(0) = 1$ is no longer valid. To enforce this condition, we rescaled the selective coefficient s by $W_{\text{effective}}(1|s_L) = e^s$, so that a novel mutation has an equivalent evolutionary effect on both fitness functions. Specifically, when $s_E = 5 \times 10^{-2}$, the scaled selective coefficient $s_L = 6.06 \times 10^{-2}$. The impact of rescaling the selective coefficient should be mild since the differences between two selective coefficients are minor and its impact would be reduced when clonal interference is strong. Details can be found in Supplementary B.3.

In the simulation, we consider the Wright-Fisher population with N haploid individuals. Each individual possesses L loci, and each locus accumulates favorable mutations with a rate of U. Parents are chosen by their fitnesses, which proportionally rely on the breeding value $sz = sk$, where k is the number of beneficial mutations and the selective coefficient $s = 0.05$. The fitness functions can be exponential, logistic, and other specified functions. Sexual reproduction occurs at a rate of r, with offspring having a equal chance to inherit from either parent at any locus. The adaptation speed is obtained from linear fitting the mutation trajectories over generations. In order to obtain a steady adaptation speed, we only fit the latter half

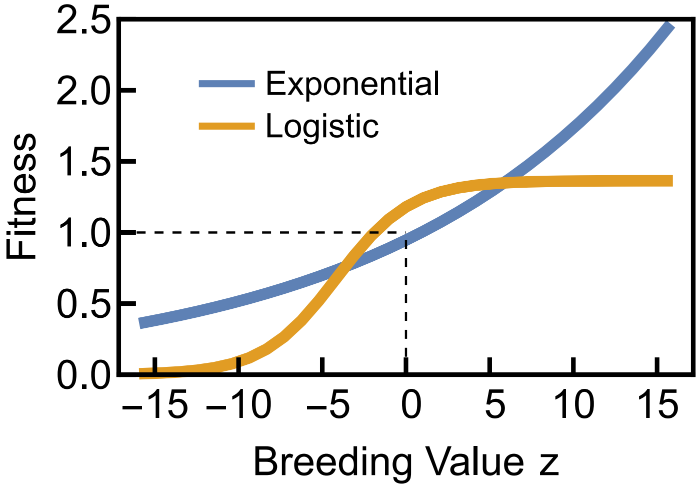


Figure 2.1: The comparison of fitness functions. The logistic fitness function restricts the dominance of the fittest individuals and promotes the fitness of the general individuals. The fitness functions are plotted in the range of $3\sigma_z$, which is obtained in Fig.2.5a.

of the trajectory.

We use simulation to investigate the adaptation speed with strong clonal interference. In this regime, the fixation probability is expected to be $\frac{1}{N} \ll P \ll 2s$. The two bounds are from the circumstances of random fixation and strong selection respectively. This condition also be written as $U \ll v \ll 2NUs$. In our simulation, the parameters are $N = 1 \times 10^6, L = 1.2 \times 10^4, S = 5 \times 10^{-2}, U = 3 \times 10^{-3}$, and $r = 2 \times 10^{-2}$. Our simulation includes beneficial and back mutations which occur randomly at any site by a geometric distribution. The simulation terminates when half of the genome is mutated when back mutation dominates.

2.3 Results

2.3.1 Adaptation speed with weak clonal interference

We consider a Wright-Fisher population with unlinked loci. Since the population has a large genome ($>= 10^6$) and each gene makes a small contribution, we assume the infinitesimal model [4, 54, 105, 120]. An individual with trait value z produces a Poisson distributed number of offspring with the expectation of $W(z)$. The offspring's trait values follow a normal distribution around the mean of their parents, with a constant variance V_O. With random mating, the next generation would have the variance $V_P/2+V_O$, while the variance of the parents is V_P. At equilibrium, $V_P \approx 2V_O$.

In our model, the mating is polygamous. Both parents are drawn with weights given by $W(z)$. The probability of fixation P can be obtained by calculating the probability of loss. It has been shown that the exponential fitness function has the adaptation speed as $v = \frac{1}{4}\mathfrak{W}(v_0) = \frac{1}{4}\mathfrak{W}(2NUs^2)$, while $\mathfrak{W}(x)$ is a product log function [120]. We show the approximated adaptation speed with the logistic fitness function in Supplementary B.2 while $z, v \ll 1$, when the clonal interference is weak. Overall, their approximated adaptation speeds are:

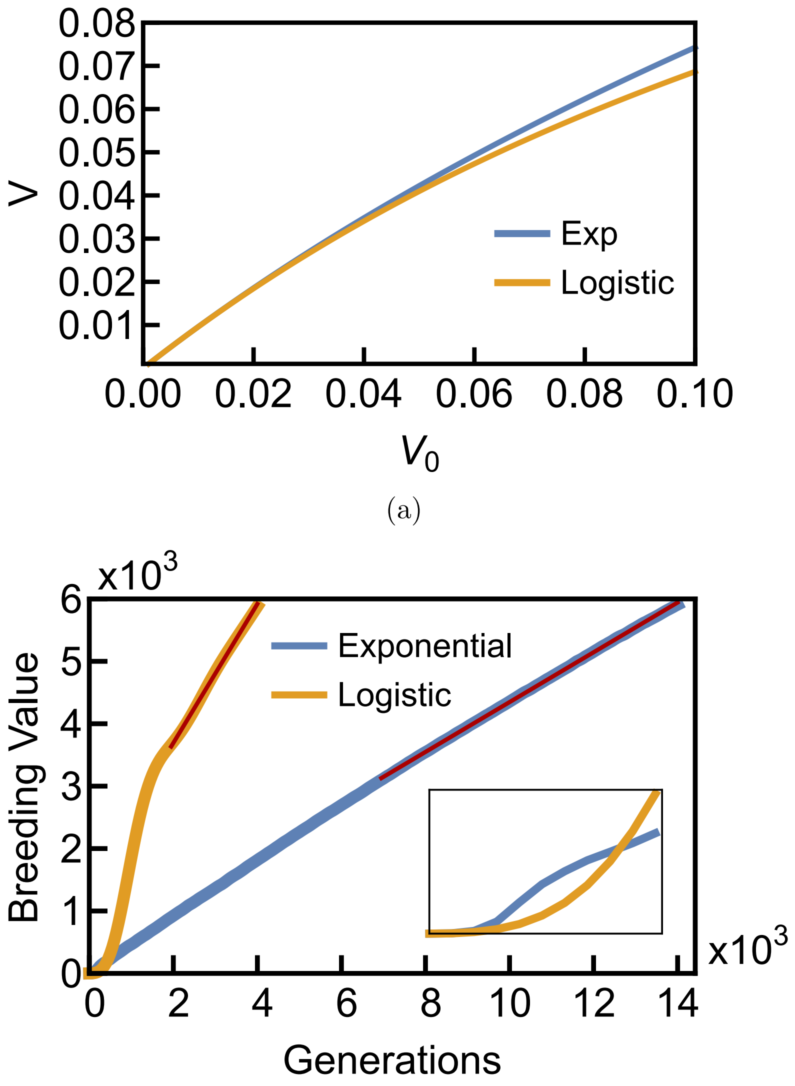
$$\begin{aligned} v_{\text{exp}} &= \frac{1}{4}\mathfrak{W}(v_0) \approx v_0 - 4v_0^2 + 24v_0^3 - \frac{512}{3}v_0^4 + \mathcal{O}(v_0^5) \\ v_{\text{logistic}} &\approx v_0 - 4v_0^2 + \frac{32}{9}\left(11 + d - d^2\right)v_0^3 + \frac{64}{27}\left(-200 - 36d + 33d^2 + 4d^3\right)v_0^4 + \mathcal{O}(v_0^5) \end{aligned} \tag{2.4}$$

As shown in Fig.2.2a, their adaptation speeds are similar.

Noted that the above studies are based on the assumption of full sexual reproductions. The results would differ when facultative sex is involved. In this case, the selective coefficient needs to be rescaled by $1/r$. Details are included in the discussion.

2.3.2 Adaptation speed with strong clonal interference

Currently, the analytic solution to the adaptation speed with strong clonal interference remains unraveled. Therefore, we use simulations to investigate the effect of limiting the best individuals. Fig.2.2b demonstrates that restricting the fitness of the best individuals generates a trade-off between short-term and long-term evolutionary dynamics. While the exponential fitness function enables swift adaptation to the new



(a)

(b)

Figure 2.2: The rate of adaptation from approximation and simulation. **(a)**. When the clonal interference is weak, the adaptation speed of populations under exponential and logistic fitness functions is relatively similar. **(b)**. When the clonal interference is strong, the population with the limited evolutionary advantage evolves faster than the one with multipliable advantages (ratio ≈ 2.79). The exponential fitness function creates a rapid adaptation in early generations. However, the logistic fitness function has a faster long-term adaptation (0.398 ± 0.002 for E and 2.79 ± 0.02 for L). Noted that the trajectories can not extend infinitely with finite genome size. The bottom inset shows the adaptation in the first 500 generations. The red lines indicate the linear regression and the adaptation speeds are acquired from their slopes.

environment, it comes at the cost of diminishing long-term adaptability. In contrast, the logistic fitness function prevents the best individuals from quickly dominating the entire population, so it exhibits slower initial adaptation. However, this maintains the genetic diversity of mediocre individuals, which can be leveraged by recombination. With recombination, diverse genetic resources can be utilized to create novel genotypes. Eventually, the logistic fitness function can have a better long-term evolution (ratio of adaptation speed ~ 2.79).

Since selection is weak ($s = 0.05$) and recombination is strong (totally random at each locus), linkage disequilibrium (LD) among alleles sweeping to fixation is expected to be negligible. Therefore, the heterozygosity $H = \sum_L p_i(1 - p_i)$ is proportional to the variance of breeding value z:

$$\sigma_z^2 = \sum_L \sigma_i^2 = H \tag{2.5}$$

while p_i is the frequency of loci i.

By Fisher's "Fundamental Theorem", the rate of increase in log fitness is given by its heritable variance; this is $v = s\sigma_z^2$. Therefore, the genetic diversities should be proportional to the adaptation speed without linkage disequilibrium (LD):

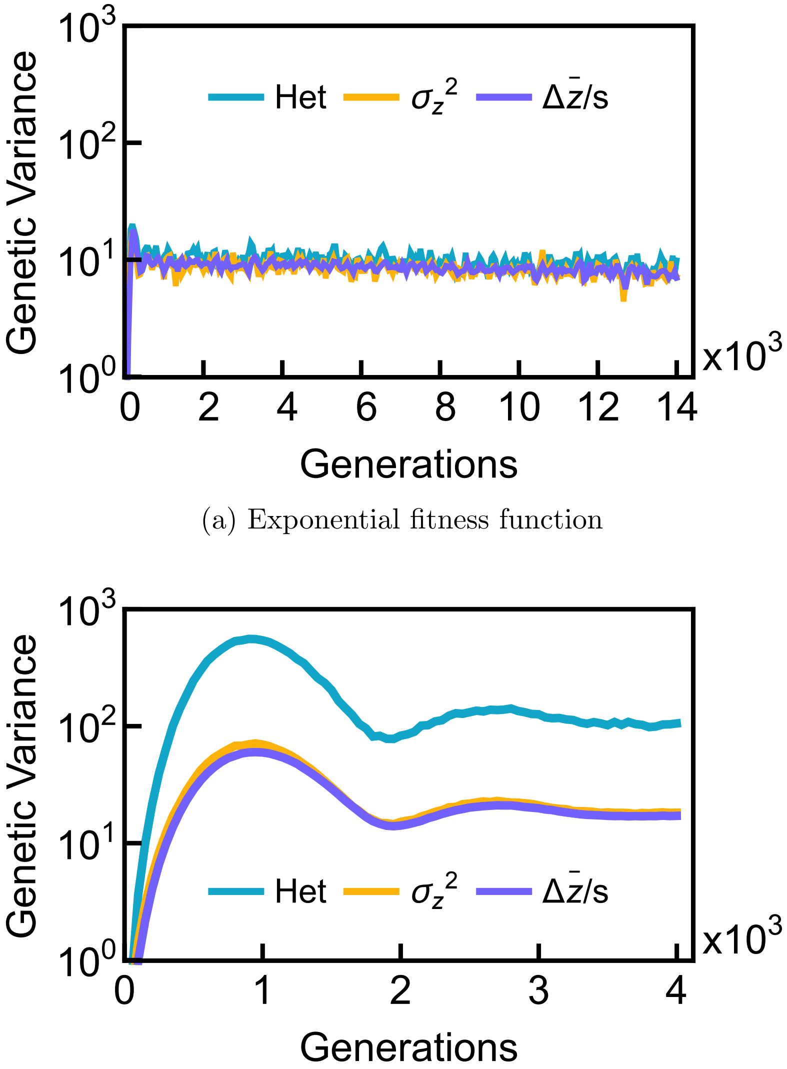
$$v/s = \sigma_z^2 = H \tag{2.6}$$

The Price Equation relates changes in the mean phenotype (Δz) to changes in the mean fitness ($\bar{w}$) and the covariance between them ($\mathrm{cov}(w_i, z_i)$). In this study, since the fitness functions are normalized and approximated as $W(z) \approx 1 + sz$ for both functions, the rate of adaptation can be expressed as $\Delta z \sim s\sigma_z^2$ when $z \sim \bar{z}$. The results presented in Fig.2.3 confirm that $\sigma_z^2 \approx \Delta\bar{z}/s$, which signifies the establishment of the Price Equation.

Furthermore, when the population is in linkage equilibrium, the heterozygosity is equivalent to the variance of mutations, which is also equivalent to the breeding value z, as shown in Eq.2.5. However, with the logistic fitness function, the heterozygosity is over 10 times higher than the variance of z. This observation indicates the existence of negative linkage disequilibrium, where individuals with similar fitnesses have distinct genotypes.

The presence of negative linkage disequilibrium implies that individuals with distinct genotypes coexist in the population. When the advantage of the fittest individuals is limited, their sweeping process slows down. This leads to strong clonal interference, where beneficial genotypes compete for fixation. The abundance of clonal interference can have a duo effect. On one hand, it reserves sufficient genetic diversity for sexual reproduction. Recombination can utilize the distinct mutations to generate novel genotypes for further adaptation. On the other hand, moderately fit individuals may impede the spread of those novel genotypes, slowing down adaptation. Generally, sufficient sexual reproduction is required to efficiently collaborate with abundant genetic diversity. As a result, the distinct mutations can be efficiently integrated into the same genotype, while the mediocre individuals would be eliminated by sex and selection.

In our simulation, Fig.2.4 implies that sexual reproduction is effective to cooperate with genetic diversity. With the logistic fitness function, the Hamming Distance between the best individuals in consecutive generations remains high, even in the stable phase. It indicates that the population experiences significant genetic changes (mean $= 213$ mutations) between generations. This is the result of the integration of existing mutations by recombination. Meanwhile, the exponential fitness function shows mild genetic changes (mean$=6.81$ mutations), which suggests sequential evolution of the same genotypes over generations. However, it is unknown if the sex rate can be further increased due to the limited computational resources, which will be discussed later in the paper.



(a) Exponential fitness function

(b) Logistic fitness function

Figure 2.3: The genetic diversity with different fitness functions. The variance of z is closely related to the rate of adaptation, which is consistent with the Price Equation. However, with the logistic function, the heterozygosity exceeds the other two types of genetic diversities, indicating negative linkage disequilibrium.

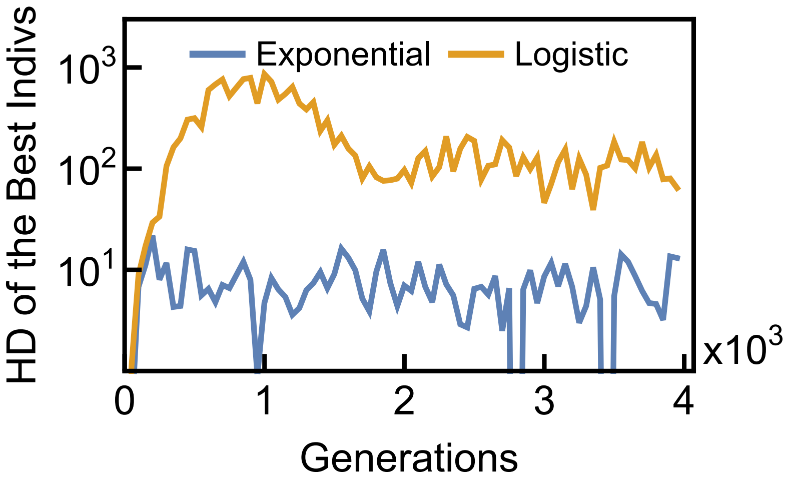


Figure 2.4: The Hamming Distance between the best individuals in two consecutive generations as measured in sites. The population under the logistic fitness function exhibits significant genetic changes (mean=213 mutations), while the population under the exponential fitness function shows smaller genetic variations (mean=6.81 mutations). The plot is truncated as the exponential fitness function reaches a stationary phase.

The infinitesimal model assumes that breeding values are normally distributed. As shown in Fig.2.5, this assumption is valid with the exponential fitness function, which results in low linkage disequilibrium (LD) (Fig.2.3a). This is because the genetic change over generations is small (Fig.2.4 allowing the recombination to break down any initial stochastic LD. Consequently, an individual's fitness is determined by the sum of many independent common alleles, leading to a normal distribution in breeding value. However, because the best individuals are promoted exponentially, the mean of breeding values quickly converges. Hence, the distribution of breeding value is narrow, and the population suffers from the lack of genetic resources.

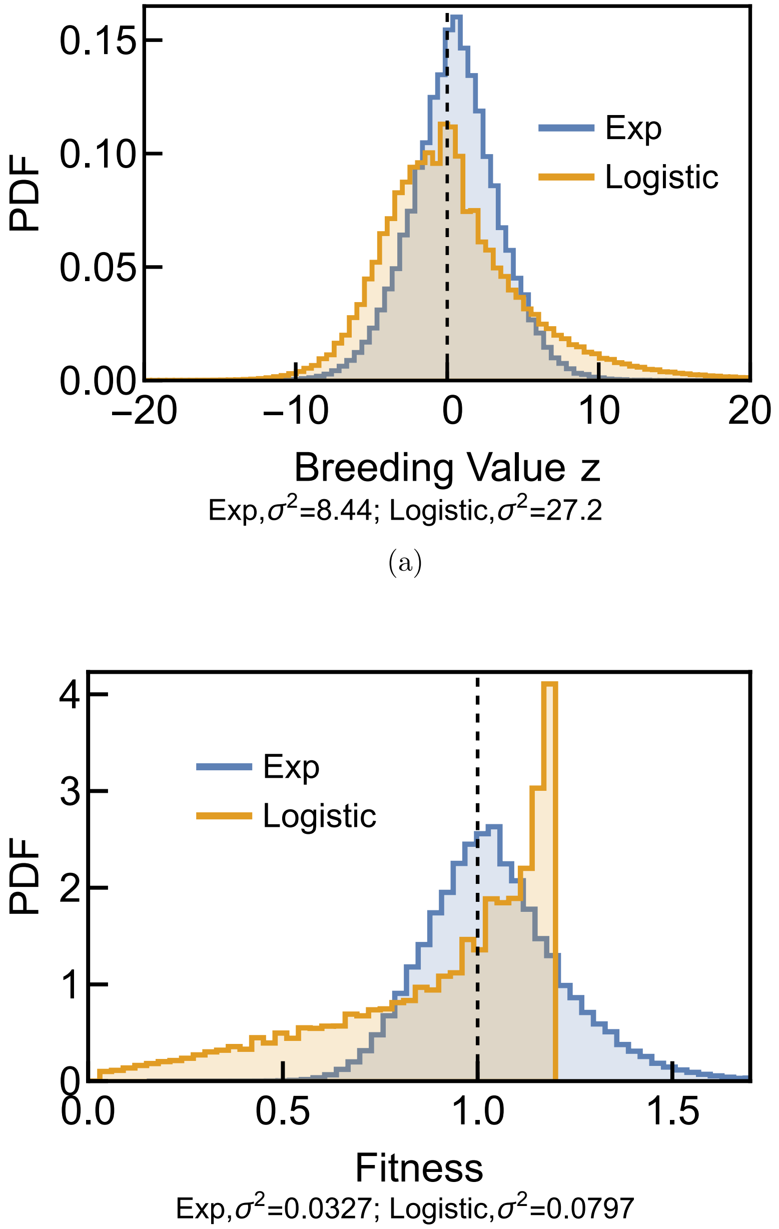
In contrast, the logistic fitness function imposes a fitness cutoff. The fittest individuals are restricted from sweeping the whole population, which allows the less-fit individuals to persist in the population. For this reason, there is a long tail on the right side of the distribution as shown in Fig.2.5a. It also produces the negative LD, as illustrated in Fig.2.3b, indicating the coexistence of multiple genotypes in the population. Accordingly, common alleles are not independent and the combination of them is not Gaussian. As a result, the logistic fitness function presents a right-skewed distribution. Those preserved genetic diversity in the right tail of the distribution provides ample mutations for recombination to create novel genotypes.

2.3.3 Alternative Fitness Functions

We propose to rescale the exponential fitness function into a saturating function by introducing a scale function $f(z)$ that satisfies certain conditions. Specifically, the scale function $f(z)$ should be approximately 1 at $z = 0$ and have an asymptotic behavior of $e^s z$ as z approaches infinity. This allows us to define a new fitness function $W(z) = e^{sz}/f(z)$ that saturates as z becomes large, while remaining linear around 0.

We consider two forms of the scale function: Cauchy and Gaussian scaled. The Cauchy scaled function takes the form $f(z) = 1 + a\frac{s^2z^2}{1+s^2z^2}e^{sz}$, while the Gaussian scaled function is $f(z) = 1 + b(1 - \exp(-s^2z^2/2))e^{sz}$. Parameters a and b are set to 0.35 and 0.5, respectively, to ensure the monotonicity of $W(z)$ and steady increases over z.

Despite all the fitness functions saturating over z, they have different effects on adaptation. The logistic fitness function significantly increases fitness (ratio ~ 2.79) by imposing a strong cut-off on the best individuals. This allows the moderately fit population to experience a sufficient increase in fitness. In contrast, the Cauchy and Gaussian scaled fitness functions are unable to further suppressed individuals on both sides while maintaining monotonicity. In fact, the Cauchy and Gaussian scaled fitness functions can limit the advantage of the fittest individuals substantially. However, the general population is still vulnerable to being swept out by preeminent individuals, which results in only mild changes in fitness (ratio ~ 1.43 and ~ 1.13, respectively). This is shown in the inset of Fig. 2.6, where general individuals have similar fitness to the exponential fitness function.



(a)

(b)

Figure 2.5: The distributions of the breeding value and fitness. The exponential fitness function yields a normally distributed breeding value z, whereas the logistic fitness function produces a right-skewed distribution. Specifically, the logistic fitness function has a wider distribution with a higher maximum value ($z_{max} = 95$ compared to $z_{max} = 23$ for the exponential function). By imposing a fitness cut-off, it accommodates mediocre individuals from those fittest individuals.

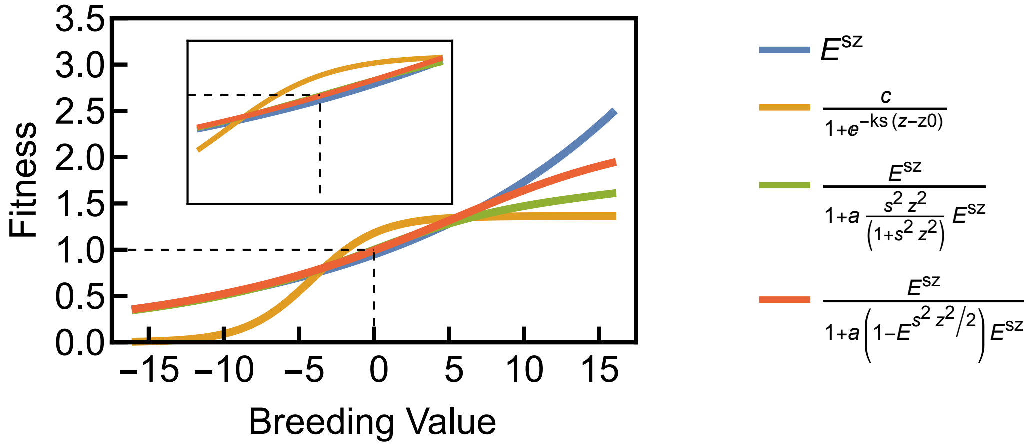


Figure 2.6: The comparison of different fitness functions. The logistic fitness function exhibits a unique fitness increase in mediocre individuals by compromising the fitness of well-adapted individuals. All fitness functions are normalized with mean$= 0$ based on their own distributions of z. The figure is plotted in the range of $3\sigma_z$ (logistic), while the inset is σ_z (logistic). σ_z is obtained in Fig.2.5a.

The goal of the new fitness function is to prevent the dominance of the fittest individuals. Therefore, an ideal sexual population should be effective in preserving genetic diversities and combining them into novel genotypes. As a result, this population should not only maintain a high level of genetic diversity but also ensure that the Hamming distances of the top performers are sufficiently large. Our study has examined four different fitness functions and found that the rate of adaptation is correlated with the strength of promotion imposed on the moderately fit individuals. The abundance of moderately fit individuals improve the efficiency of sexual reproduction by providing distinct lineages, resulting in a bigger genetic changes in the fittest individuals (Supplementary B.4, Fig.B.4 and Fig.B.5). This implies that the degree of heterozygosity and the Hamming distance of the best individuals can be indicators for the effectiveness of sexual reproduction.

2.4 Discussion

2.4.1 Summary

In this paper, we investigate the impact of limiting the evolutionary advantages of outstanding individuals in sexual populations with either weak or strong clonal interference. Our findings reveal that the logistic fitness function can significantly enhance the rate of adaptation compared with the original exponential fitness function when clonal interference is strong. The exponential fitness function exploits the best genotypes to promote the genetic background for future mutations. However, this comes with the cost of eliminating other genotypes with lower fitness when multiple lineages coexist. On the other hand, the logistic fitness function can harbor those moderately fit genotypes, leading to negative linkage disequilibrium and profound genetic diversity. This exceptional genetic diversity can be leveraged by recombination to explore novel genotypes. Those novel genotypes drive the evolution and increase the rate of adaptation by a factor of 2.79. This is consistent with its high genetic changes (~ 200 mutations) in two consecutive generations, while the exponential selection shows a mild transformation of the best genotypes (~ 10 mutations). We also investigate other fitness functions, demonstrating that the faster adaptation is related with the effectiveness in preserving genetic diversity as well as in recombining them. During this process, the moderately fit individuals plays a key role in providing genetic deviations. Our study highlights the significance of genetic diversity in the evolution of a sexual population, where strong selection can slow down the adaptation by sweeping out maladapted individuals.

2.4.2 Short-term and long-term evolution

Initially, the genetic background of a population is homogeneous. The exponential fitness function rapidly promotes the best genotype as the new genetic background at the cost of losing other genotypes. Thus, it produces faster adaptation in the early generations (<400 gens in our simulations), since short-term adaption is sensitive to the strength and precision of the selection [54]. However, when multiple genotypes coexist and the clonal interference becomes stronger, the logistic fitness function becomes more effective in protecting individuals with moderately fit fitness from being eliminated. This condition breaks when dominant genotypes emerge. At this time, the cycle restarts as the genetic background is back to homogeneity to some extent in a period of ~ 1000 generations. When the interaction between clonal interference and recombination reaches equilibrium, the fluctuation damps, and the population enters the stationary phase, which is the phase of focus in this study. During this phase, the effective mutation rate is proportional to the number of unmutated sites. The adaption speed is insensitive to the effective mutation rate, only at the level of the logarithm. However, when half of the sites are mutated, the effective mutation rate is 0. After that, the adaptation slows down as the available sites start to deplete.

In practice, the optimal strategy depends on the subject studied, and the length of the experiments. For microbial experiments, the time scale can range from a few generations (~ 10) to dozens of thousands of generations ($\sim 60,000$), which may cover the whole range of different phases [11, 61, 67]. Even though limiting advantages of preponderant individuals could be potent for long-term evolution, the exponential selection for breeding values can be effective for short-term adaptation. Additionally, the fastest adaptation occurs when clonal interference is intermediate and the logistic selection is present. This phase lasts for a considerable duration, making it optional for mid-range experiments. Overall, the optimal strategy for adaptation depends on the specific research being conducted.

2.4.3 Population subdivision

The population structure interacts with the clonal interference and affects the rate of adaptation. Even though the fixation probability of a beneficial allele can be uninfluenced by population subdivision when each deme contributes proportionally to its size [39, 82, 123], this can change when multiple sweeps occur simultaneously. In asexual populations, the population subdivision would limit the adaptation by preventing the fixation of beneficial alleles in different sub-populations [78, 123]. However, long-range migration and recombination can alleviate local clonal interference [78]. Specifically, recombination can overcome and utilize clonal interference in structured populations [37, 86]. This would allow a structured population adapt faster than a well-mixed population, as discussed in the last chapter.

Population structure acts as a physical barrier that slows down the sweeping process of those ascendant individuals. This is because individuals in different sub-populations are restricted by migration, subject to local genetic background and local competition. Hence, the mechanism of the sexual structured population is similar to limiting the evolutionary advantages of the best individuals, which is to reserve genetic diversity for recombination. This indicates that preserving novel mutations from being swept out by natural selection could be a universal strategy for long-term evolution.

2.4.4 Limits on the higher frequency of sex

Sexual reproduction can cope with genetic diversity and decrease clonal interference by aggregating the mutations into the same genotype. Therefore, the frequency of sexual reproduction is vital to the rate of adaptation. It has been shown that facultative sex would rescale selective coefficients by a factor of $1/r$, resulting in replacing the speed baseline v by v/r^2 [88, 120]. Hence, in order to simulate the unlinked loci, it requires complete sexual reproductions. In our simulations, the frequency of sex remains low ($r = 10^{-3} \sim 10^{-2}$). It is hard to dramatically increase due to estimated memory usage. In order to understand that, when interference is strong, the rate of adaptation $v_0/r^2 \sim 1$, which can also be written as $v/v_0 \sim \frac{(r/s)^2}{NU}$. To maintain other parameters, the population size would increase with r^2 which linearly increases the memory usage. When sex is free, estimated memory usage would be around 2500 times more than the current usage. In addition to this, the mutation accumulation would be even faster and it requires a much larger genome size to enter the stationary phase, which further requires additional demand on memory. Therefore, it is technically difficult to simulate a clonal interference with free recombination.

Chapter 3

Investigating the evolutionary origins of the first three SARS-CoV-2 variants of concern

Abstract

The emergence of Variants of Concern (VOCs) of SARS-CoV-2 with increased transmissibility, immune evasion properties, and virulence poses a great challenge to public health. Despite unprecedented efforts to increase genomic surveillance, fundamental facts about the evolutionary origins of VOCs remain largely unknown. One major uncertainty is whether the VOCs evolved during transmission chains of many acute infections or during long-term infections within single individuals. We test the consistency of these two possible paths with the observed dynamics, focusing on the clustered emergence of the first three VOCs, Alpha, Beta, and Gamma, in late 2020, following a period of relative evolutionary stasis. We consider a range of possible fitness landscapes, in which the VOC phenotypes could be the result of single mutations, multiple mutations that each contribute additively to increasing viral fitness, or epistatic interactions among multiple mutations that do not individually increase

viral fitness—a “fitness plateau”. Our results suggest that the timing and dynamics of the VOC emergence, together with the observed number of mutations in VOC lineages, are in best agreement with the VOC phenotype requiring multiple mutations and VOCs having evolved within single individuals with long-term infections.

3.1 Introduction

For the first 8 months of the SARS-CoV-2 pandemic, the virus exhibited a very slow pace of adaptation, with D614G being the only persistent adaptive substitution that appears to have resulted in increased transmissibility of the virus [55, 99, 132]. However, during the second half of 2020, three designated variants of concern (VOCs) of SARS-CoV-2, Alpha, Beta, and Gamma, emerged independently and in quick succession [1, 35, 113]. No other VOC emerged until Delta and Omicron in 2021 which appear to be very different, both genetically and phenotypically, from the three original VOCs [103, 112]. The VOCs are characterized by a large number of mutations relative to the genetic background from which they first emerged, and exhibit altered phenotypes resulting in varying combinations of increased transmissibility, virulence, and immune evasion [15, 30, 35, 111].

Phylogenetic analyses show that a large number of mutations, mostly located in the spike protein, have independently evolved in multiple lineages of SARS-CoV-2 including the Alpha, Beta and Gamma variants and are likely playing a key role in the adaptive evolution of the SARS-CoV-2 [79, 112]. Experimental measurements and molecular dynamics simulations also show that some of these mutations have synergistic interactions for important functional traits [90, 131], indicating that they may have greater combined fitness benefit to the virus. Some of the distinctive mutations in the VOCs, including the E484K and N501Y mutations found in the first three VOCs, have also been observed in chronic infections such as those in certain immunocompromised individuals [18, 58, 62], suggesting that the VOCs may have arisen from such infections. Some of the other possible explanations for the emergence of VOCs include prolonged circulation of the virus in areas of the world with poor genomic surveillance or reverse-zoonosis from other animals such as rodents followed by sustained transmission and adaptive evolution within the animal population and a spill over back to the humans (see [95] for a recent review on the possible origins of variants of SARS-CoV-2).

While finding the evolutionary process(es) that may have led to the emergence of VOCs has profound consequences for understanding the fate of the SARS-CoV-2 pandemic, there have currently been no systematic investigations to assess the likelihood of any particular evolutionary pathway that would lead to the emergence of VOCs. In this work, we investigate whether the emergence of VOCs was the result of evolution via sustained transmission chains between acutely infected individuals or prolonged infections and evaluate plausible fitness landscapes. We also discuss the potential implications of our results for the future of the pandemic and potential measures that might lower the rate at which new VOCs emerge.

Personal contributions

I have actively participated and made significant contributions throughout the entire duration of this project. My involvement encompasses various stages, including preliminary mathematical analysis, simulation analysis, and the final draft preparation. Specifically, I assume full responsibility for the development of the simulation code, which constituted a substantial portion of the project. Additionally, I actively contributed to the analysis of the results and made substantial contributions to the writing process, including the drafting of sections within the thesis and the manuscript revision.

I affirm that this author contribution disclaimer accurately represents my involvement and contributions to the research project and its subsequent documentation.

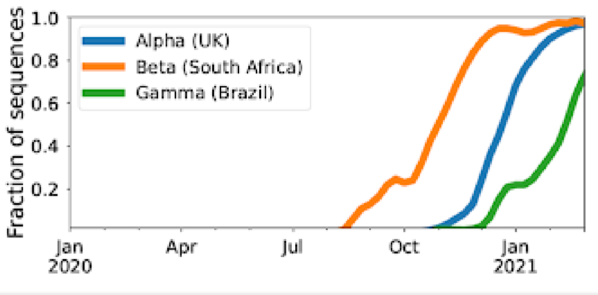
3.2 Results

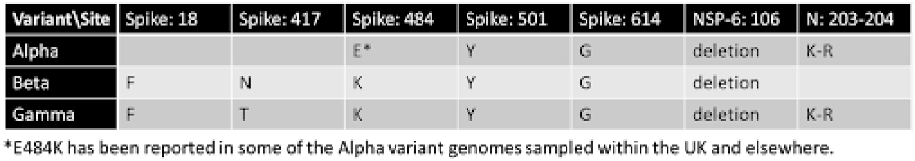
3.2.1 Emergence of VOCs: an evolutionary puzzle

The Alpha, Beta, and Gamma VOCs arose independently and in quick succession, with several shared mutations, in three different countries and began to spread globally (Figure 3.1). This long waiting time followed by clustered emergence of a handful of lineages was not predicted by any simple evolutionary theories. Typically, one would assume that either the beneficial mutation supply is small, in which case one expects a long waiting time for the first VOC but also long gaps before subsequent VOCs, or the mutation supply is large, in which case one expects many VOCs with only a short waiting time [57]. Moreover, each VOC had $> 6 - 10$ mutations distinguishing it from then-dominant genotypes, which was also unexpected. One of the key evolutionary questions is whether VOCs evolved over the course of many acute infections or within single chronic infected hosts. Both possibilities have serious issues. The many-acute-infections hypothesis needs to explain how the virus acquired so many changes, as the mutant lineages would have had to remain at frequencies below the detection threshold in different countries for several months. The chronic-infection hypothesis needs to explain both why adaptation to the within-host environment led to a transmission advantage between hosts, and why there was no 'leakage' of some intermediate mutations at the between-host level before the emergence of the VOCs, i.e., why genotypes with some of the VOC mutations did not escape from the chronically infected patients earlier.

3.2.2 Between-host model of VOC emergence

We assume the effective virus population size is $N_e = N/\sigma^2$ where N is the number of infectious individuals worldwide and σ^2 is the variance in offspring number (secondary cases). We treat each acute infection as one generation, assuming a tight transmis-





Variant\Site	Spike: 18	Spike: 417	Spike: 484	Spike: 501	Spike: 614	NSP-6: 106	N: 203-204
Alpha			E*	Y	G	deletion	K-R
Beta	F	N	K	Y	G	deletion	
Gamma	F	T	K	Y	G	deletion	K-R

*E484K has been reported in some of the Alpha variant genomes sampled within the UK and elsewhere.

Figure 3.1: The three initial Variants of Concern arose in quick succession after a long period of limited adaptation. For each VOC, the curve shows its frequency among the SARS-CoV-2 sequences collected each week from its country of origin. The table shows the amino acid changes across the SARS-CoV-2 genome that are shared between at least two of the three VOCs [112]. *E484K has been detected in some Alpha sequences.

sion bottleneck of a single virion [9, 73, 80]. Viruses mutate at rate μ per base per generation (see Section 3.4). For a mutant virus population with selective advantage s relative to the background, the average number of secondary cases increases by a factor $1 + s$. We also assume that the number of secondary cases approximately follows a negative binomial distribution with mean Rt and dispersion parameter k, so that $\sigma^2 \approx R_t(1+R_t/k)$. There is substantial uncertainty in the amount of overdispersion in the pandemic, and consequently similar uncertainty in the effective population size. Therefore, we consider a range of values for k to see if any would be consistent with the observed dynamics of the VOC emergence. We also note that while the importance of spatial structure is clearly visible in the spatially restricted initial spread of the VOCs from real-world data, we expect that we can neglect it when analyzing their emergence. This is because spatial structure should not have a large impact on viral dynamics until a lineage becomes locally common, and the specific mutations differentiating the VOCs were all locally rare prior to their emergence.

3.2.3 Within-host model of VOC emergence

Unlike tracking the between-host evolution of SARS-CoV-2 where an unprecedented effort has led to huge numbers of consensus genome sequences [50], our current knowledge of the within-host evolutionary dynamics of SARS-CoV-2 is still very limited, particularly in those with chronic infections. Because there is very limited data with which to constrain the within-host evolutionary dynamics of chronic infections with SARS-CoV-2, we simply treat it as a 'black box' and assume with some probability, P_f, that a new infection is chronic and may lead to the production of a VOC (Table 3.1; section 3.4). We also assume that within-host substitutions required for the production of the VOC occur at a constant rate μ_C per generation (see Table 3.1). (Here a generation is still defined as the typical length of an acute infection.) Given that we know only three VOC lineages emerged by late 2020, we expect $T_{obs}NP_f \sim 3$ where $T_{obs} \sim 180 - 317$ days is the expected time to the emergence of the first VOC since the beginning of the pandemic based on phylogenetic estimates (see Table 3.1). Therefore, given the typical variation in the population size throughout the pandemic for biologically relevant parameter combinations $N \sim 1 \times 10^6 - 1 \times 10^7$, we expect that values of $P_f \sim 5 \times 10^{-9} - 1 \times 10^{-7}$ will maximize the likelihood of the within-host model and focus on these.

3.2.4 Fitness landscapes

One possible explanation for the temporal clustering of VOCs with large numbers of mutations is that the underlying fitness landscape may have some structure that causes the dynamics to deviate from our usual expectations. Unfortunately, the full space of possible fitness landscapes is enormous and impossible to explore exhaustively. To investigate the possible effects of the landscape on the dynamics, we therefore focus on three limiting local fitness landscapes that span a range of biologically plausible scenarios (Figure 3.2a). Importantly, these landscapes describe only between-host fitness, which could be very different from within-host fitness. As mentioned above, we treat within-host dynamics implicitly using an effective substitution rate and so do not need an explicit fitness landscape for it. In all three landscapes, the peak is a VOC phenotype with fitness advantage s over the ancestor. We assume that Alpha, Beta, and Gamma are similar enough that they can be approximately described by the same landscape and the same value of s, which we infer from the early rate of increase of the VOCs (see Methods). Landscape 1 is the simplest possibility: a single mutation on the ancestral background is sufficient to confer the full advantage. In Landscape 2, we test whether simply increasing the number of mutations involved can explain the temporal clustering. In this landscape, the VOC phenotype is produced by a combination of $K > 1$ mutations, each providing an independent fitness benefit s/K. In Landscape 3, we test whether epistasis may have an effect: the VOC phenotype again requires K mutations, but we now assume that they provide no fitness benefit until the full combination is acquired, i.e., the population must cross a fitness plateau. As mentioned above, there is experimental evidence for this form of epistasis among the VOC mutations [90, 131]. We expect that shallow fitness valleys will produce similar dynamics to Landscape 3, as will shallow upward slopes with a large jump in fitness at the end [122]. Note that mutations in all the three landscapes can be acquired via the between- or within-host evolutionary pathways (Figure 3.2b).

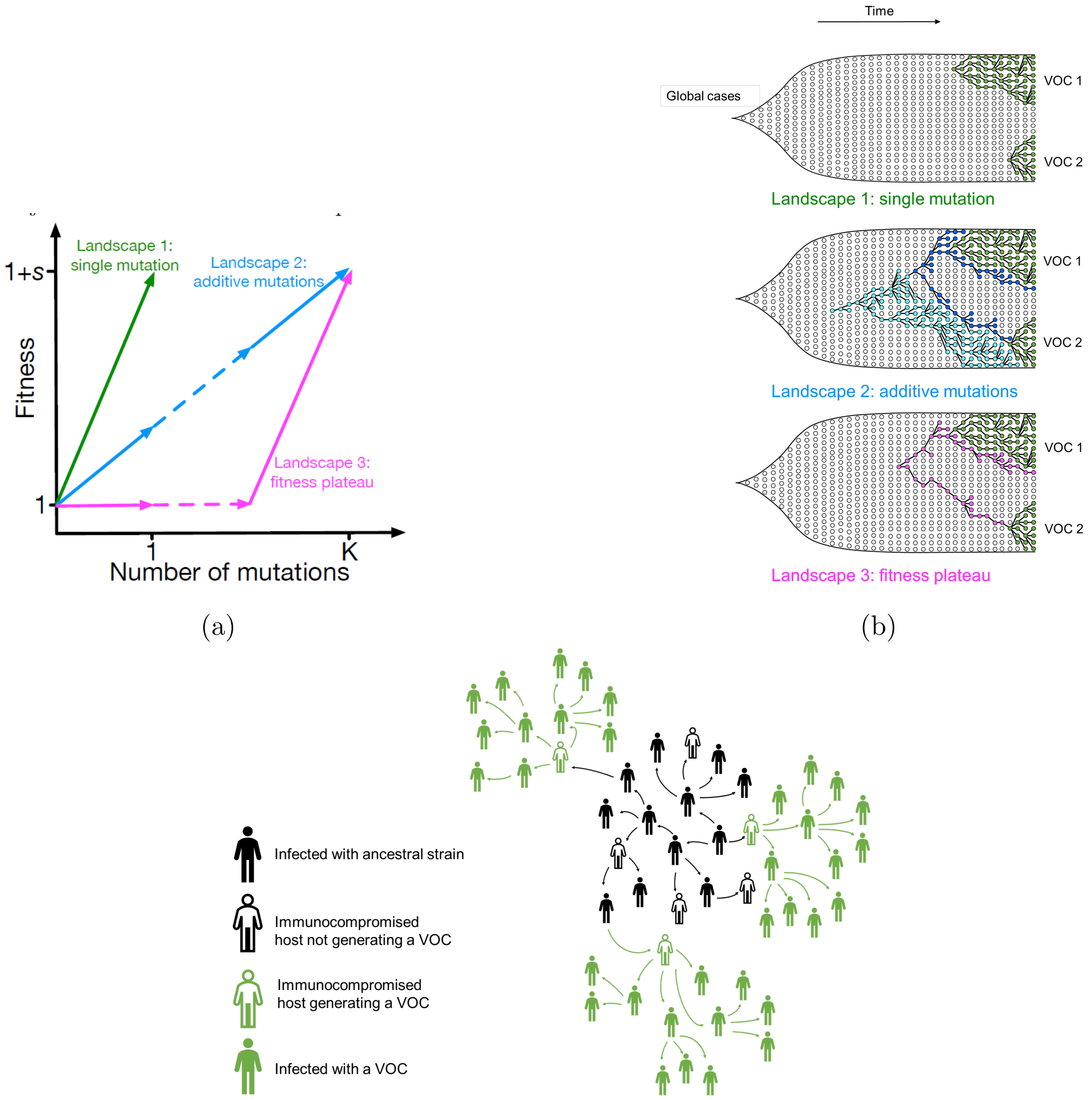
Table 3.1: Model parameters

Symbol	Description	Value(range)	Source
t	Time in units of generations	assuming 5.2 days per generation	[36]
IFR	Global median infection fatality rate of COVID-19	$0.5\%(0.2\% - 1.5\%)$	[68]
N	Number of daily infectious individuals worldwide	daily confirmed deaths / median global IFR	-
μ	Mutation rate per nucleotide per generation	$1.0(0.87 - 2.0) \times 10^{-5}$	[43]
s	Selective advantage of the VOCs	(0.3 - 1.1)	
k	Dispersion in distribution of number of secondary infections	0.1 (0.05 - 0.2)	[34]
T_{obs}	Time to the emergence of the first VOC (number of days since 2020-01-03)	(180 - 317) days	[1, 35, 74, 113]
ΔT_{obs}	Time between the emergence of the first and second VOC	(0 - 137) days	[1, 35, 74, 113]
P_f	Probability of a chronic SARS-CoV-2 infection in an ICI producing a VOC	–	–
μ_c	Within-host fixation rate of VOC mutations per generation	–	–

For each evolutionary scenario, we test whether there are parameter values consistent with the data on the timing of the emergence of Alpha, Beta, and Gamma variants of SARS-CoV-2 (see Section 3.4; Table 3.1). For these parameter values, we further investigate whether they correspond to biologically reasonable scenarios in terms of the frequencies of the intermediate mutations prior to the emergence of VOCs, total number of mutations required to produce VOCs, total number of successful VOC lineages produced over time, and the timing between the emergence of different VOC lineages.

3.2.5 Landscape 1: Single mutations

We start with the simplest possible fitness landscape, in which a single mutation conferring a fitness advantage s relative to the genetic background of circulating lineages is required for the emergence of VOCs. We first consider the between-host evolutionary pathway. As long as the effective population size of the pandemic was not much smaller than the census size (i.e., overdispersion was not too large), the mutation supply $N_e\mu$ became large early in 2020. At this point, numerous lineages would have emerged over a short period of time (see the $k = 0.2$ scenario in Figure 3.3a), inconsistent with the observed dynamics. We can therefore rule out this scenario.

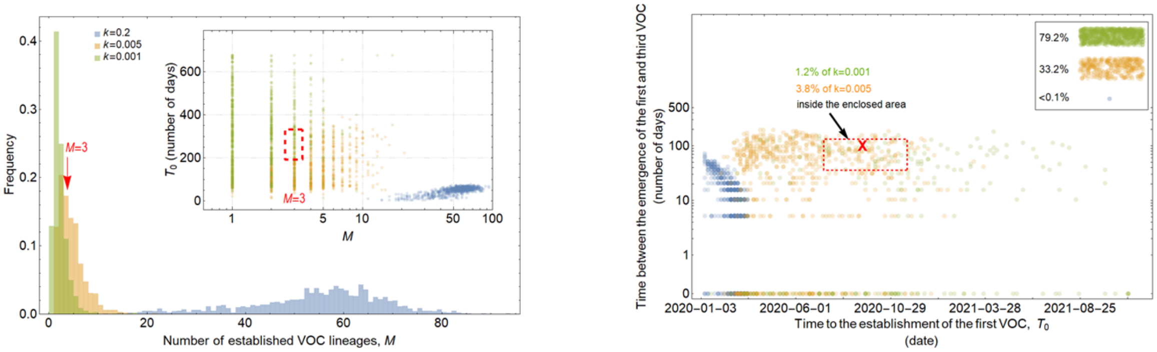


(a)

(b)

(c)

Figure 3.2: Possible evolutionary pathways to the emergence of SARS-CoV-2 VOCs. (**a**, left) The three limiting fitness landscapes for the emergence of VOCs as a function of the relevant number of mutations required, K. (**a**, right) VOCs can emerge from either a single advantageous mutation (green) or multiple mutations that each contribute independently to increasing fitness (blue) or only in combination (magenta). **(b).** Emergence of VOCs via the within-host evolutionary path such that an infectious individual passes on a wild-type variant of the virus to an immunocompromised individual where the virus may acquire the relevant mutations during the chronic phase of the infection and later be passed on to the rest of the population.



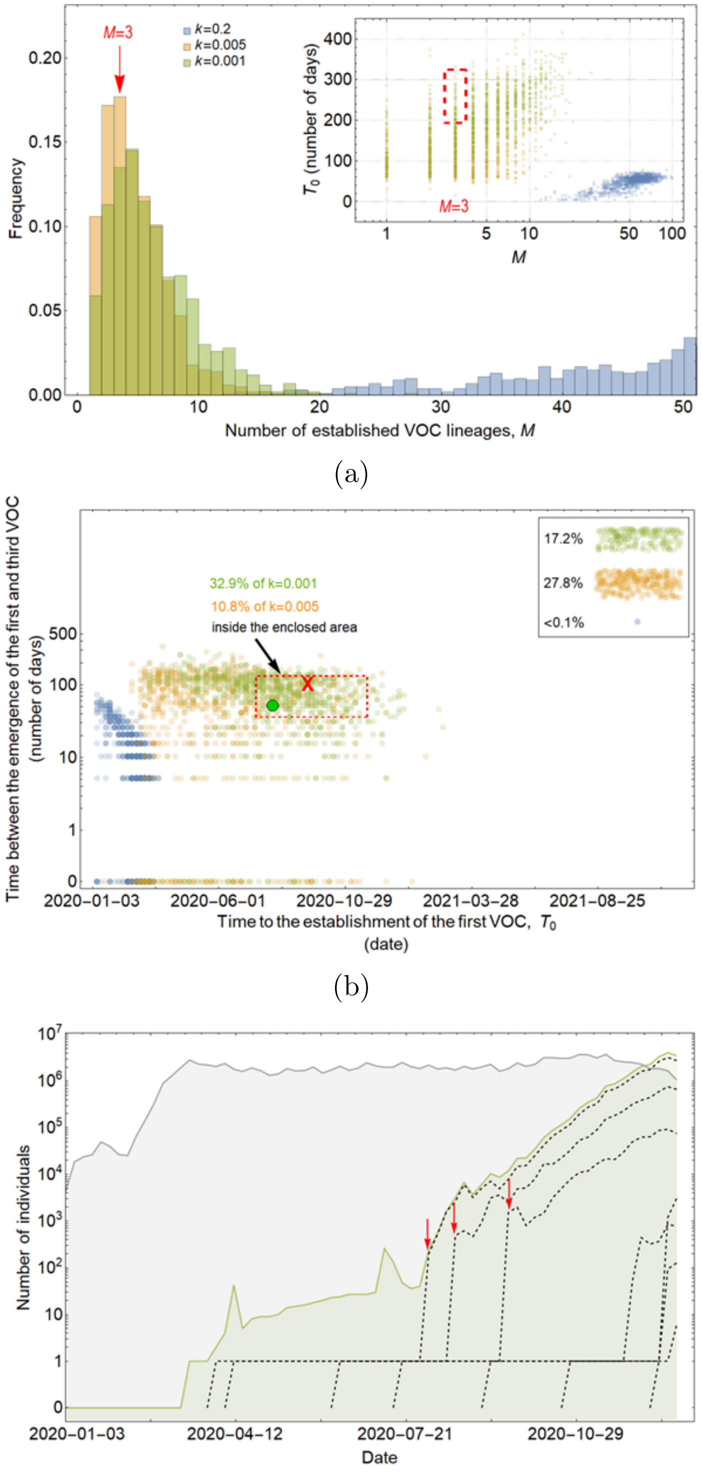
(a)

(b)

Figure 3.3: Evolution between hosts on a single-mutation landscape ($K = 1$) rarely reproduces the observed VOC dynamics, even with extreme overdispersion. **(a).** Total number of established VOC lineages (M) measured under varied levels of overdispersion, k, such that IFR= 1.5%, $\mu = 0.87 \times 10^{-5}$, $K = 1$, and $s = 0.4$. The inset shows M with respect to the waiting time for the establishment of the first VOC lineage since the start of the pandemic, T_0. The region corresponding to the waiting time for the emergence of the first three SARS-CoV-2 VOC is highlighted in red. Under low levels of overdispersion (blue, $k = 0.2$), too many VOC lineages are produced very early on in the pandemic. On the other hand, as we increase overdispersion (orange and green), fewer VOCs can establish in the population. It also takes them much longer to establish and reach high frequencies in the population. **(b).** Evaluating the temporal clustering of the first three VOC lineages. For each simulation run, represented by a point on the graph, we measure T_0 and the time difference between the establishment of the first and third successful VOC lineages. The red dashed rectangle shows the region corresponding to the emergence of the first three SARS-CoV-2 VOCs with the cross sign ("X") representing the mean value. We see that as the level of overdispersion increases, the emergence time of VOCs are more scattered and rarely exhibit temporal clustering in late 2020 – Only 9.1% and 2.9% of the evolutionary dynamics corresponding to overdispersion $k = 0.005$ and $k = 0.001$ fall inside the enclosed area, respectively. The inset shows that 33.2% and 79.2% of the runs for $k = 0.005$ and $k = 0.001$ scenarios produce fewer than three successful VOC lineages by the end of the simulation period. Each run stops once the frequency of the VOC population reaches 75%. See also Appendix Figure C.1.

If overdispersion were very large, it could have kept $N_e\mu$ low through the establishment of the VOCs (see the $k = 0.005$ and 0.001 scenarios in Figure 3.3a). Figure 3.3a shows that under extremely high levels of overdispersion ($k = 0.005$ and 0.001) this model can match the long waiting time for the emergence of the first VOC. However, such high levels of overdispersion are not supported by any existing epidemiological studies on SARS-CoV-2 transmission [34]. Moreover, Figure 3.3b shows that this model rarely produces an evolutionary dynamics that would fit the joint waiting time distribution for all three VOCs (also see Appendix Figure C.1). Under these mutation-limited conditions, there is an approximately exponential waiting time for the arrival of each VOC lineage (once we reach the point where COVID-19 becomes a pandemic in March 2020). Thus, it predicts similarly long waiting times for the emergence of Alpha, Beta, and Gamma, inconsistent with the observed temporal clustering. Therefore, there is no biologically reasonable combination of parameters that result in the clustered emergence of VOCs in late 2020 via the Landscape 1 between-host evolutionary pathway.

On the other hand, if VOCs arose from chronic infections, then their emergence was a two-step process: first, chronic infections had to occur, and then the VOC mutation had to arise in them. The waiting time for the first step is determined by NP_f; note that the number of chronic infections depends on the census size N rather than N_e, i.e., it is insensitive to the amount of overdispersion. The second step follows an exponential distribution within each chronic host, with rate μ_C. The third step, the spread of the VOC from the original chronic host to the rest of the population, then takes much less time than the first two. Figure 3.4 shows that to match observed VOC dynamics we must assume that the level of overdispersion is very high (i.e., very low mutation supply, $N_e\mu$), effectively blocking the between-host evolutionary pathway, while simultaneously assuming that chronic infections are very frequently produced in the population (i.e., $NP_f \sim 1$) and that there is a relatively long waiting time before the production of each VOC mutation ($\mu_C \sim 0.01$). However, like the between-host pathway, this scenario requires very high levels of overdispersion which makes the Landscape 1 within-host evolutionary pathway also an unlikely explanation for the emergence of VOCs (see Appendix Figure C.2).



M=3
k=0.2
k=0.005
k=0.001
Frequency
0.00
0.05
0.10
0.15
0.20
0
10
20
30
40
50
Number of established VOC lineages, M
T0 (number of days)
0
100
200
300
400
1
5
10
50
100
M
M=3
Time between the emergence of the first and third VOC (number of days)
0
1
10
100
500
32.9% of k=0.001
10.8% of k=0.005
inside the enclosed area
X
17.2%
27.8%
<0.1%
2020−01−03
2020−06−01
2020−10−29
2021−03−28
2021−08−25
Time to the establishment of the first VOC, T0 (date)
Number of individuals
0
1
10
10^2
10^3
10^4
10^5
10^6
10^7
2020−01−03
2020−04−12
2020−07−21
2020−10−29
Date

(a)

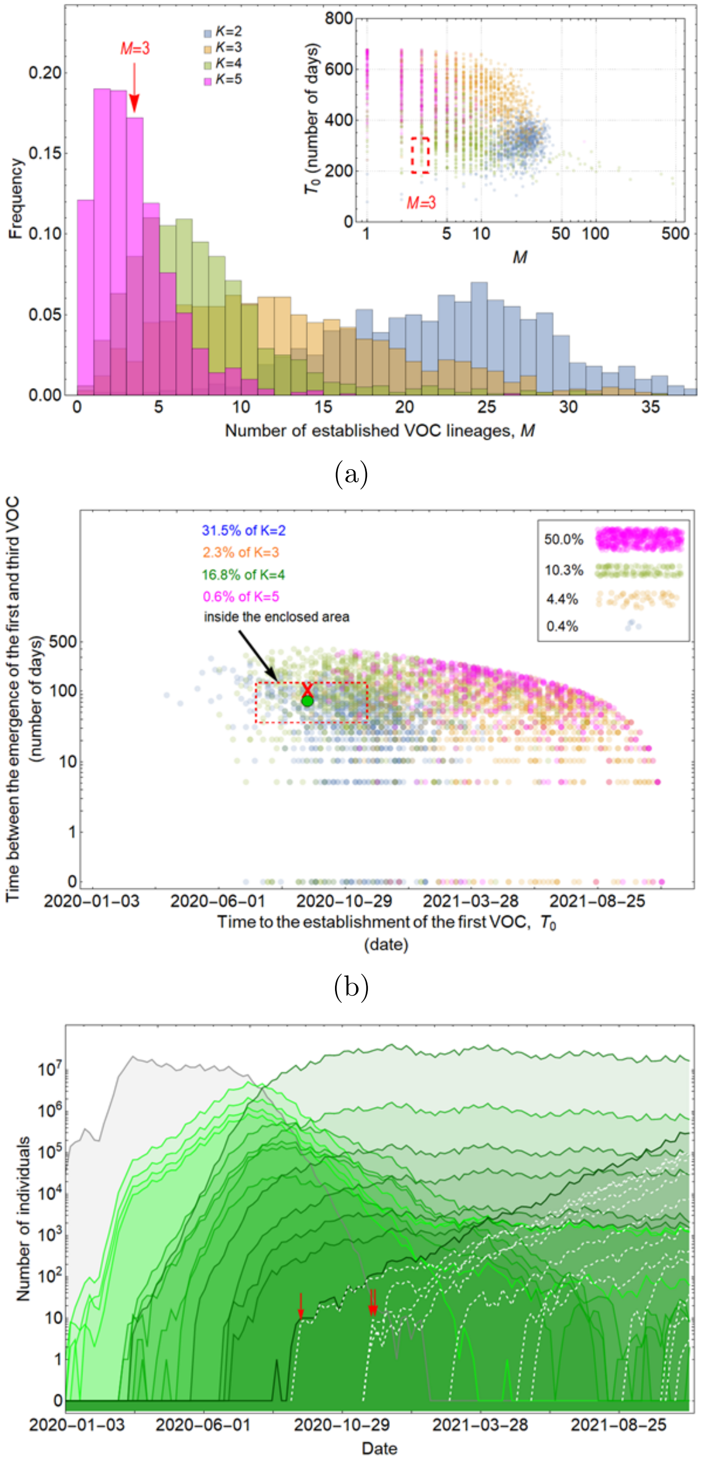
(b)

(c)

Figure 3.4: Evolution within hosts on a single-mutation ($K = 1$) landscape can match the observed VOC dynamics, but only with extreme overdispersion to prevent between-host evolution. **(a).** Total number of established VOC lineages (M) measured under varied levels of overdispersion, k, where the within-host parameters for the $k = 0.2$ scenario (blue) are $P_f = 5 \times 10^{-10}$ and $\mu_C = 0.001$. For the $k = 0.005$ scenario (orange), $P_f = 6 \times 10^{-8}$ and $\mu_C = 0.1$. Finally, for the $k = 0.001$ scenario (green), $P_f = 6 \times 10^{-6}$ and $\mu_C = 0.01$. For all the three scenarios, the between-host parameters $\mu = 0.87 \times 10^{-5}$, IFR$= 1.5\%$, $K = 1$, and $s = 0.4$ are the same. The inset shows M with respect to the waiting time for the establishment of the first VOC lineage since the start of the pandemic, T_0. The region corresponding to the waiting time for the emergence of the first three SARS-CoV-2 VOC is highlighted in red. We see that under low levels of overdispersion, $k = 0.2$ (blue), too many VOC lineages are produced very early on in the pandemic. On the other hand, as we increase overdispersion (orange and green), fewer VOC lineages can establish in the population, and it generally takes longer for them to do so. **(b).** Evaluating the temporal clustering of the first three VOC lineages. For each simulation run, represented by a point on the graph, we measure the time that it takes for a single adaptive mutation to establish in the population and the time difference between the establishment of the first and third successful VOC lineage. The red dashed rectangle shows the region of the parameter space corresponding to the emergence of the first three SARS-CoV-2 VOCs with the cross sign ("X") representing the mean value. The graph shows that by increasing the level of overdispersion and lowering the evolutionary contribution from the between-host pathway, multiple VOCs can emerge in quick succession via the within-host pathway such that a larger fraction of the simulation runs yield the correct timing for the emergence of the first three VOCs in late 2020 (i.e., they fall inside the enclosed area). The inset shows that 27.8% and 17.2% of the runs for $k = 0.005$ and $k = 0.001$ scenarios produce fewer than three successful VOC lineages by the end of the simulation period. Each run stops once the frequency of the VOC population reaches 75%. **(c).** A typical evolutionary trajectory corresponding to the $k = 0.001$ scenario (green) highlighted with a bold green circle in panel **(b).**. The graph shows the VOC population (green) along with the individual VOC lineages (black dashed lines) emerging from the background population (gray). Red vertical arrows show the establishment time of the first three VOCs. We see that the VOC mutation is first produced in a single individual within the population (chronically infected case) for a relatively long time before successfully spreading to the rest of the population. See also Appendix Figure C.2.

3.2.6 Landscape 2: Additive mutations

Landscape 2 corresponds to an evolutionary pathway in which there were $K > 1$ major mutations involved in the emergence of VOCs, each making an additive contribution of $\approx s/K$ to fitness. If evolution occurred at the whole-population level, Figure 3.5 and Appendix Figure C.3 show that, for a range of parameter combinations, the additive fitness landscape requiring up to four mutations can create evolutionary dynamics with appropriately long waiting times before the arrival of the first successful VOC lineage, while for combinations of more than four mutations, VOC lineages do not emerge by late 2020 under any biologically reasonable parameter combinations for effective population size, mutation rate, and selective coefficient. However, while $K \leq 4$ can match the observed waiting for the first VOC lineage, for the $K = 2$ and 3, this first VOC is usually followed by the establishment of nearly a dozen VOC lineages that emerge in quick succession (see $K = 2$ and 3 scenarios in Figure 3.5a; also see Appendix Figure C.3), inconsistent with the observation of only three VOC lineages emerging in late 2020. However, while for $K = 4$ fewer VOC lineages are produced, a closer examination of a typical evolutionary trajectory that matches the long waiting time before the establishment of the first VOC further reveals that the intermediate single-, double-, or triple-mutants reach high frequencies before the emergence of the first successful (quadruple-mutant) VOC lineage (Figure 3.5c). The sequential fixation of adaptive mutations at the population level would imply that the intermediate mutations were detectable many months prior to the emergence of VOCs, again inconsistent with the genomic surveillance data from around the world. The inconsistency is also visible phylogenetically. The sequential fixation dynamics predicted by the model create a ladder-like phylogenetic relationship between the background and mutant populations whereby every new VOC mutation becomes dominant in the population before giving rise to lineages with additional mutations. Even though such phylogenetic relationships may emerge in SARS-CoV-2 over longer evolutionary timescales [as have been observed in human coronaviruses [33]], they do not resemble the observed topology of the phylogeny of the VOCs of SARS-CoV-2, which is more star-like.


M=3
K=2
K=3
K=4
K=5
Frequency
0.00
0.05
0.10
0.15
0.20
0
5
10
15
20
25
30
35
Number of established VOC lineages, M
T_0 (number of days)
0
200
400
600
800
1
5
10
50
100
500
M
M=3
31.5% of K=2
2.3% of K=3
16.8% of K=4
0.6% of K=5
inside the enclosed area
50.0%
10.3%
4.4%
0.4%
Time between the emergence of the first and third VOC (number of days)
0
1
10
100
500
2020-01-03
2020-06-01
2020-10-29
2021-03-28
2021-08-25
Time to the establishment of the first VOC, T_0 (date)
Number of individuals
0
1
10
10^2
10^3
10^4
10^5
10^6
10^7
2020-01-03
2020-06-01
2020-10-29
2021-03-28
2021-08-25
Date

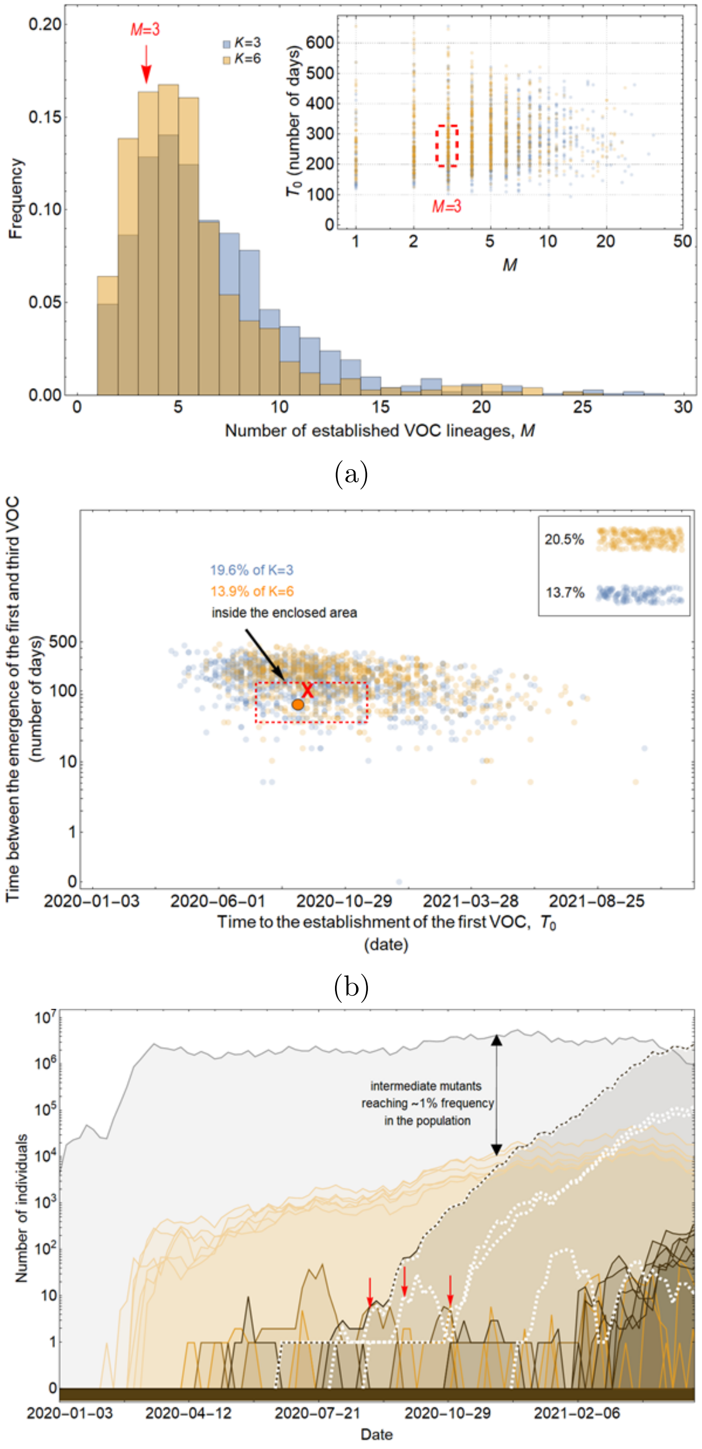
(a)

(b)

(c)

Figure 3.5: Between-host evolution on an additive fitness landscape can match the observed VOC dynamics, but only by having intermediate mutants reach unrealistically high frequencies. **(a).** Total number of established VOC lineages (M) for different number of mutations, K, involved in the production of a VOC. For the $K = 2$ scenario (blue), IFR= 1.5%, $\mu = 0.87 \times 10^{-5}$, $k = 0.05$, and $s = 0.3$. For the $K = 3$ scenario (orange), IFR= 0.5%, $\mu = 0.87 \times 10^{-5}$, $k = 0.05$, and $s = 0.5$. For both the $K = 4$ (green) and $K = 5$ (magenta) scenarios, IFR= 0.2%, $\mu = 2 \times 10^{-5}$, $k = 0.2$, and $s = 1.0$. The inset shows M with respect to the waiting time for the establishment of the first VOC lineage since the start of the pandemic, T_0. Under $K = 2$ and 3, a very large number of successful VOC lineages are produced by late 2020, with the $K = 2$ scenario producing, on average, more than 20 VOC lineages that establish in the population. On the other hand, for the $K = 5$ scenario, on average, fewer than three lineages are produced. It also takes much longer for them to establish in the population. **(b).** Evaluating the temporal clustering of the first three VOC lineages. For each simulation run, represented by a point on the graph, we measure the time that it takes for a single adaptive mutation to establish in the population and the time difference between the establishment of the first and third successful VOC lineage. The red dashed rectangle shows the region of the parameter space corresponding to the emergence of the first three SARS-CoV-2 VOCs with the cross sign ("X") representing the mean value. We see a noticeable overlap between the $K = 2$ and 4 scenarios and the red rectangle suggesting that a larger fraction of the simulation runs exhibit temporal clustering dynamics for VOC emergence. The inset shows that 10.3% of the runs for the $K = 4$ scenario produce fewer than three successful VOC lineages by the end of the simulation period. Each run stops once the frequency of the VOC population reaches 75%. **(a).** A typical evolutionary trajectory corresponding to the $K = 4$ scenario highlighted with a bold green circle in panel **(b).**. The graph shows the background population in gray and the i-mutant populations ($1 < i \leq K$) in different shades of green from light (fewer mutations) to dark (more mutations). Note that for the $K = 4$ scenario, there are four single-mutant, six double-mutant, four triple-mutant, and one quadruple-mutant genotypes. The dashed lines show the dynamics of all the established VOC lineages over time. Red vertical arrows show the establishment time of the first three VOCs. We can see that some of the intermediate mutant genotypes reach close to fixation before giving rise to the VOC population. See also Appendix Figure C.3.

For the chronic-infection pathway, on the other hand, the intermediate mutants could have fixed within the host while remaining at undetectable frequencies at the between-host level until the production of the VOCs. Figure 3.6 shows that for a combination of parameters requiring K=3 and 6 mutations where the mutation supply is low and the strength of selection is relatively weak such that the intermediate mutants cannot reach fixation before the emergence of the VOC population, the Landscape 2 within-host pathway can lead to the clustered emergence of a few VOC lineages by late 2020. However, if the selective coefficient s/K on single mutants is too high, they will reach observable frequencies before the VOCs emerge, as we discussed above with the between-host pathway. Effectively, this means that there is a minimal K of at least 3 needed so that the strength of selection on each mutant allele is not too strong. Alternatively, lower K is possible but requires extremely large overdispersion, as in the $K = 1$ case.


M=3
K=3
K=6
Frequency
Number of established VOC lineages, M
T_0 (number of days)
M
M=3
20.5%
13.7%
19.6% of K=3
13.9% of K=6
inside the enclosed area
Time between the emergence of the first and third VOC (number of days)
Time to the establishment of the first VOC, T_0 (date)
2020–01–03
2020–06–01
2020–10–29
2021–03–28
2021–08–25
intermediate mutants reaching ~1% frequency in the population
Number of individuals
Date
2020–01–03
2020–04–12
2020–07–21
2020–10–29
2021–02–06

(a)

(b)

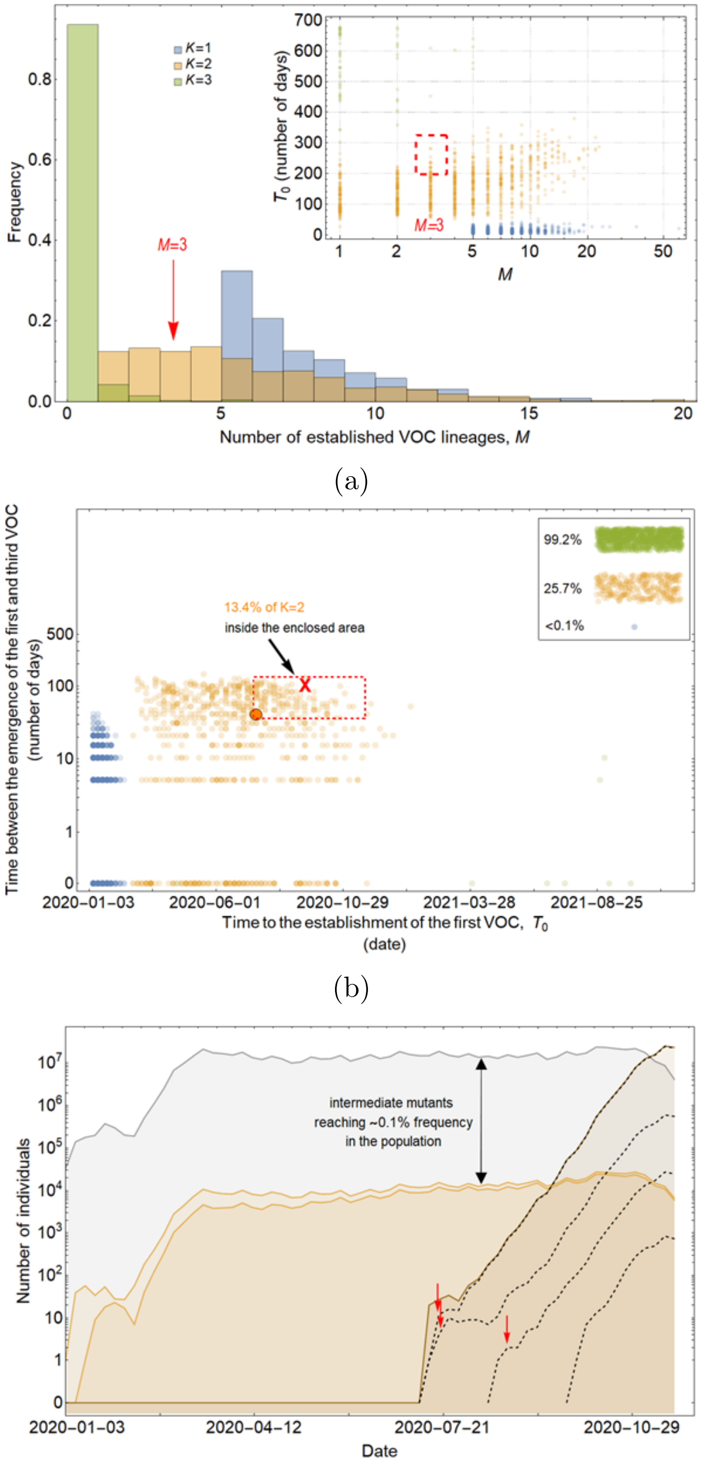
(c)

Figure 3.6: Evolution within hosts on an additive fitness landscape can match the observed VOC dynamics as long as K is large enough that between-host evolution is ineffective. **(a).** Total number of established VOC lineages (M for different number of mutations, K, involved in the production of a VOC. For $K = 3$ (blue), the within-host parameters are $P_f = 3.5 \times 10^{-8}$, and $\mu_C = 0.15$. For $K = 6$ (orange), $P_f = 3 \times 10^{-8}$, and $\mu_C = 0.3$. In both scenarios, the between-host parameters $\mu = 0.87 \times 10^{-5}$, IFR= 1.5%, $k = 0.05$, and $s = 0.3$ are the same. The inset shows M with respect to the waiting time for the establishment of the first VOC lineage since the start of the pandemic, T_0. The region corresponding to the waiting time for the emergence of the first three SARS-CoV-2 VOC is highlighted in red. Both scenarios produce roughly the same of number of VOC lineages. However, on average, T_0 is slightly longer for the K=6 scenario. **(b).** Evaluating the temporal clustering of the first three VOC lineages. For each simulation run, represented by a point on the graph, we measure the time that it takes for a single adaptive mutation to establish in the population and the time difference between the establishment of the first and third successful VOC lineage. The red dashed rectangle shows the region of the parameter space corresponding to the emergence of the first three SARS-CoV-2 VOCs with the cross sign ("X") representing the mean value. We can see that by having a combination of relatively high level of overdispersion, high IFR, and low between-host mutation rate, there is a lower chance of intermediate mutations reaching fixation via the between-host path. Instead, multiple VOCs can emerge in quick succession during chronic infections such that a relatively large fraction of the simulation runs yield a temporal clustering that matches the emergence of the first three VOCs in late 2020 (i.e., they fall inside the enclosed area). The inset shows that 20.5% and 13.7% of the runs for $K = 3$ and 6 scenarios produce fewer than three successful VOC lineages by the end of the simulation period, respectively. Each run stops once the frequency of the VOC population reaches 75%. **(c).** A typical evolutionary trajectory corresponding to the K=6 scenario highlighted with a bold orange circle in panel **(b).**. The graph shows the background population in gray and the i-mutant populations ($1 < i \leq K$) in different shades of green from light (fewer mutations) to dark (more mutations). The dashed lines show the dynamics of all the established VOC lineages over time. Red vertical arrows show the establishment time of the first three VOCs. We can see that the single-mutant genotypes (lines in light orange) are produced via the between-host pathway but never reach above 1% prevalence before the emergence of the VOCs (white dashed lines). See also Appendix Figure C.4.

3.2.7 Landscape 3: Fitness plateau crossing

As in Landscape 2, Landscape 3 describes an evolutionary pathway where there are $K > 1$ major mutations involved in the generation of VOCs, but in this case, only the full K-mutant VOC genotype has a substantial selective advantage relative to the background population, while the selective advantages of the intermediate genotypes are negligible. This does not necessarily imply that the selective coefficients of the intermediate genotypes are small in the standard weak-selection sense (small relative to $1/N_e$), but only that they are too small to substantially affect the dynamics of the production of the first successful K-mutant VOC lineage, a weaker condition that depends on the mutation rate [122].

For the between-host model of VOC emergence, our analysis suggests that only a plateau-crossing of size $K = 2$ may be consistent with the timing of the emergence of SARS-CoV-2 VOCs (Figure 3.7; Appendix Figure C.5). Extended plateaus requiring $K > 2$ mutations take much longer to cross and for most parameter combinations either zero or one VOC lineage is produced before the end of 2020 (Figure 3.7a). For a typical $K = 2$ plateau-crossing trajectory, single-mutant genotypes grow linearly over time and reach a frequency of $\lessapprox 0.1\%$ before producing $\sim 1 - 5$ successful VOC lineages that emerge in quick succession (Figure 3.7c). Therefore, unlike the between-host evolutionary pathway in Landscape 2, a fitness plateau could have led to the clustered emergence of several VOCs after a long waiting time during which none of the intermediate mutations reached high frequency. However, the fact that for biologically plausible parameter values only a narrow plateau of $K = 2$ mutations can be crossed seems inconsistent with the high number of mutations found in the VOCs and particularly with the high number of similar mutations shared across unrelated VOC lineages. This inconsistency may be partly reconciled with the possibility of compounded evolutionary effects following the plateau-crossing event such as the emergence of hyper-mutability traits across certain sites or strong within-host selection following the acquisition of the K mutations.


K=1
K=2
K=3
Frequency
0.0
0.2
0.4
0.6
0.8
M=3
0
5
10
15
20
Number of established VOC lineages, M
700
600
500
400
300
200
100
0
T0 (number of days)
M=3
1
2
5
10
20
50
M
Time between the emergence of the first and third VOC (number of days)
99.2%
25.7%
<0.1%
13.4% of K=2
inside the enclosed area
X
500
100
10
1
0
2020-01-03
2020-06-01
2020-10-29
2021-03-28
2021-08-25
Time to the establishment of the first VOC, T0 (date)
Number of individuals
10^7
10^6
10^5
10^4
10^3
10^2
10
1
0
intermediate mutants
reaching ~0.1% frequency
in the population
2020-01-03
2020-04-12
2020-07-21
2020-10-29
Date

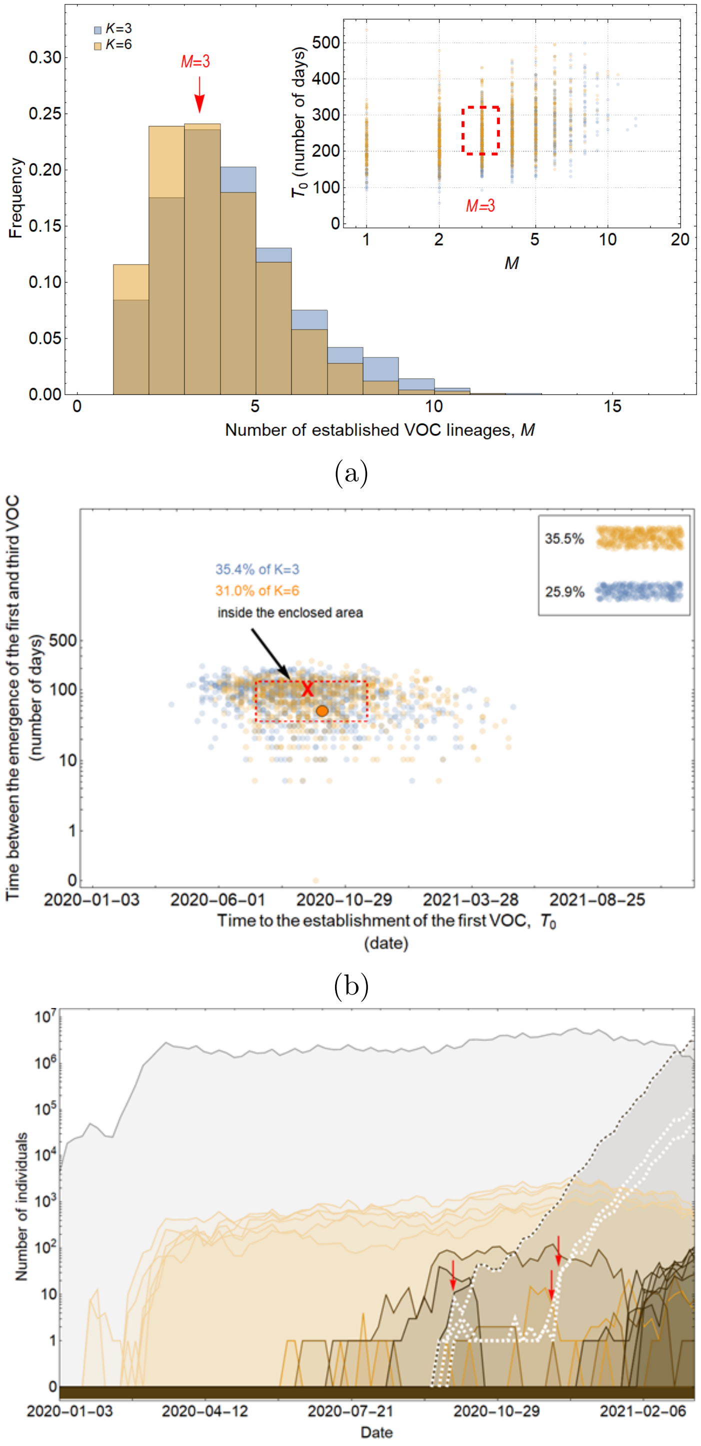
(a)

(b)

(c)

Figure 3.7: Between-host evolution on a fitness plateau can match the observed VOC dynamics, but only for $K = 2$. **(a).** Total number of established VOC lineages (M) for different number of mutations, K, involved in the production of a VOC, such that IFR= 0.2%, $\mu = 2 \times 10^{-5}$, $k = 0.2$, and $s = 1.0$. The inset shows M with respect to the waiting time for the establishment of the first VOC lineage since the start of the pandemic, T_0. For $K = 1$ and 3 scenarios, there are too many and too few VOC lineages are produced by late 2020. Only for the $K = 2$ scenario we can see an intermediate number of VOC lineages being produced in the right time span. **(b).** Evaluating the temporal clustering of the first three VOC lineages. For each simulation run, represented by a point on the graph, we measure the time that it takes for a single adaptive mutation to establish in the population and the time difference between the establishment of the first and third successful VOC lineage. The red dashed rectangle shows the region of the parameter space corresponding to the emergence of the first three SARS-CoV-2 VOCs with the cross sign ("X") representing the mean value. We see a noticeable overlap between the $K = 2$ scenario and the red rectangle suggesting that a fraction of the simulation runs exhibit temporal clustering dynamics for VOC emergence. The inset shows that 99.2% and 25.7% of the runs for the $K = 3$ and two scenarios produce fewer than three successful VOC lineages by the end of the simulation period. Each run stops once the frequency of the VOC population reaches 75%. **(c).** A typical evolutionary trajectory corresponding to the $K = 6$ scenario highlighted with a bold orange circle in panel **(b).**. The graph shows the background population in gray, single-mutants in light orange, and double-mutants in dark orange. Note that for the $K = 2$ scenario, there are two single-mutant and one double-mutant genotypes. The dashed lines show the dynamics of all the established VOC lineages over time. Red vertical arrows show the establishment time of the first three VOCs. We can see that the single-mutant genotypes reach close to 0.1% before giving rise to the VOC population. See also Appendix Figure C.5.

If the VOCs arose from chronic infections, the intermediate VOC mutations (which are neutral at the between-host level of selection but may be selected within-host) can rapidly fix within a host, allowing much wider plateaus to be crossed compared to the between-host evolutionary pathway. Unlike Landscape 2 within-host pathway, the early leakage of intermediate mutations to the population is much less likely as they have no strong selective advantage over the background population. Figure 3.88 shows that the within-host evolutionary pathway of Landscape 3 creates evolutionary trajectories that are consistent with the clustered emergence of ~ 3 VOCs in late 2020. There is also less seeding of new chronic infections with intermediate mutations, leading to fewer VOC lineages compared to Landscape 2 (also see Appendix Figure C.6).


K=3
K=6
M=3
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Frequency
0
5
10
15
Number of established VOC lineages, M
T0 (number of days)
500
400
300
200
100
0
M=3
1
2
5
10
20
M
35.4% of K=3
31.0% of K=6
inside the enclosed area
35.5%
25.9%
X
Time between the emergence of the first and third VOC (number of days)
500
100
10
1
0
2020-01-03
2020-06-01
2020-10-29
2021-03-28
2021-08-25
Time to the establishment of the first VOC, T0 (date)
Number of individuals
10^7
10^6
10^5
10^4
10^3
10^2
10
1
0
2020-01-03
2020-04-12
2020-07-21
2020-10-29
2021-02-06
Date

(a)

(b)

(c)

Figure 3.8: Within-host evolution on a fitness plateau can match the observed VOC dynamics for a large range of plateau widths. **(a).** Total number of established VOC lineages (M for different number of mutations, K, involved in the production of a VOC. For $K = 3$ (blue), the within-host parameters are $P_f = 2 \times 10^{-8}$, and $\mu_C = 0.1$. For $K = 6$ (orange), $P_f = 4.5 \times 10^{-8}$, and $\mu_C = 0.25$. In both scenarios, the between-host parameters $\mu = 1 \times 10^{-5}$, IFR= 0.5%, $k = 0.1$, and $s = 0.7$ are the same. The inset shows M with respect to the waiting time for the establishment of the first VOC lineage since the start of the pandemic, T_0. The region corresponding to the waiting time for the emergence of the first three SARS-CoV-2 VOC is highlighted in red. Both scenarios produce roughly the same of number of VOC lineages. However, on average, T_0 is slightly longer for the $K = 6$ scenario. **(b).** Evaluating the temporal clustering of the first three VOC lineages. For each simulation run, represented by a point on the graph, we measure the time that it takes for a single adaptive mutation to establish in the population and the time difference between the establishment of the first and third successful VOC lineage. The red dashed rectangle shows the region of the parameter space corresponding to the emergence of the first three SARS-CoV-2 VOCs with the cross sign ("X") representing the mean value. We can see that a noticeable fraction of simulation runs for both scenarios yield a temporal clustering that matches the emergence of the first three VOCs in late 2020 (i.e., they fall inside the enclosed area). The inset shows that 35.5% and 25.9% of the runs for $K = 3$ and 6 scenarios produce fewer than three successful VOC lineages by the end of the simulation period, respectively. Each run stops once the frequency of the VOC population reaches 75%. **(c).** A typical evolutionary trajectory corresponding to the $K = 6$ scenario highlighted with a bold orange circle in panel **(b).**. The graph shows the background population in gray and the i-mutant populations ($1 < i \leq K$) in different shades of orange from light (fewer mutations) to dark (more mutations). The dashed lines show the dynamics of all the established VOC lineages over time. Red vertical arrows show the establishment time of the first three VOCs. We can see that the single-mutant genotypes (lines in light orange) are produced via the between-host pathway from very early on in the pandemic but are at very low prevalence before the emergence of the VOCs (white dashed lines). See also Appendix Figure C.6.

3.3 Discussion

The global spread of the Omicron variant of SARS-CoV-2 has given a renewed attention to the underlying evolutionary mechanisms that lead to the emergence of VOCs. Practically, we would like to know whether to expect future VOCs to arise, and if so when and whether there will be early warning signs. Answering this question is not only important for understanding the fate of the pandemic but also may have major public health implications for how to best develop strategies for controlling the spread of the disease. In this study, we provided a quantitative framework for investigating the likelihood of different evolutionary pathways that can give rise to VOCs of SARS-CoV-2. We found that VOCs are unlikely to be driven by a single adaptive change at the population level as this would require significantly high levels of overdispersion which is not supported by any existing epidemiological study on SARS-CoV-2 transmission [34]. We also showed that if multiple VOC mutations combine additively for advantage, they can only emerge on the background of a chronic infection, otherwise individual VOC mutations would reach high frequencies from the early stages of the pandemic and, therefore, would have been picked up from genomic surveillance data. If individual VOC mutations were acquired during chronic infections and had a strong advantage relative to the then-dominant genotypes, they may have still been leaked to the population at large before the emergence of VOCs. Therefore, we showed that only additive mutations with relatively small fractional contribution to VOC fitness may yield evolutionary dynamics that resembles the clustered emergence of SARS-CoV-2 VOCs in late 2020. On the other hand, we showed that cryptic circulation of a mutant lineage for sustained periods of time before producing VOCs is possible via a fitness plateau-crossing landscape. While at the between-host level such a landscape may not yield more than two mutations in excess of the background population over a period of 7-12 months under biologically relevant parameter combinations, many more mutations can be accumulated during a chronic infection, for example such as those found in certain immunocompromised individuals, without ever being leaked to the rest of the population. We found that the pattern of the timing of VOC emergence via the fitness plateau-crossing landscape under both the within- and between-host pathways are aligned with the timing of the clustered emergence of Alpha, Beta, and Gamma variants in late 2020.

Studies have shown that the N501Y mutation alone in Alpha have resulted in its enhanced transmissibility in hamsters while other mutations were either neutral or deleterious when expressed individually [70]. This may imply that the evolutionary pathway towards the emergence of Alpha may have been a result of a mixture between Landscape 1 and 2 whereby some of the intermediate mutants had a selective advantage compared to the ancestral state while others were effectively neutral. Furthermore, it is possible that other VOCs such as Delta have taken up a similar evolutionary path where some of the intermediate mutations reached high frequencies in the population before the constellation of mutations appeared in the Delta clade [76].

Phylogenetic studies have also provided evidence for the detection of intermediate Alpha- and Gamma-like genomes which are highly divergent and are ancestral to the Alpha and Gamma clades [46, 51]. This is aligned with the possibility that some of the intermediate mutants which are potentially highly divergent leaked through to the rest of the population before the constellation of K mutation was produced in a chronically infected individual, as predicted by our model.

Finally, it is important to note that in all of the within-host evolutionary pathways (i.e., Landscapes 1, 2, and 3), we found parameter combinations that can re-create the clustered emergence of the first three SARS-CoV-2 VOCs. In particular, we showed that either because of having very few mutations that are selectively beneficial at the population level (i.e., Landscape 1) or the low prevalence of intermediate mutations before the emergence of the VOCs (i.e., Landscapes 2 and 3), we would expect the

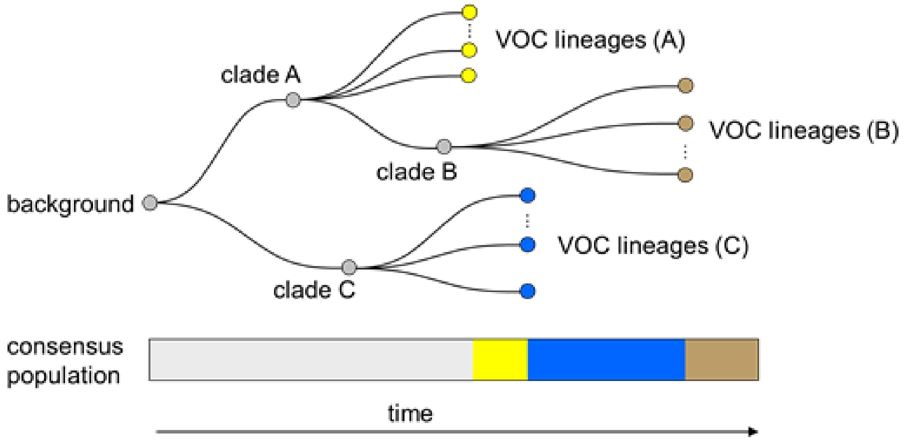


Figure 3.9: Within-host evolution reproduces the star-like genealogy of the VOCs. For a wide range of parameter combinations in Landscape 1, 2, and 3, we showed that the within-host evolutionary dynamics can become virtually uncoupled from the evolution at the between-host level enabling each VOC lineage to arise independently on the background of a different clade which leads to a star-like tree topology.

phylogenetic relationship between the VOC lineages and background populations to manifest itself with a long evolutionary distance branch connected to deeper internal nodes of the tree with each VOC clade independently emerging from a unique genetic background and be subsequently replaced by another VOC clade from an entirely different background (Figure 3.9). This creates a phylogenetic relationship between VOC clades that is similar to what we observe for Alpha, Beta, and Gamma variants [1, 35, 113].

3.3.1 Cryptic transmission of VOCs in humans

Another possibility for why the VOCs were not detected until mid to late 2020 is that they may have been circulating cryptically in areas of the world with poor genomic surveillance before becoming globally dominant. While variants of SARS-CoV-2 with multiple spike mutations have been detected through travel surveillance from passengers travelling from areas with little to no genomic surveillance [21, 31], if they were highly transmissible variants and had a potential to become a VOC, given the interconnectivity of the human interactions, it should not take very long before they become globally dominant. Therefore, as we showed in our analysis of Landscape 1, this scenario of VOC emergence seems to be only possible under significant levels of overdispersion such that it prohibits the selectively beneficial mutations from immediately taking off globally relative to other variants.

3.3.2 Possibility of reverse zoonosis

A somewhat similar idea to the cryptic transmission of variants in human populations is the possibility of a lineage (or multiple lineages) of SARS-CoV-2 jumping from humans to other mammals such as white-tailed deer, mink, hamster, and mouse where they circulate and evolve without being detected for a relatively long period before they jump back to the human population [16, 96, 118, 130]. In particular, some recent studies have reported the detection of multiple spillovers of SARS-CoV-2 from humans and onward transmission in deer population with highly divergent genomes being detected in deer population with potential deer-to-human transmission [66, 102]. However, the genomic composition of these divergent genomes in deer population are different from the VOCs with a much lower ratio of non-synonymous to synonymous changes which suggests they may be following a completely different evolutionary path. Nevertheless, these studies indicate that it is possible for a highly divergent set of genomes to evolve in another species with a potential for deer-to-human transmission without ever being detected. Mink and hamster sequences offer some of the more compelling examples of transmission from humans to a non-human species and back, supported by phylogenetic evidence [96, 130]. None of the currently identified sequences from animals appear as sister taxa to any of the circulating VOCs. While we cannot rule out evolution in an animal reservoir, one might expect the contribution of animals to human transmission chains to be dwarfed by the amount of human-to-human transmission currently happening.

3.3.3 Role of recombination

Recombination can bring together mutations from different backgrounds, potentially expediting the rate of adaptation by creating viable and more pathogenic hybrid new variants of a pathogen. Coronaviruses are also known to recombine with one another during mixed infections [47, 75]. While during the early stages of the pandemic SARS-CoV-2 sequences typically differed by only a handful of mutations from each other thereby making the effects of recombination indistinguishable from those of recurrent mutation [84], as more viral genetic diversity built-up in the population, the generation and transmission of interlineage recombinants of SARS-CoV-2 in humans were reported in multiple studies [49, 56]. Even though there is currently no definitive evidence for recombination being involved in the emergence of VOCs, including the emergence of ancestral Omicron lineage [117], we would expect the role of recombination to be more pronounced as the virus continues accumulate more genetic diversity by sustained circulation around the world. It has also been suggested that the emergence of the BA.3 lineage was a result of an ancestral recombination event between BA.1 and BA.2 [117]. Also, the emergence of the newly identified BA.4/BA.5 lineages was likely through a prior interlineage recombination event [114].

3.3.4 Shifting landscape

We have assumed a static fitness landscape prior to the emergence of the first three VOCs; here we consider the plausibility of that assumption. During the first year of the pandemic, a novel virus was spreading in an immunologically naïve population [79]. As more individuals became infected and developed natural immunity, it is possible that the fitness landscape for the virus shifted as selection for immune escape increased [3]. However, by the time the first three VOCs emerged in late 2020, the majority of the world's population were still susceptible to the disease and may not have even been exposed to it. Therefore, it is unlikely that the build-up of natural immunity alone was the reason behind their increased selective advantage. In contrast, the global dominance of Omicron in late 2021 was largely due to its immune escape properties relative to previous variants of SARS-CoV-2 [103] and, therefore, its emergence was likely the result of a changing viral fitness landscape.

Based on our model assumptions, one may expect that Omicron should have also emerged along with the first three VOCs in late 2020. However, it is possible that the rapid rise in infections globally by early 2021 resulted in a shift in the evolutionary landscape of the virus, creating another long waiting time before the emergence of Omicron. In that respect, our modelling framework may also explain the clustered emergence of BA.1 and BA.2 lineages of Omicron around the same time with largely similar sets of mutations. It is possible that after acquiring the necessary constellation of mutations, they leaked to the rest of the population as BA.1 while the virus was still evolving in a chronically infected host diverged further away from BA.1 by acquiring rapid sequential adaptive substitutions leading to the BA.2 lineage. Therefore, we may expect to see similar shifts in the landscape in 1-2 years from now.

3.3.5 Delta, Omicron, and future VOCs

Emergence of new VOCs with increased transmissibility, immune evasion properties, and virulence poses a great challenge to managing SARS-CoV-2. While we have focused on the first three VOCs, our basic finding that chronic infections greatly increase the virus' ability to explore the fitness landscape suggests such infections are likely to also have been sources of the highly divergent Delta and Omicron variants and may well also be sources of future VOCs. Apart from the N501Y and E484K mutations, other spike mutations such as ΔH69–V70, P681H, and H655Y have also been found in several VOCs including Delta and Omicron as well as some chronically infected individuals [124, 125], making it plausible that these VOC mutations have also emerged from chronic infections which have also been repeatedly favored by natural selection [2].

If chronic infections are indeed the main source of generating VOCs, then finding and treating chronic infections should be a top priority, not just for the benefit of chronically ill patients but also from a public health standpoint. One of the main challenges with assessing the likelihood of VOC emergence during chronic infections would be to quantify the prevalence of immunosuppressed individuals within a population and determine which forms of immunosuppression are associated with chronic infections.

Several studies have now shown evidence of recurrent SARS-CoV-2 mutations in immunocompromised patients [18, 58, 62], with some suggesting the detection of a variant-like lineage which arose from a chronic infection that spilled over into a local population [45, 124]. Another major implication of our work is that we can now quantitatively explain the possibility of such events and find the expected time that it takes for a new VOC to emerge from a within-host evolutionary pathway that involves any number of mutations. We showed that a typical within-host plateau-crossing or additive mutation pathway involving 3-6 mutations requires a within-host fixation rate of $\mu_C \sim 0.1 - 0.3$ per generation which corresponds to a period of 50-300 days since the start of a chronic infection. The timing of such an event aligns with the time frame over which some of the major mutations involved in VOCs have been observed in patients with chronic infections [18, 58, 62]. This also implies that if a VOC emerges from the within-host evolutionary pathway, it is more likely to reflect the genetic diversity of the virus population from several months ago. It can explain why, for instance, Delta was not descended from an earlier VOC, and even more strikingly, the Omicron variants were not descendants of Delta, which was the most prevalent variant at the time of emergence of Omicron. It also suggests that while the next VOC could emerge from the prevalent Omicron background, it could also come from, e.g., a chronic infection with Delta that started prior to the Omicron wave.

Some of the key remaining questions involve how much more of the fitness landscape the virus will be able to explore as more chronic cases accumulate and existing chronic cases last longer. For instance, if it has already crossed a 6-mutant fitness-plateau, how much longer would it take to explore 7-mutant fitness plateaus?

3.4 Materials and methods

3.4.1 Between-host model of VOC emergence

Effective population size

We approximate the between-host evolution of SARS-CoV-2 as a haploid population of size $N(t)$ which is equal to the number of daily infectious individuals with SARS-CoV-2 worldwide. Since the number of confirmed cases is often a significant underestimation of the true number of infections [e.g., see [14, 42]], we use the number of daily confirmed deaths [127] to back-calculate the number of infectious individuals, $N(t)$, from the global median infection fatality rate (IFR) of COVID-19 [68]. We note this approach is still subject to several potential sources of bias including variation in IFR over time (e.g. due to various pharmaceutical interventions) and across different demographics [92]. Using confirmed COVID-19-related deaths may still underestimate the true number of deaths associated with the disease due to under-reporting of deaths particularly in areas of the world where there is limited testing from suspect cases [59]. Nevertheless, by allowing for a wide variation in global IFR (from 0.2% to 1.5%), we can capture most of the uncertainty in the number of infectious individuals worldwide. We also note that for the timespan of interest in our work (i.e., start of the pandemic until the emergence of the first three VOCs), the impact of pharmaceutical interventions such as vaccination on lowering the global IFR is likely to have been negligible given that vaccination campaigns mostly started in 2021. The confirmed global deaths started being reported from 2020-01-23. Assuming a 20-day delay from the onset of symptoms to death [129], we set 2020-01-03 as the first timepoint in the simulation.

Advantage of mutants

The selective coefficient of a mutant individual depends on its number of mutations and the fitness landscape (see Figure 3.2a). For instance, in the case of an additive fitness landscape of size $K = 3$, the fitness advantage of the single-, double-, and triple-mutants are $s/3$, $2s/3$, and s relative to the wild-type population, respectively. During one generation, the frequency f_i of individuals with genotype i and selective advantage s_i relative to the wild-type increases by a factor $(1+s_i)$, along with further adjustments to their frequency due to mutations from/to other genotypes. Upon normalization ($\sum_i f_i = 1$), these frequencies are used for the Dirichlet-multinomial sampling step. After the sampling step, the numbers of cases are converted to frequencies for sampling in the next generation.

Epidemic spread

Due to a high degree of individual-level variation in the transmission of SARS-CoV-2 (i.e., overdispersion) [34, 71], we use a Dirichlet-multinomial (instead of a multinomial) distribution to assign offspring in generation $t + 1$ to parents from generation t. The Dirichlet-multinomial is parametrized by $N(t + 1)$ (the number of offspring to draw for the next generation) and $A\overrightarrow{f}$, the weights of the different genotypes, where $\overrightarrow{f}$ is the normalized vector giving their frequencies in the current generation. The scalar A controls the amount of dispersion, with smaller A corresponding to increased demographic noise. To match it to observations, we note that under the Dirichlet-multinomial model, the number of secondary cases produced by an infection approximately follows a negative binomial distribution with mean

$R_t = N(t+1)/N(t)$. In terms of the Dirichlet multinomial parameters, the variance of this negative binomial is $\sigma^2 = \frac{N(t+1)}{N(t)}\left(1 - \frac{1}{N(t)}\right)\frac{N(t+1)+A}{1+A} \approx R_t\frac{N(t+1)+A}{1+A}$. This should match the variance in the number of secondary cases written in terms of the dispersion parameter K, $\sigma^2 = R_t(1 + R_t/k)$. Equating these two expressions gives $A = kN(t)(1 - \frac{1}{N(t+1)}) - 1 \approx kN(t)$.

Mutation rate

Assuming a constant generation time 5.2 days for all variants of SARS-CoV-2 over time [36], we use the phylogenetically estimated substitution rate per site per year [43] to calculate the mutation rate per site per generation time, μ. We also note that generation time may vary over time depending on the behavioral changes in the population and emergence of variants, which is why we allow for some variation in the mutation rate parameter in our model $(0.87 - 2.0) \times 10^{-5}$ based on phylogenetic estimates. We assume each site has two states: wild-type and mutant. Therefore, for a group of K sites, there are $2K$ genotypes. The reason for choosing this binarization of the mutation states is given the relatively short evolutionary timescales over which we are examining the evolution of SARS-CoV-2, we do not expect to see more than one nucleotide change at any given site. We assume that at each site there is only one possible mutation that contributes to the VOC phenotype, so alternative mutations can be neglected.

Inferring the selective advantage of VOCs

Finally, the selective advantage s of the VOCs is determined by fitting an exponential function, $f(t)$, of the form, $f(t) = a\mathrm{e}^{\mathrm{st}}$, to the proportion of Alpha, Beta, and Gamma variants sampled in the country where they were first detected (i.e., UK, South Africa, and Brazil) using the NonlinearModelFit function in Mathematica 11.0 [126]. We find that the selective advantage s for Alpha, Beta, and Gamma are 0.37

(95% confidence interval: 0.33 - 0.41), 0.74 (95% confidence interval: 0.65 - 0.83), and 0.84 (95% confidence interval: 0.58 - 1.08), respectively (see Appendix Figure C.7). The confidence interval is obtained by multiplying the standard error by the value of Student's t for the given confidence level and degrees of freedom. Given the uncertainty in our estimates due to noise in the observations, potential sampling bias, and spatio-temporal heterogeneities, we make the assumption that the value of s is roughly the same for the different VOCs and use the same estimate for all three trajectories (Table 3.1).

3.4.2 Within-host model of VOC emergence

Each VOC mutation is fixed within the host at rate μ_C such that the fixation time is an exponentially distributed number with mean $1/\mu_C$. Each mutation may then spread to the rest of the population with a probability that is proportional to its fitness as determined by the Dirichlet-multinomial sampling. At any time during the pandemic, a chronic infection can be seeded by other infectious individuals within the population, $N(t)$, with a probability P_f. Therefore, at every generation, the number of chronic infections is given by a binomial distribution with success probability P_f. Once a chronic infection emerges, it remains in the population for the remainder of the simulation period.

3.4.3 Simulation setup

For both within-host and between-host models of VOC emergence, we run each evolutionary scenario for a given combination of model parameters 1,000 times. We then measure total number of established VOC lineages, M, the time that it takes for the establishment of the first VOC, T_0, and the time between the establishment of the i^{th} and $(i+1)^{th}$ VOC, $T_{i:(i+1)}$, for the first six established VOC lineages in each scenario. An established VOC lineage is defined as a lucky lineage with selective advantage s

that survives drift upon reaching a size $1/s$. Similarly, the establishment time of a VOC lineage is defined as the time that it takes for that lineage to reach size $1/s$. Each run stops once the frequency of the VOC population reaches 75%.

Conclusion

In this dissertation, we have investigated various aspects of evolution: optimal population dynamics, mode of selection, and the emergence of SARS-CoV-2 variants of concern (VOCs). The findings from the three independent projects shed light on different mechanisms and factors influencing adaptation and genetic changes.

Collectively, these three projects contribute to our understanding of the complexities of evolution and provide insights into the mechanisms underlying adaptation and genetic changes. By considering population dynamics, selection modes, and real-life scenarios such as viral evolution, this research enhances our knowledge of the factors shaping evolutionary processes. Further investigations building upon these findings will continue to advance our understanding of evolution and its implications for various biological systems and practical challenges.

Synchronization at the population and individual level

In Chapter 1, we focused on investigating the synchronization of population and individual levels with respect to population structure. Our study revealed how population structure can facilitate the accumulation of genetic diversity, and how this can combine with synchronized dispersal and sexual reproduction to produce very fit recombinant individuals.

The novel mechanism we discovered shows how population structure can expedite adaptation on smooth fitness landscapes without the presence of epistasis by minimizing clonal interference among alleles. This finding has significant implications for evolutionary experiments, as structured populations are often preferred when studying scenarios involving epistasis, as they allow for exploration across the genome space [22, 65, 87, 116].

The bursty population dynamics we observed are particularly effective in generating novel genotypes during periods of environmental stress, facilitating significant moves within the genome space. This characteristic is especially advantageous when crossing fitness valleys becomes necessary [25]. Moreover, the proposed synchronization approach is practical and feasible for evolutionary experiments. Microbial experiments have already investigated the rate of adaptation in structured populations [65, 101]. By incorporating a periodic stressful environment into the bursty mixing setup, it can be easily implemented using existing experimental protocols [65]. Simply introducing a stressful stimuli during the mixing procedure to induce facultative sexual reproduction allows for the creation of the desired synchronization. Therefore, our model holds promise for further experimental investigations.

Exploring the role of bursty population dynamics in the context of complex fitness landscapes, including landscapes with epistasis and ruggedness, would be a fascinating area of investigation. This would help elucidate the specific conditions under which bursty populations are most advantageous and shed light on the interplay between population structure, synchronization, and the exploration of the genome space.

Saturating fitness function

In Chapter 2, we explored the introduction of the logistic fitness function to limit the reproductive advantage of the fittest individuals. This approach promoted the coexistence of competing lineages, generating abundant genetic diversity. Comparing different forms for the fitness function revealed that the key to enhancing adaptation was promoting the reproduction of moderately fit individuals.

However, it is important to note that there is a lack of direct results regarding the contribution to adaptation of individuals from different parts of the fitness distribution. Previous studies have indicated that moderate-strength truncation selection ($\sim 15\% - 50\%$) can be optimal for long-term evolution [52, 53, 105]. This suggests that the contribution of the best individuals directly impacts short-term evolution, while moderately fit individuals play a crucial role in long-term evolution. The impact of maladapted individuals remains unclear. On one hand, they can hinder adaptation by reducing the efficiency of exploiting beneficial genotypes. On the other hand, they may carry novel mutations that occur independently of their genetic background. In the bursty population framework, the genetic variation present in maladapted individuals can be used during periods of very frequent sexual reproduction, freeing beneficial mutations from poor genetic backgrounds. Future studies could focus on investigating the evolutionary contribution of these less-fit individuals under consistent environmental conditions.

Moreover, it is worth considering the potential integration of the optimal evolution strategies explored in this chapter and in Chapter 1. In Chapter 1, we introduced a population dynamics that effectively uses the genetic diversity preserved by population structure. In Chapter 2, we introduced a mode of selection that greatly increases genetic diversity. By combining these two approaches, we propose a more effective evolution strategy. The fitness function can allow genetic diversity to build up during typical generations, and then these genetic resources can be converted to phenotypic diversity during bursts of frequent sexual reproduction. These bursts could also include periods during which the fitness function is switched back to a more standard exponential form to allow for increased exploitation. This integrated strategy holds promise for adaptation in scenarios characterized by strong clonal interference.

The emergence of the first three SARS-CoV-2 variants of concern

In Chapter 3, we introduced a quantitative framework to study the evolutionary pathway leading to the emergence of Variants of Concern (VOCs) of SARS-CoV-2, focusing on the contrast between evolution occurring while being transmitted from host to host and within-host evolution during chronic infection. Our findings highlight that VOCs are primarily driven by combinations of mutations originating from chronic infections. Consequently, treating chronic infections becomes crucial in reducing the rate of future VOC emergence. While our simulations are consistent with multiple clinical studies (see Chapter 3, Section 3.3), they rely on inferred clinical parameters. The actual mechanism of within-host evolution, in particular, remains poorly understood. We treat this process as a "black box" and assume a constant within-host substitution rate, independent of the genetic background and health of the host. Additionally, there is a lack of studies on two key parameters, P_f and μ_c, which we indirectly infer. It is important for subsequent studies to improve the understanding of the actual mechanisms and corresponding parameters to refine the accuracy of the predictions. It remains completely unclear, for instance, how the within-host fitness landscape relates to the between-host one, i.e., why evolution within a single host should sometimes select for increased ability to transmit between hosts.

Regarding between-host evolution, there has been unprecedented attention given to this specific process for COVID-19. As a result, we have multiple sources to infer the parameters (Table 3.1). However, there are simplifications in this mechanism. We treat susceptible individuals as a well-mixed population, neglecting the spatial structure of SARS-CoV-2 transmission. This simplification may impact the spread of intermediate mutants that do not benefit from complete fitness increases. We argue that this effect might be minor in the early stages when the mutant frequency is still low. Lineages can still experience exponential growth in frequency based on fitness advantages before encountering limitations imposed by locality. But the observed incidence curves of the VOCs in fact deviate substantially from simple exponentials. The reasons for this remain unclear, as does the reason for the prolonged period of stasis in overall case numbers, and for the large fluctuations observed in local allele frequencies.

Appendix A

Population structure can reduce clonal interference under synchronized recombination and dispersal

A.1 The rate of adaptation w/ and w/o sex

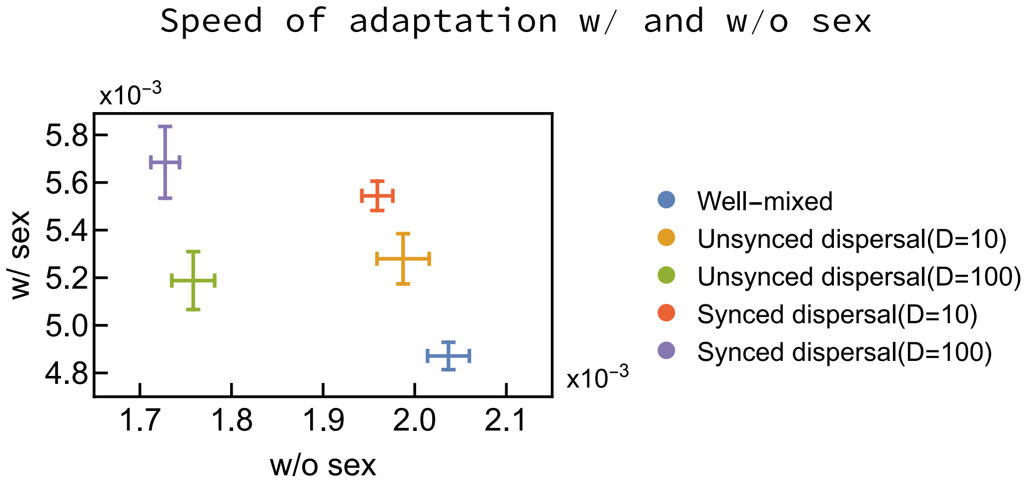


Figure A.1: Structured populations benefit more from sexual reproduction than well-mixed populations. The vertical and horizontal axes are the adaptation speeds with and without sexual reproduction, respectively. Sexual reproduction is assumed to be unsynchronized. Without sexual reproduction, the well-mixed population adapts the fastest, while with sexual reproduction, it is the slowest.

A.2 The adaptation speeds of different population structures

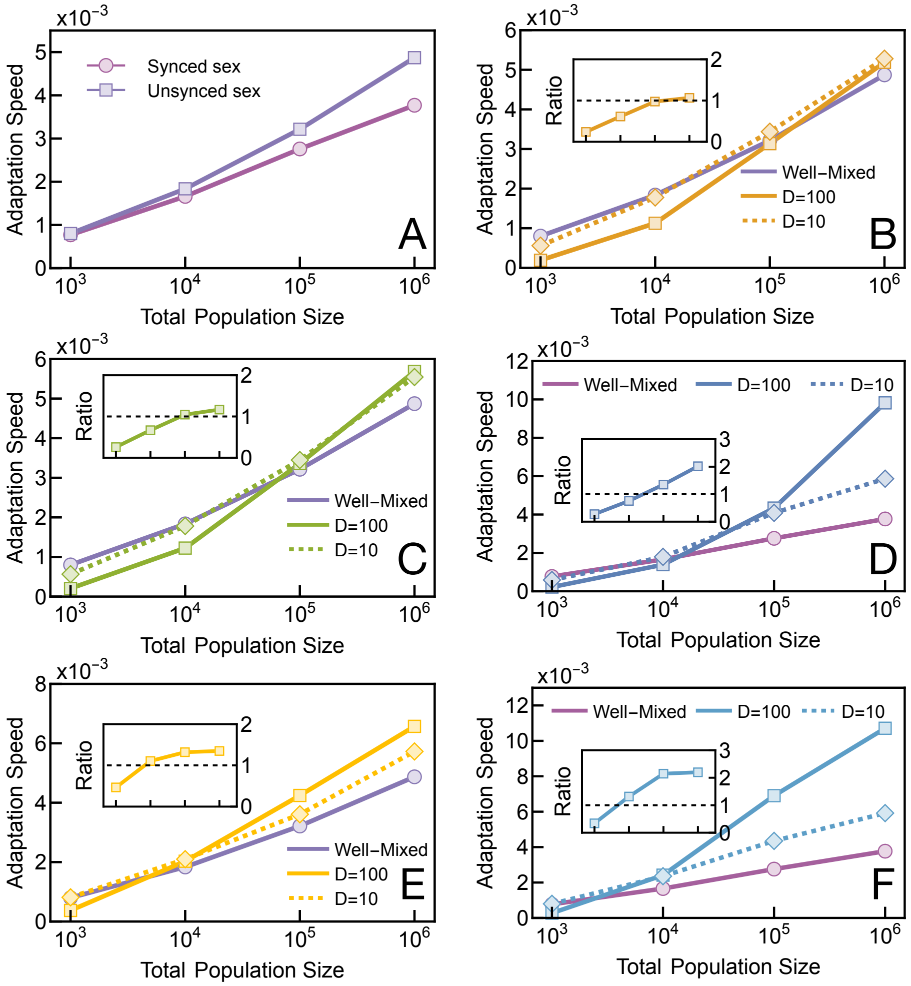


Figure A.2: The adaptation speeds of different population structures. Ratio is compared to a corresponding well-mixed population (w/ or w/o synced sex). **A.** Well-mixed population. **B.** Unsynced structured population. **C.** Structured population with unsynced sex and population synced dispersal. **D.** Population synchronization (synced sex and dispersal). **E.** Individual synchronization. Derived from Fig. A.2B. **F.** Population and individual synchronization structured. Derived from Fig. A.2D. **(B-F)**. The violet lines are the corresponding well-mixed population (w/ or w/o synced sex) from Fig. A.2A.

A.3 Genetic difference of the best individuals with the population synchronization and unsynced sex

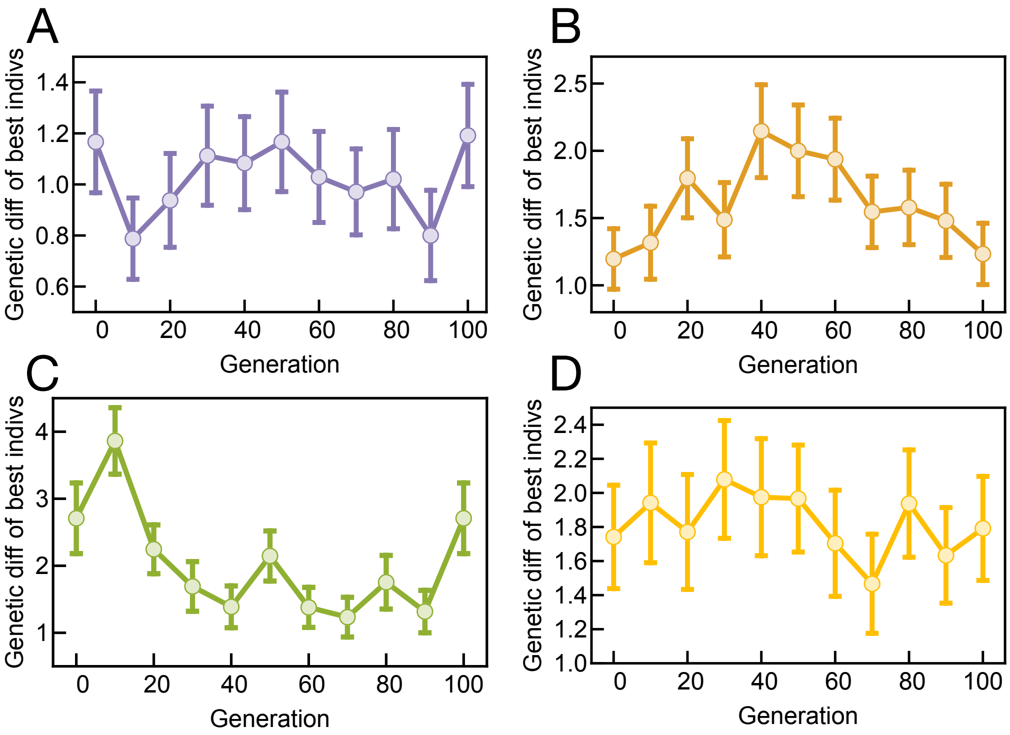


Figure A.3: Genetic differences of the best individuals with the population synchronization and unsynced sex. **(A)**. Well-mixed population. **(B)**. Unsynced structured population **(C)**. Structured population with unsynced sex and population synced dispersal. **(D)**. Individual synchronization. Derived from Fig. A.3B. **(B-C)** are specifications of Fig. 1.2B.

A.4 Trend of global genetic diversity

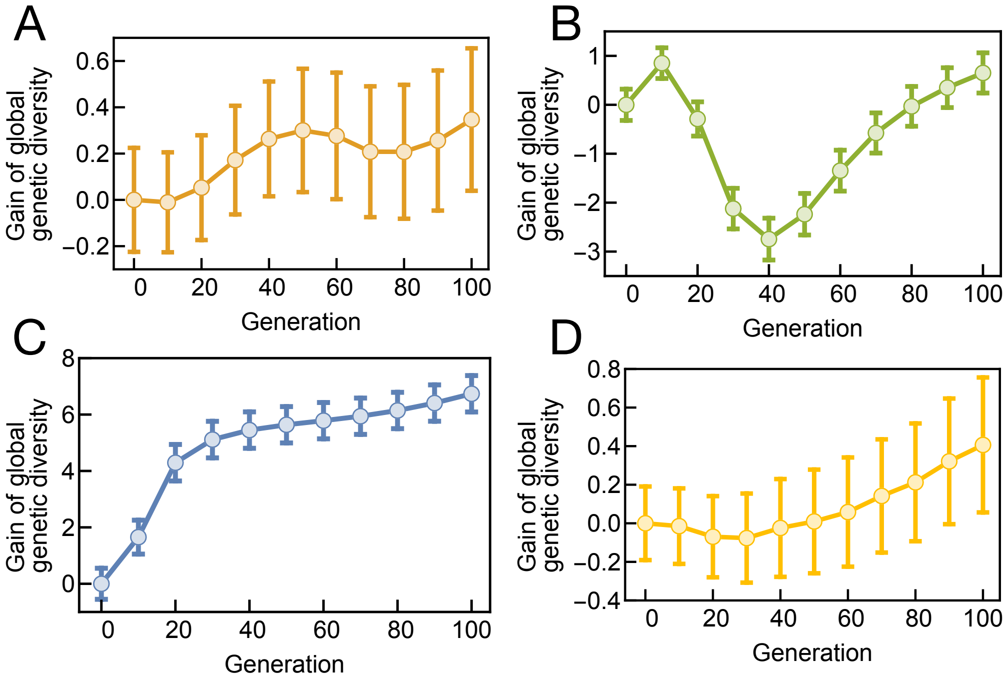


Figure A.4: Trend of global genetic diversity of the population synchronizations: unsynced (A), population synced dispersal only (B), population synced sex and dispersal (C) and individual synced only (D). Data are set to start with 0 at each t_{gap} cycle. The trend is 'S' curved with only synced dispersal (**B**), while it is constantly increasing trend with synced sex and dispersal (**C**).

A.5 Maximum fitnesses and their frequency

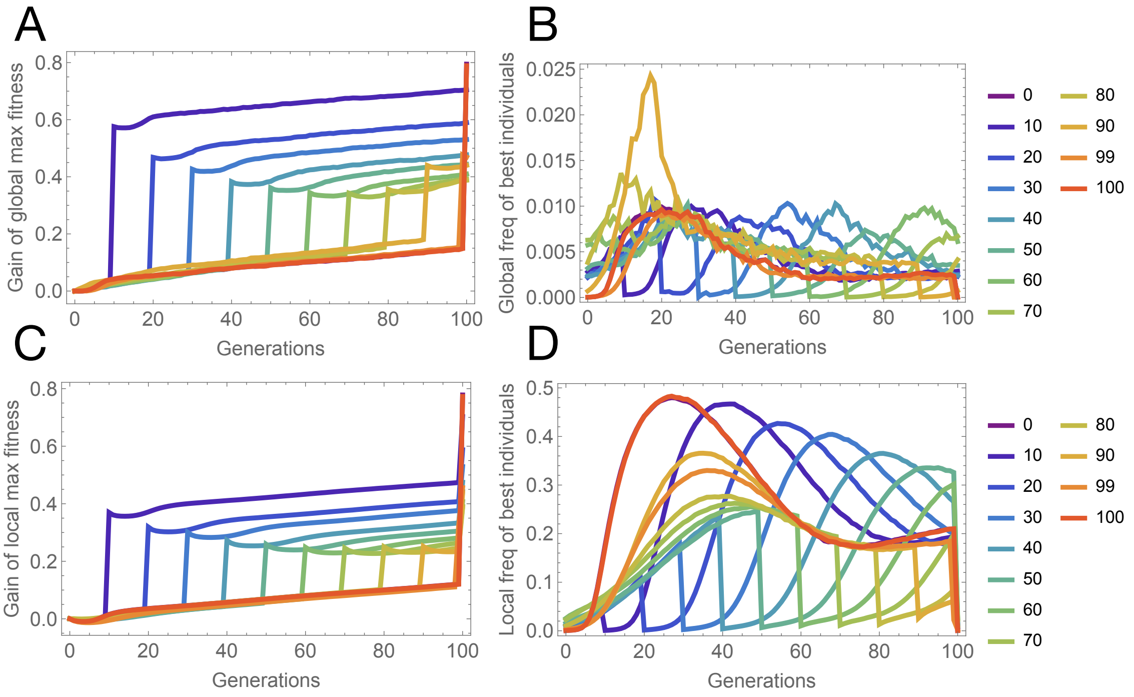


Figure A.5: Maximum fitnesses and their frequency in the offset simulation, as in Fig. 1.3. Maximum fitnesses and their frequency globally (A-B), and locally(C-D). When sex is introduced, the frequency of best individuals are usually very low $\sim 0.1\%$. So, the novel genotypes are vulnerable.

Appendix B

Promoting moderately fit individuals can increase adaptation speed under strong clonal interference

B.1 Normalize the fitness function

Since the weight of each individual being selected as a parent is proportional to its fitness $W(z)$, the mean of all fitnesses should be normalized to 1 when the population size is constant $E[W(z)] = 1$. With an exponential fitness function, this is avoided by shifting the distribution of z [120]. However, the form of a logistic fitness function $W(z) = \frac{c}{1+\mathrm{e}^{-\mathrm{ks}(\mathrm{z}-\mathrm{z}_0)}} = \frac{2(1-d)}{1-2d+e^{-2s(1-d)z}}$ has not yet be normalized.

We consider the approximation logistic function [26]. A shifted logistic function can be approximated by the error function $\mathrm{erfc}(x) = 1 - \mathrm{erf}(x) = 1 - \frac{2}{\sqrt{\pi}} \int_0^x \mathrm{e}^{-\mathrm{t}^2} \mathrm{dt}$:

$$L(x|k, x_0) = \frac{1}{1 + \exp\left(-ks(x - x_0)\right)} \approx \frac{1}{2} \mathrm{erfc}\left(-\frac{ks\sqrt{\pi}}{4}(x - x_0)\right) \tag{B.1}$$

Also, the product of an error function and an exponential function can be integrated as [91]:

$$\int_{-\infty}^{+\infty} \operatorname{erf}(x)\exp(-ax+b)^2 dx = -\frac{\sqrt{\pi}}{a}\operatorname{erf}(\frac{b}{\sqrt{a^2+1}}), \quad \Re(a) > 0$$

This can be further extended to the expectation under the normal distribution:

$$\begin{aligned}
E[L(z|k,z_0)] &= \int_{-\infty}^{+\infty} L(z|k,z_0)\mathcal{N}(y|z_N,\sigma)dz \\
&= \int_{-\infty}^{+\infty} L(z|k,0)\mathcal{N}(z|z_N - z_0,\sigma)dz \\
&\approx \int_{-\infty}^{+\infty} \frac{1}{2}\operatorname{erfc}(-\frac{ks\sqrt{\pi}}{4}z)\mathcal{N}(z|z_N - z_0,\sigma)dz \\
&= \frac{1}{2}\operatorname{erfc}\left(-\frac{ks\sqrt{\pi}}{4\gamma}(z_N - z_0)\right); \quad \gamma = \sqrt{1+\frac{k^2s^2\pi\sigma^2}{8}} \\
&\approx L\left(z_N - z_0|\frac{k}{\gamma},0\right) = \frac{1}{1+\exp\left(-\frac{ks}{\gamma}(z_N - z_0)\right)}
\end{aligned} \tag{B.2}$$

Thus, the fitness function $W(z) = c \times L(z|k,z_0)$ should be normalized to $W_{\text{effective}}(z) = c' * L(z|k,x_0)$ when $z \sim \mathcal{N}(0,\sigma^2)$:

$$\begin{aligned}
W_{\text{effective}}(z) &= c' * L(z|k,z_0) = L(z|k,x_0)/L(-z_0|k/\gamma,0) \\
&= \frac{1+\exp(ksz_0/\gamma)}{1+\exp\left(-ks(z-z_0)\right)}; \quad \gamma = \sqrt{1+\frac{k^2s^2\pi\sigma^2}{8}} \\
&= \frac{\left(\frac{1}{1-2d}\right)^{\frac{\sqrt{2}}{\sqrt{\pi(d-1)^2s^2v+2}}}+1}{\frac{e^{2(d-1)sz}}{1-2d}+1}
\end{aligned} \tag{B.3}$$

The variance of the breeding value z is not a constant. The fitnesses will automatically be normalized by the sampling methods. Therefore, the actual fitness function is $W_{\text{effective}}(z)$.

B.2 Adaptation speed with unlinked loci

Previous study has studied the the adaptation speed with exponential fitness function [120]. Here, we extend the formula of adaptation speed to a general fitness function. According to the infinitesimal model, the breeding value z is the result of a vast number of alleles with additive effects. The breeding values x of offspring of two individuals with breeding values y, z would follow a normal distribution $x \sim \mathcal{N}((y+z)/2, \sqrt{v/2}) = \phi(x|y,z)$. As discussed in the main text, v is the variance in z at equilibrium and the rate of increase in z based on Fisher's "Fundamental Theorem". Therefore, we have the distribution of breeding values in the current generation following a normal distribution $z \sim \mathcal{N}(0, \sqrt{v}) = \psi(z)$.

In order to obtain the fixation probability of an allele, we consider it occurs in a genetic background z at $t = 0$, which results in the fitness as $(1+s)W(z)$. That means this individual would produce a Poisson-distributed number of offspring with the expectation of $W(z)$. With random mating, each offspring would have the other parent draw from the distribution $\psi(z)W(z)$.

The fixation probability at $t = 0$ is $P_0(z)$, and the chance of losing this allele is $Q_0(z) = 1 - P_0$. In general, $P_t(z) = P_0(z - vt)$. The probability of loss is determined by the chances of loss for all offspring (with), considering all possible mates:

$$
\begin{aligned}
Q_0(z) &= \sum_{j=0}^{\infty} \frac{\mathrm{e}^{-\lambda}\lambda^{\mathrm{j}}}{j!} (\int_{-\infty}^{\infty} \psi(y)W(y) \int_{-\infty}^{\infty} \phi(x|y,z)Q_1(x)dxdy)^j \\
&= \exp\left(-\lambda \int_{-\infty}^{\infty} \psi(y)W(y) \int_{-\infty}^{\infty} \phi(x|y,z)P_i(x)dxdy\right)
\end{aligned}
\tag{B.4}
$$

Where $\lambda = (1+s)W(z)$ is the fitness with the allele.

Therefore, the probability of fixation is

$$\begin{aligned} P_0(z) &= 1 - Q_0(z) \\ &= 1 - \exp\left(-\lambda \int_{-\infty}^{\infty} \psi(y)W(y) \int_{-\infty}^{\infty} \phi(x|y,z)P_1(x)dxdy\right) \\ &= 1 - \exp(-\lambda P_I(z)) \end{aligned} \tag{B.5}$$

Assuming $s \ll 1$ and the clonal interference is strong $P \ll 2s$, the probability of fixation can be approximated by $P \sim \mathcal{O}(s)$:

$$\begin{aligned} P_0(z) &= 1 - Q_0(z) \\ &= 1 - \exp\left(-\lambda \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \psi(y)W(y)\phi(x|y,z)P_1(x)dxdy\right) \\ &= 1 - \exp\left(-(1+s)\tilde{P}_0(z)\right) \\ &\approx (1+s)\tilde{P}_0 - \frac{1}{2}\tilde{P}_0(z)^2 + \mathcal{O}(s^3) \end{aligned} \tag{B.6}$$

where $\tilde{P}_0$ is given by

$$\begin{aligned} \tilde{P}_0 &= W(z) \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \psi(y)W(y)\phi(x|y,z)P_1(x)dxdy \\ &= W(z) \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \psi(y)W(y)\phi(x|y,z)P_0(x-v)dxdy \end{aligned} \tag{B.7}$$

Since $P_0(z) = \tilde{P}_0 + \mathcal{O}(s^2)$, $P_0(z)$ is obtained to the first order in s for a specific fitness function $W(z)$. However, the coefficient of $P_0(z)$ can not be determined from Eq. B.7. Instead, we can infer it from Eq. B.6 by taking the expectations:

$$2s\bar{P} = \overline{P(z)^2} \tag{B.8}$$

By combining both Eq. B.7 and Eq. B.8, we have the fixation probability. Substituting the expression for v, it becomes $v = NU\bar{P}\log(1+s)$.

For our logistic fitness function, we use polynomial approximation $W(z) = 1 +$

$z + dz^2$ to calculate its fixation probability. Its effective fitness function can be

$$W_{\text{effective}}(z) = 1 + z + dz^2 - d^3v^2 \tag{B.9}$$

Thus, the fixation probability $P(z)$ can be approximated

$$\begin{aligned} P(z) &= c_1(1 + c_2 z + c_3 z^2) + \mathcal{O}(z^3) \\ &\approx c_1\left(1 + 2z + \frac{4}{3}(1+d)z^2\right) \end{aligned} \tag{B.10}$$

where

$$c_1 = \frac{2s(4dv + 4v + 3)}{16d^2v^2 + 32dv^2 + 8dv + 16v^2 + 20v + 3}$$

.

Since $v = NU\bar{P}\log(1+s) \approx v_0\bar{P}/2s$, the overall adaptation speed is given by

$$(3 + 4(1+d)v)^2 v_0 - (9 + 12(5+2d)v + 48(1+d)^2v^2)v = 0 \tag{B.11}$$

The solution to this function can be approximated by:

$$\begin{aligned} v \approx & v_0 - 4v_0^2 + \frac{32}{9}\left(-11 - d + d^2\right)v_0^3 \\ & + \frac{64}{27}\left(-200 - 36d + 33d^2 + 4d^3\right)v_0^4 \end{aligned} \tag{B.12}$$

In comparison, the adaptation speed with the exponential fitness function has the Taylor series as $v \approx v_0 - 4v_0^2 + 24v_0^3 - \frac{512\mathrm{v0}^4}{3}$.

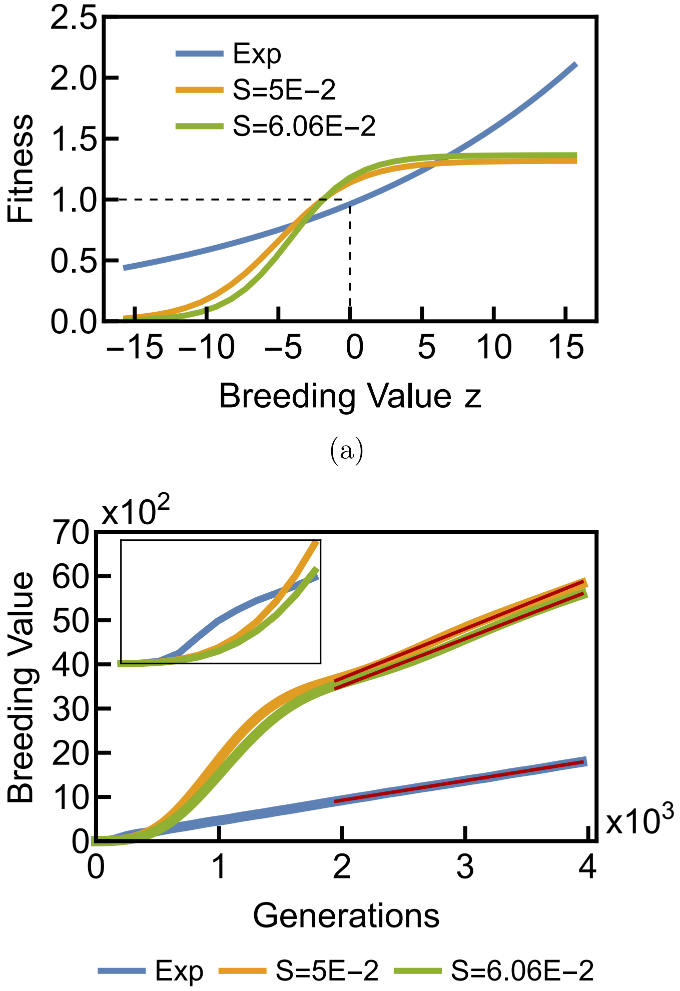
B.3 Rescale the selective coefficient

One of the conditions for a fitness function is that a novel mutation should have a fixed evolutionary benefit in a uniform genetic background, which can be expressed as $W'(0) = 1$ or $W(1) = \mathrm{e}^{\mathrm{s}}$.

However, this condition is no longer valid when the effective fitness function $W_{\text{effective}}(z)$ is scaled from the general fitness function $W(z)$, as the mean of the effective fitness function should be 1. This rescale is unavoidable, since the population size is fixed and it occurs automatically in the sampling method. Therefore, we manually scale the selective coefficient to validate the condition

$$W_{\text{effective}}(z, s_L) = \mathrm{e}^{\mathrm{sE}}$$

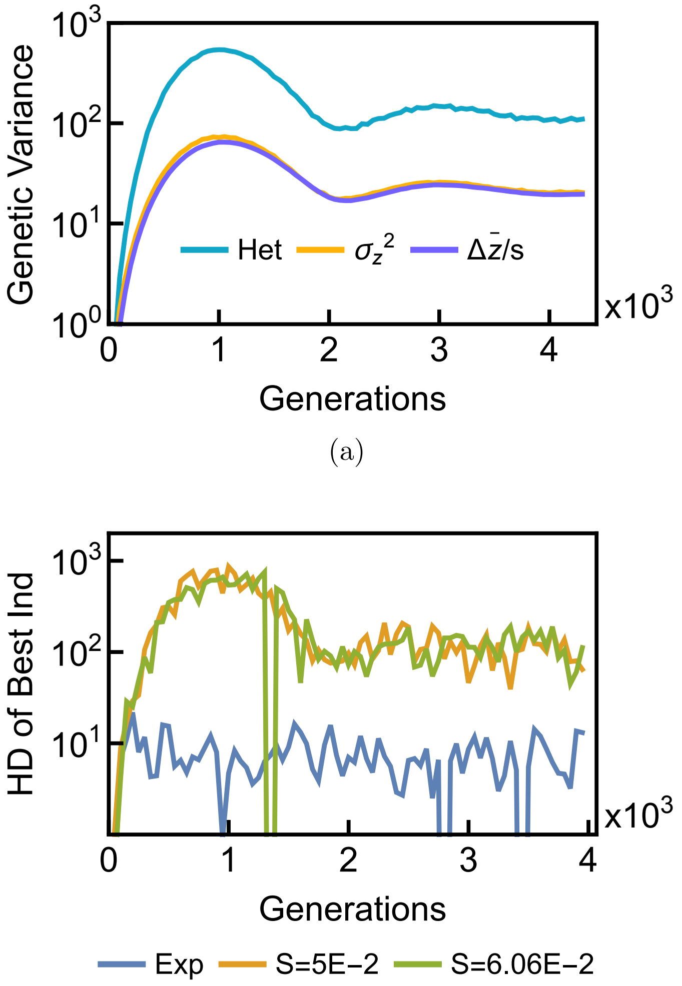
Specifically, when $s_E = 5\mathrm{e} - 2$, the scaled selective coefficient is $s_L = 6.06\mathrm{e} - 2$. The differences between these two selective coefficients are minor, which can be further reduced when clonal interference is strong. As shown in Fig.B.1, the fitness functions with either selective coefficient are similar. The overall simulation results, including the adaptation speeds, genetic diversities, and the Hamming Distances of the fittest individuals, are highly comparable (Fig.B.1 and Fig.B.2). Therefore, the faster rate of adaptation in logistic fitness function is not a result of a higher selective coefficient, but rather a result of the form of the fitness function.



(a)

(b)

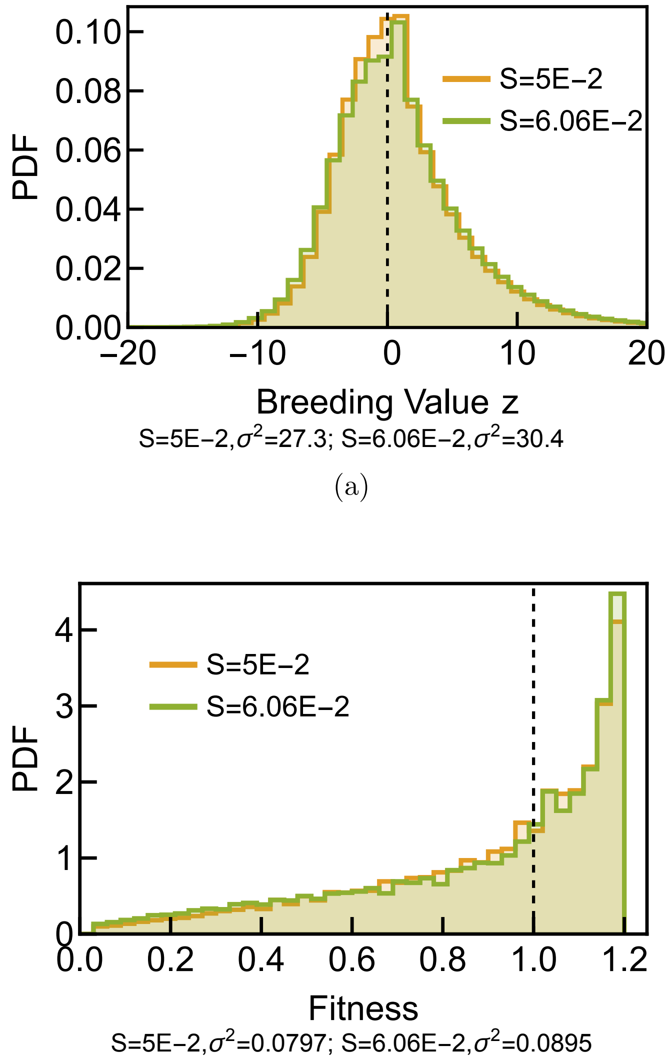
Figure B.1: The fitness functions with different selective coefficients are similar. Therefore, they yield similar adaptation trajectories in the simulations. The ratios of the adaptation speed compared with the exponential fitness function are 2.68 for $s = 5\text{E} - 2$ and 2.79 for $s = 6.06\text{E} - 2$.



(a)

(b)

Figure B.2: The observed genetic variance and relatedness among the fitness functions with different selective coefficients are highly comparable.



(a)

(b)

Figure B.3: The distributions of breeding values and fitnesses remain similar whether using the unscaled and scaled selective coefficients.

B.4 Cauchy and Gaussian scaled fitness functions

We propose a general form to rescale the exponential fitness function

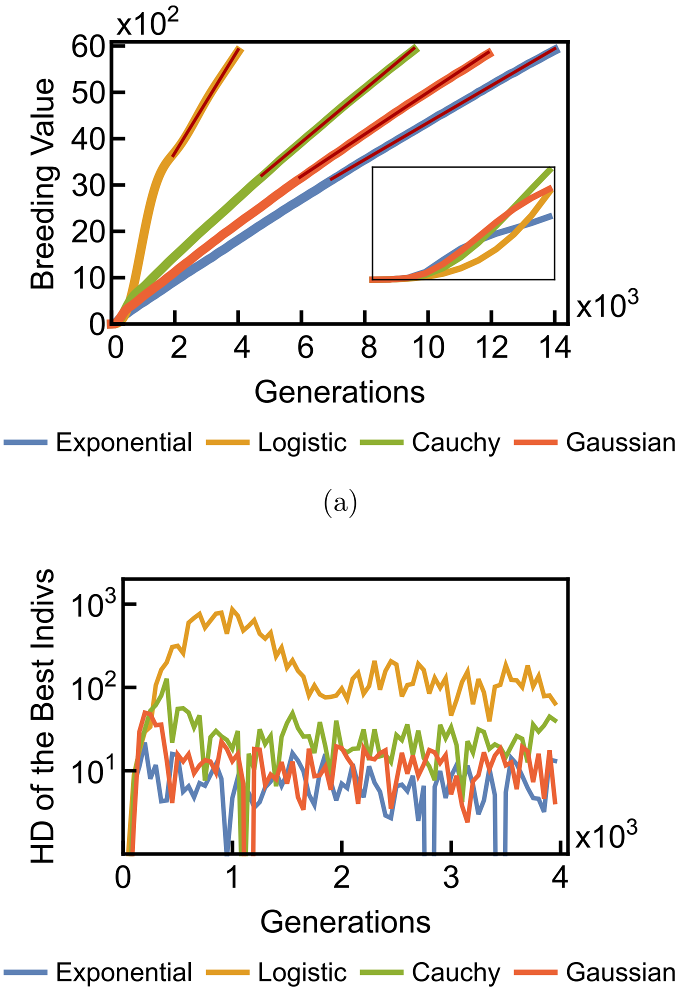
$$W(z) = \mathrm{e}^{\mathrm{sz}}/\mathrm{f(z)} \tag{B.13}$$

and the scale function $f(z)$ should satisfy the following conditions:

$$\begin{cases} f(0) \approx & 1 \\ f(\infty) \sim & \mathrm{e^s z} \end{cases} \tag{B.14}$$

In general, the scale fitness function should be saturated to a certain value when z is large in order to limit the beneficial individuals.

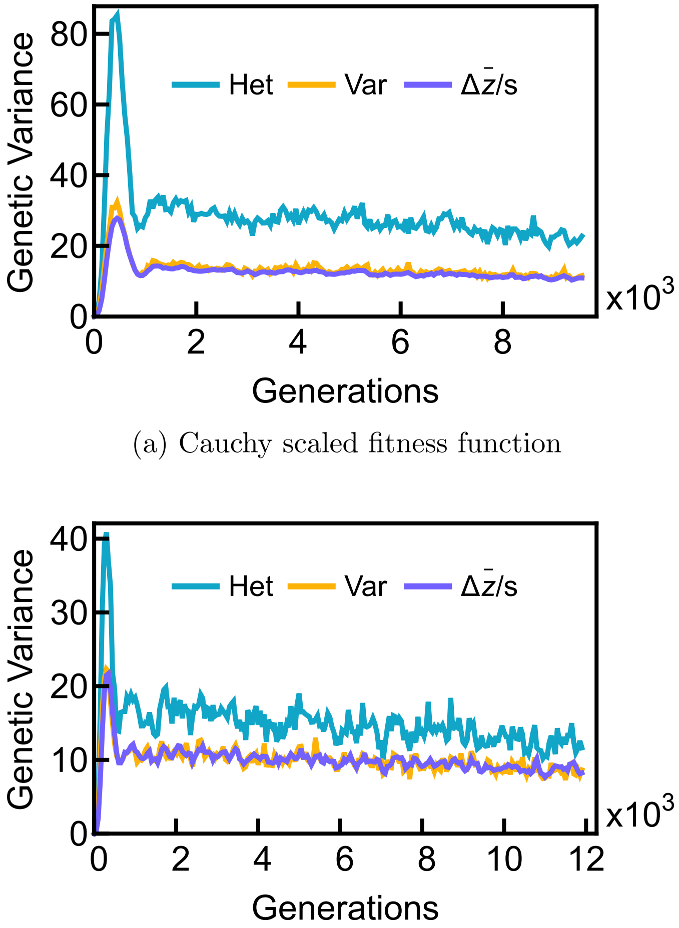
We consider Cauchy and Gaussian as two scale functions. Specifically, the Cauchy scaled function is $f(z) = 1 + 0.5\frac{s^2z^2}{1+s^2z^2}\mathrm{e}^{\mathrm{sz}}$, and the Gaussian scaled function is $f(z) = 1+0.35(1-\exp(\frac{-s^2z^2}{2}))\mathrm{e}^{\mathrm{sz}}$. In our simulations, these two fitness function creates a mild increase in the rate of adaptation, compared with the exponential fitness function. As discussed in the main text, the effectiveness of a fitness function can be indicated by the intensity of the genetic diversity and the Hamming distance of the fittest individuals, when sexual reproduction is present. This matches the simulation results in Fig.B.4 and Fig.B.5.



(a)

(b)

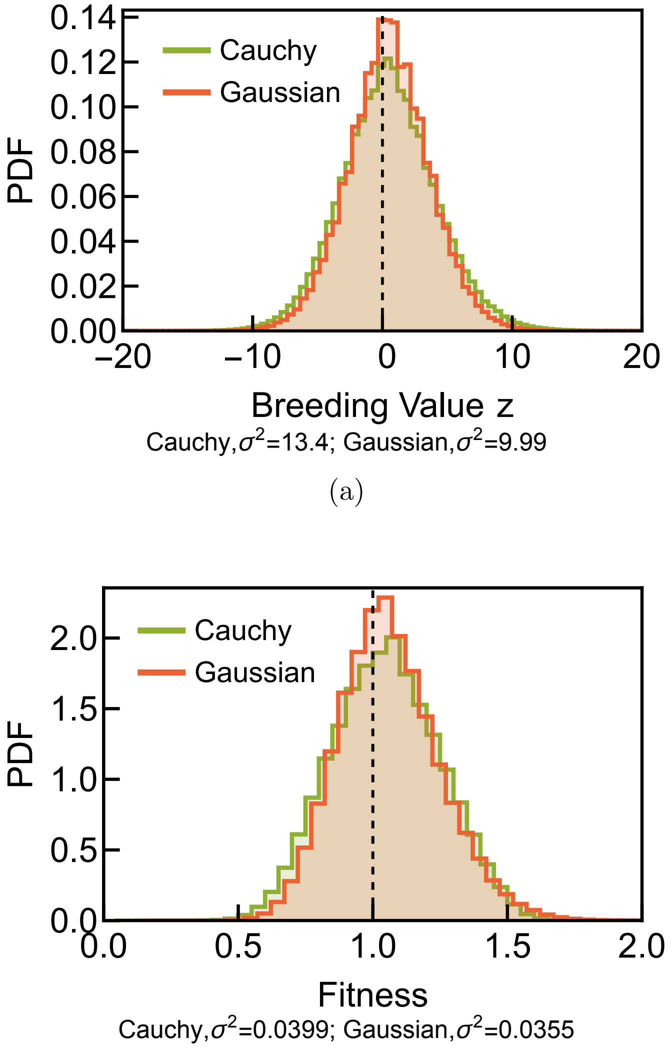
Figure B.4: The adaptation trajectories and relatedness for Cauchy and Gaussian scaled fitness functions. **(a.)** Compared to the exponential fitness function, the ratios of adaptation speeds for logistic, Cauchy, and Gaussian scaled fitness functions are 2.79 ± 0.02 1.43 ± 0.01 and 1.13 ± 0.01 respectively. **(b.)** The Hamming Distance of the best individuals between two consecutive generations is positively correlated with the adaptation speeds. This could be an indicator of the effectiveness of sexual reproduction.



(a) Cauchy scaled fitness function

(b) Gaussian scaled fitness function

Figure B.5: The genetic diversity of populations subject to Cauchy and Gaussian scaled fitness functions is positively correlated with adaptation speed. This suggests that the faster adaptation speed observed with the saturating fitness function is a consequence of more effective sexual reproduction, as evidenced by the increased genetic diversity.



(a)

(b)

Figure B.6: The breeding values are normally distributed for both Cauchy and Gaussian scaled fitness functions. The maximum breeding value for Cauchy and Gaussian scaled fitness functions are 39.15 and 30.16 respectively.

Appendix C

Investigating the evolutionary origins of the first three SARS-CoV-2 variants of concern

C.1 Distribution of waiting times with the between-host pathway assuming a fitness landscape with a single adaptive mutation.

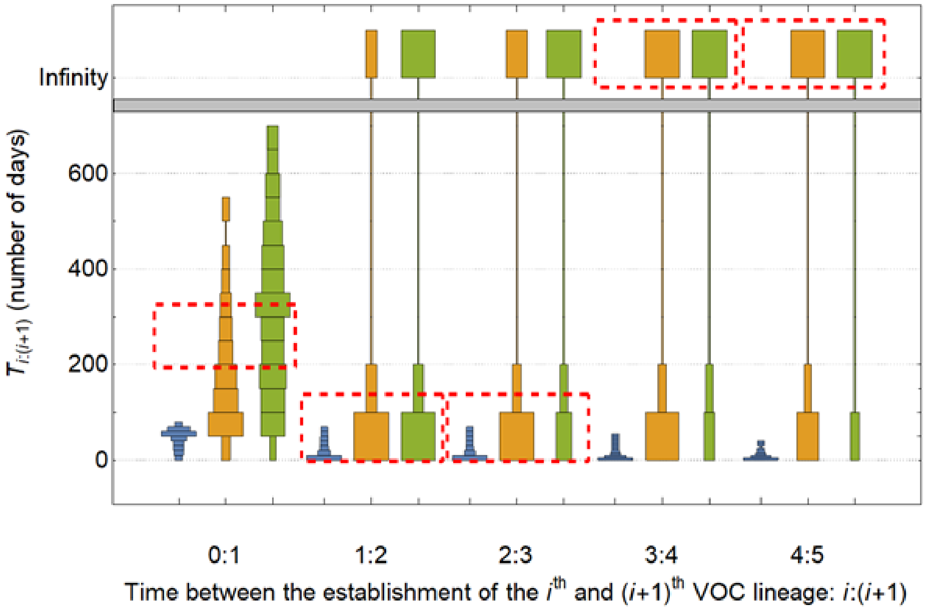


Figure C.1: Distribution of waiting times for the establishment of consecutive pairs of VOC lineages via the between-host pathway assuming a fitness landscape with a single adaptive mutation. The distribution of times that it takes between the production of thei^{th} and $(i+1)^{th}$ lineage, $T_{i:(i+1)}$, for the first 5 established VOC lineages described in Figure 3.3. $T_0 : 1$ is the waiting time for the establishment of the first VOC lineage (equivalent to T_0).

C.2 Distribution of waiting times with the within-host pathway assuming a fitness landscape with a single adaptive mutation.

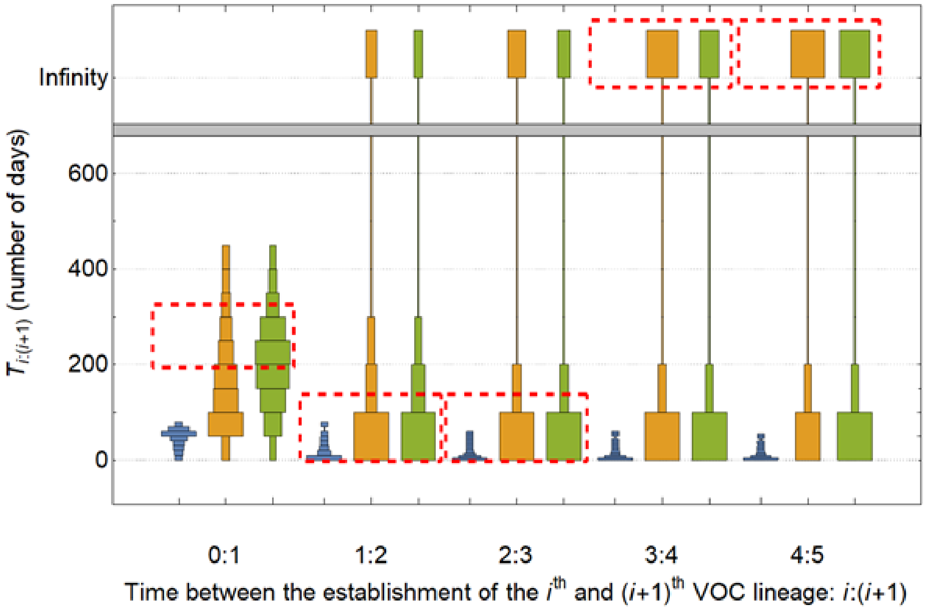


Figure C.2: Distribution of waiting times for the establishment of consecutive pairs of VOC lineages via the within-host pathway assuming a fitness landscape with a single adaptive mutation. The distribution of times that it takes between the production of the i^{th} and $(i+1)^{th}$ lineage, $T_{i:(i+1)}$, for the first 5 established VOC lineages described in Figure 3.4. $T_0 : 1$ is the waiting time for the establishment of the first VOC lineage (equivalent to T_0).

C.3 Distribution of waiting times with the between-host pathway assuming an additive fitness landscape.

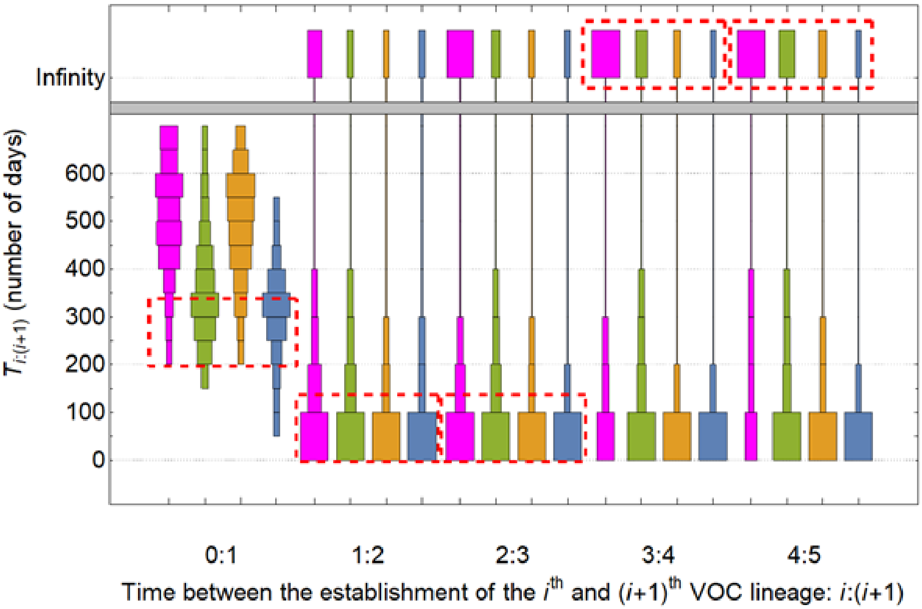


Figure C.3: Distribution of waiting times for the establishment of consecutive pairs of VOC lineages via the between-host pathway assuming an additive fitness landscape. The distribution of times that it takes between the production of the i^{th} and $(i+1)^{th}$ lineage, $T_{i:(i+1)}$, for the first 5 established VOC lineages described in Figure 3.5. $T_0 : 1$ is the waiting time for the establishment of the first VOC lineage (equivalent to T_0).

C.4 Distribution of waiting times with the within-host pathway assuming an additive fitness landscape.

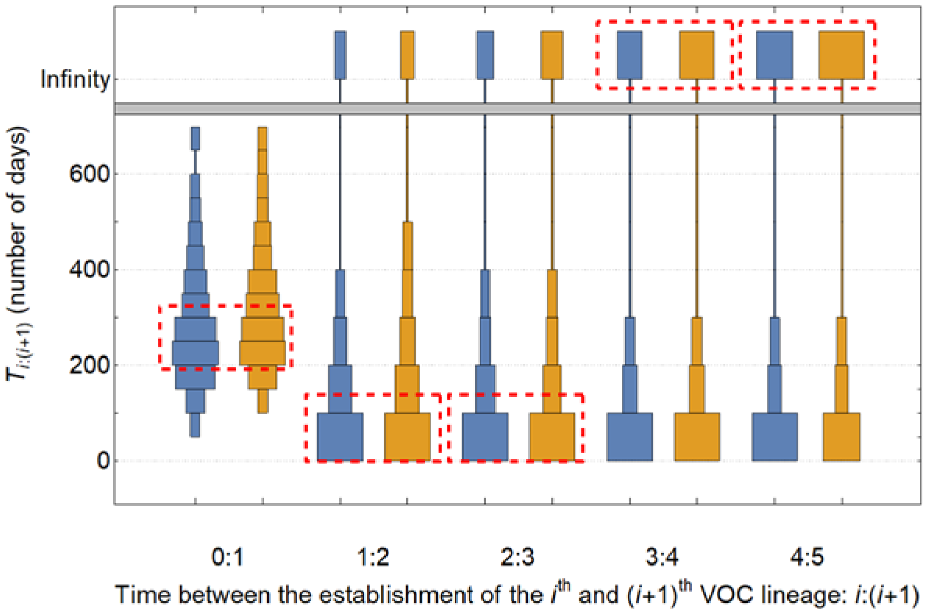


Figure C.4: Distribution of waiting times for the establishment of consecutive pairs of VOC lineages via the within-host pathway assuming an additive fitness landscape. The distribution of times that it takes between the production of the i^{th} and $(i+1)^{th}$ lineage, $T_{i:(i+1)}$, for the first 5 established VOC lineages described in Figure 3.6. $T_0 : 1$ is the waiting time for the establishment of the first VOC lineage (equivalent to T_0).

C.5 Distribution of waiting times with the between-host pathway assuming a fitness plateau landscape.

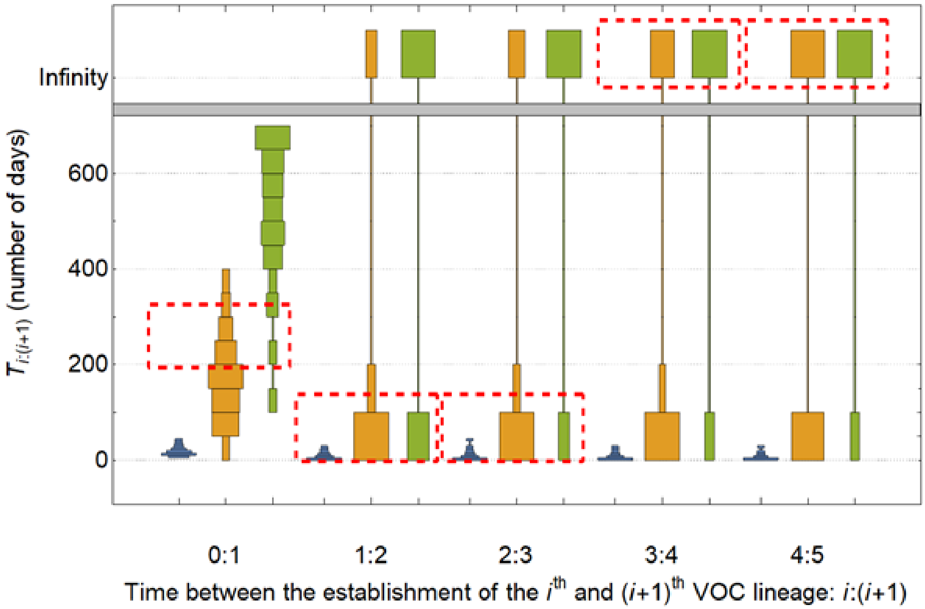


Figure C.5: Distribution of waiting times for the establishment of consecutive pairs of VOC lineages via the between-host pathway assuming a fitness plateau landscape. The distribution of times that it takes between the production of the i^{th} and $(i+1)^{th}$ lineage, $T_{i:(i+1)}$, for the first 5 established VOC lineages described in Figure 3.7. $T_0 : 1$ is the waiting time for the establishment of the first VOC lineage (equivalent to T_0).

C.6 Distribution of waiting times with the within-host pathway assuming a fitness plateau landscape.

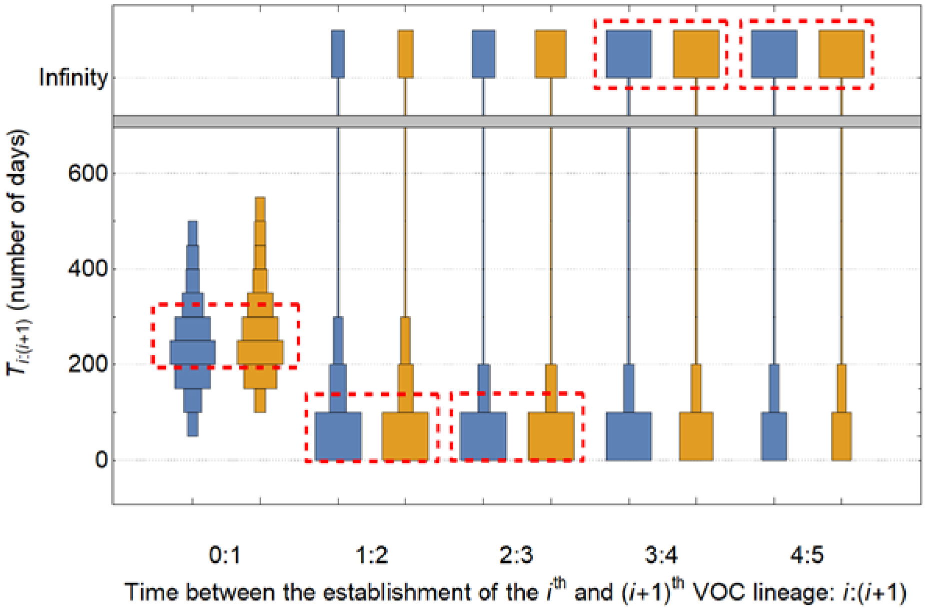
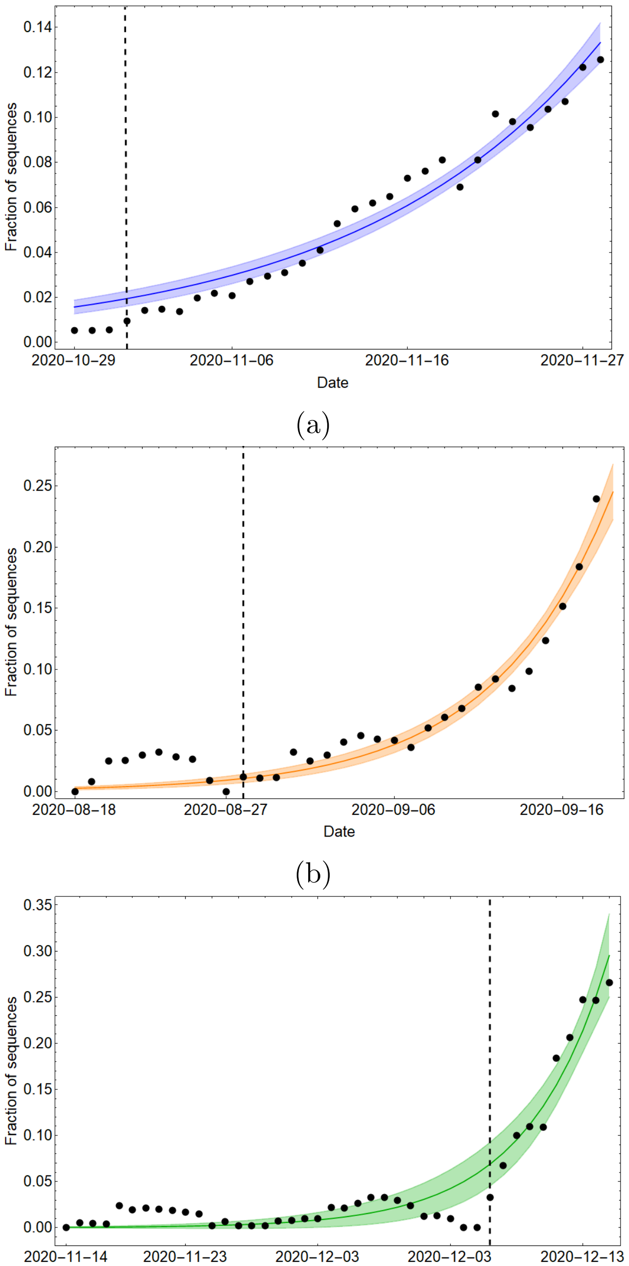


Figure C.6: Distribution of waiting times for the establishment of consecutive pairs of VOC lineages via the within-host pathway assuming a fitness plateau landscape. The distribution of times that it takes between the production of the i^{th} and $(i+1)^{th}$ lineage, $T_{i:(i+1)}$, for the first 5 established VOC lineages described in Figure 3.8. $T_0 : 1$ is the waiting time for the establishment of the first VOC lineage (equivalent to T_0).

C.7 Exponential fit for selective coefficient



(a)

(b)

(c)

Figure C.7: Exponential model fits to the frequency of individual SARS-CoV-2 VOC sequences sampled in its country of origin. **(a)-(c)**. Fitting an exponential function of the form, $f(t) = a e^{\mathrm{bt}}$, to the frequency of Alpha, Beta, and Gamma sequences sampled in the UK, South Africa, and Brazil, respectively. Vertical dashed line shows the starting timepoint used for the fitting. The shaded area shows the mean prediction bands.

Bibliography

[1] Andrew Rambaut, Nick Loman, Oliver Pybus, Wendy Barclay, Jeff Barrett, Alesandro Carabelli, Tom Connor, Tom Peacock, David L Robertson, and Erik Volz. Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations. https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563, December 2020.

[2] Stephen W. Attwood, Sarah C. Hill, David M. Aanensen, Thomas R. Connor, and Oliver G. Pybus. Phylogenetic and phylodynamic approaches to understanding and combating the early SARS-CoV-2 pandemic. *Nature Reviews. Genetics*, 23(9):547–562, September 2022. ISSN 1471-0064. doi: 10.1038/s41576-022-00483-8.

[3] R Barrett and D Schluter. Adaptation from standing genetic variation. *Trends in Ecology & Evolution*, 23(1):38–44, January 2008. ISSN 01695347. doi: 10.1016/j.tree.2007.09.008.

[4] N. H. Barton, A. M. Etheridge, and A. Véber. The infinitesimal model: Definition, derivation, and implications. *Theoretical Population Biology*, 118:50–73, December 2017. ISSN 0040-5809. doi: 10.1016/j.tpb.2017.06.001.

[5] Robert Belas, Rachel Schneider, and Michael Melch. Characterization of Pro-

teus mirabilis Precocious Swarming Mutants: Identification of rsbA, Encoding a Regulator of Swarming Behavior. *Journal of Bacteriology*, 180(23):6126–6139, December 1998. ISSN 0021-9193.

[6] C Bernstein and V Johns. Sexual reproduction as a response to H2O2 damage in Schizosaccharomyces pombe. *Journal of Bacteriology*, 171(4):1893–1897, April 1989. ISSN 0021-9193.

[7] Ernesto Berríos-Caro, Tobias Galla, and George W. A. Constable. Switching environments, synchronous sex, and the evolution of mating types. *bioRxiv*, page 2020.07.31.230482, July 2020. doi: 10.1101/2020.07.31.230482.

[8] Anne-Florence Bitbol and David J. Schwab. Quantifying the Role of Population Subdivision in Evolution on Rugged Fitness Landscapes. *PLOS Computational Biology*, 10(8):e1003778, August 2014. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1003778.

[9] Katarina M. Braun, Gage K. Moreno, Cassia Wagner, Molly A. Accola, William M. Rehrauer, David A. Baker, Katia Koelle, David H. O'Connor, Trevor Bedford, Thomas C. Friedrich, and Louise H. Moncla. Acute SARS-CoV-2 infections harbor limited within-host diversity and transmit via tight transmission bottlenecks. *PLOS Pathogens*, 17(8):e1009849, August 2021. ISSN 1553-7374. doi: 10.1371/journal.ppat.1009849.

[10] J. R. Brisbane and J. P. Gibson. Balancing selection response and rate of inbreeding by including genetic relationships in selection decisions. *Theoretical and Applied Genetics*, 91(3):421–431, August 1995. ISSN 0040-5752, 1432-2242. doi: 10.1007/BF00222969.

[11] Molly K. Burke, Joseph P. Dunham, Parvin Shahrestani, Kevin R. Thornton, Michael R. Rose, and Anthony D. Long. Genome-wide analysis of a long-term

evolution experiment with Drosophila. *Nature*, 467(7315):587–590, September 2010. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature09352.

[12] Nathan W. Burke and Russell Bonduriansky. Sexual Conflict, Facultative Asexuality, and the True Paradox of Sex. *Trends in Ecology & Evolution*, 32(9): 646–652, September 2017. ISSN 0169-5347. doi: 10.1016/j.tree.2017.06.002.

[13] A. Caballero, E. Santiago, and M. A. Toro. Systems of mating to reduce inbreeding in selected populations. *Animal Science*, 62(3):431–442, June 1996. ISSN 1357-7298, 1748-748X. doi: 10.1017/S1357729800014971.

[14] Angela M. Caliendo, David N. Gilbert, Christine C. Ginocchio, Kimberly E. Hanson, Larissa May, Thomas C. Quinn, Fred C. Tenover, David Alland, Anne J. Blaschke, Robert A. Bonomo, Karen C. Carroll, Mary Jane Ferraro, Lisa R. Hirschhorn, W. Patrick Joseph, Tobi Karchmer, Ann T. MacIntyre, L. Barth Reller, Audrey F. Jackson, and Infectious Diseases Society of America (IDSA). Better tests, better care: improved diagnostics for infectious diseases. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, 57 Suppl 3(Suppl 3):S139–170, December 2013. ISSN 1537-6591. doi: 10.1093/cid/cit578.

[15] Sandile Cele, Inbal Gazy, Laurelle Jackson, Shi-Hsia Hwa, Houriiyah Tegally, Gila Lustig, Jennifer Giandhari, Sureshnee Pillay, Eduan Wilkinson, Yeshnee Naidoo, Farina Karim, Yashica Ganga, Khadija Khan, Mallory Bernstein, Alejandro B. Balazs, Bernadett I. Gosnell, Willem Hanekom, Mahomed-Yunus S. Moosa, Network for Genomic Surveillance in South Africa, COMMIT-KZN Team, Richard J. Lessells, Tulio de Oliveira, and Alex Sigal. Escape of SARS-CoV-2 501Y.V2 from neutralization by convalescent plasma. *Nature*, 593(7857): 142–146, May 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03471-w.

[16] Jeffrey C. Chandler, Sarah N. Bevins, Jeremy W. Ellis, Timothy J. Linder, Rachel M. Tell, Melinda Jenkins-Moore, J. Jeffrey Root, Julianna B. Lenoch, Suelee Robbe-Austerman, Thomas J. DeLiberto, Thomas Gidlewski, Mia Kim Torchetti, and Susan A. Shriner. SARS-CoV-2 exposure in wild white-tailed deer (Odocoileus virginianus). *Proceedings of the National Academy of Sciences*, 118(47):e2114828118, November 2021. doi: 10.1073/pnas.2114828118.

[17] Joshua L Cherry and John Wakeley. A Diffusion Approximation for Selection and Drift in a Subdivided Population. *Genetics*, 163(1):421–428, January 2003. ISSN 1943-2631. doi: 10.1093/genetics/163.1.421.

[18] Bina Choi, Manish C. Choudhary, James Regan, Jeffrey A. Sparks, Robert F. Padera, Xueting Qiu, Isaac H. Solomon, Hsiao-Hsuan Kuo, Julie Boucau, Kathryn Bowman, U. Das Adhikari, Marisa L. Winkler, Alisa A. Mueller, Tiffany Y.-T. Hsu, Michaël Desjardins, Lindsey R. Baden, Brian T. Chan, Bruce D. Walker, Mathias Lichterfeld, Manfred Brigl, Douglas S. Kwon, Sanjat Kanjilal, Eugene T. Richardson, A. Helena Jonsson, Galit Alter, Amy K. Barczak, William P. Hanage, Xu G. Yu, Gaurav D. Gaiha, Michael S. Seaman, Manuela Cernadas, and Jonathan Z. Li. Persistence and Evolution of SARS-CoV-2 in an Immunocompromised Host. *The New England Journal of Medicine*, 383(23):2291–2293, December 2020. ISSN 1533-4406. doi: 10.1056/NEJMc2031364.

[19] C. T. Chung, S. L. Niemela, and R. H. Miller. One-step preparation of competent Escherichia coli: transformation and storage of bacterial cells in the same solution. *Proceedings of the National Academy of Sciences*, 86(7):2172–2175, April 1989. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.86.7.2172.

[20] Jean-Pierre Claverys, Marc Prudhomme, and Bernard Martin. Induction of Competence Regulons as a General Response to Stress in Gram-Positive Bacte-

ria. *Annual Review of Microbiology*, 60(1):451–475, 2006. doi: 10.1146/annurev.micro.60.080805.142139.

[21] Philippe Colson, Jérémy Delerce, Emilie Burel, Jordan Dahan, Agnès Jouffret, Florence Fenollar, Nouara Yahi, Jacques Fantini, Bernard La Scola, and Didier Raoult. Emergence in southern France of a new SARS-CoV-2 variant harbouring both N501Y and E484K substitutions in the spike protein. *Archives of Virology*, 167(4):1185–1190, April 2022. ISSN 1432-8798. doi: 10.1007/s00705-022-05385-y.

[22] Jacob D. Cooper and Benjamin Kerr. Evolution at 'Sutures' and 'Centers': Recombination Can Aid Adaptation of Spatially Structured Populations on Rugged Fitness Landscapes. *PLOS Computational Biology*, 12(12):e1005247, December 2016. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1005247.

[23] Arthur W. Covert and Claus O. Wilke. Intermediate Migration Yields Optimal Adaptation in Structured, Asexual Populations, April 2014.

[24] Jerry A. Coyne, Nicholas H. Barton, and Michael Turelli. PERSPECTIVE: A CRITIQUE OF SEWALL WRIGHT'S SHIFTING BALANCE THEORY OF EVOLUTION. *Evolution*, 51(3):643–671, June 1997. ISSN 0014-3820. doi: 10.1111/j.1558-5646.1997.tb03650.x.

[25] Kristina Crona. Recombination and peak jumping. *PLOS ONE*, 13(3):e0193123, March 2018. ISSN 1932-6203. doi: 10.1371/journal.pone.0193123.

[26] Gavin E Crooks. Logistic approximation to the logistic-normal integral. http://threeplusone.com/logistic-normal, 2013.

[27] James F. Crow and Motoo Kimura. Evolution in Sexual and Asexual Populations. *The American Naturalist*, 99(909):439–450, November 1965. ISSN 0003-0147. doi: 10.1086/282389.

[28] Ruth Daniels, Jos Vanderleyden, and Jan Michiels. Quorum sensing and swarming migration in bacteria. *FEMS Microbiology Reviews*, 28(3):261–289, June 2004. ISSN 1574-6976. doi: 10.1016/j.femsre.2003.09.004.

[29] John Davey. Fusion of a fission yeast. *Yeast*, 14(16):1529–1566, 1998. ISSN 1097-0061. doi: 10.1002/(SICI)1097-0061(199812)14:16⟨1529::AID-YEA357⟩3.0.CO;2-0.

[30] Nicholas G. Davies, Sam Abbott, Rosanna C. Barnard, Christopher I. Jarvis, Adam J. Kucharski, James D. Munday, Carl A. B. Pearson, Timothy W. Russell, Damien C. Tully, Alex D. Washburne, Tom Wenseleers, Amy Gimma, William Waites, Kerry L. M. Wong, Kevin van Zandvoort, Justin D. Silverman, CMMID COVID-19 Working Group, COVID-19 Genomics UK (COG-UK) Consortium, Karla Diaz-Ordaz, Ruth Keogh, Rosalind M. Eggo, Sebastian Funk, Mark Jit, Katherine E. Atkins, and W. John Edmunds. Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science*, 372 (6538):eabg3055, April 2021. doi: 10.1126/science.abg3055.

[31] Tulio de Oliveira, Silvia Lutucuta, John Nkengasong, Joana Morais, Joana Paula Paixão, Zoraima Neto, Pedro Afonso, Julio Miranda, Kumbelembe David, Luzia Inglês, Amilton Pereira Agostinho Paulo Raisa Rivas Carralero, Helga Reis Freitas, Franco Mufinda, Sofonias Kifle Tessema, Houriiyah Tegally, Emmanuel James San, Eduan Wilkinson, Jennifer Giandhari, Sureshnee Pillay, Marta Giovanetti, Yeshnee Naidoo, Aris Katzourakis, Mahan Ghafari, Lavanya Singh, Derek Tshiabuila, Darren Martin, and Richard J. Lessells. A novel variant of interest of SARS-CoV-2 with multiple spike mutations detected through travel surveillance in Africa, April 2021.

[32] Michael M. Desai and Daniel S. Fisher. Beneficial Mutation–Selection Balance

and the Effect of Linkage on Positive Selection. *Genetics*, 176(3):1759–1798, July 2007. ISSN 0016-6731, 1943-2631. doi: 10.1534/genetics.106.067678.

[33] Rachel T. Eguia, Katharine H. D. Crawford, Terry Stevens-Ayers, Laurel Kelnhofer-Millevolte, Alexander L. Greninger, Janet A. Englund, Michael J. Boeckh, and Jesse D. Bloom. A human coronavirus evolves antigenically to escape antibody immunity. *PLOS Pathogens*, 17(4):e1009453, April 2021. ISSN 1553-7374. doi: 10.1371/journal.ppat.1009453.

[34] Akira Endo, Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group, Sam Abbott, Adam J. Kucharski, and Sebastian Funk. Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China. *Wellcome Open Research*, 5:67, 2020. ISSN 2398-502X. doi: 10.12688/wellcomeopenres.15842.3.

[35] Nuno R. Faria, Thomas A. Mellan, Charles Whittaker, Ingra M. Claro, Darlan da S. Candido, Swapnil Mishra, Myuki A. E. Crispim, Flavia C. S. Sales, Iwona Hawryluk, John T. McCrone, Ruben J. G. Hulswit, Lucas A. M. Franco, Mariana S. Ramundo, Jaqueline G. de Jesus, Pamela S. Andrade, Thais M. Coletti, Giulia M. Ferreira, Camila A. M. Silva, Erika R. Manuli, Rafael H. M. Pereira, Pedro S. Peixoto, Moritz U. G. Kraemer, Nelson Gaburo, Cecilia da C. Camilo, Henrique Hoeltgebaum, William M. Souza, Esmenia C. Rocha, Leandro M. de Souza, Mariana C. de Pinho, Leonardo J. T. Araujo, Frederico S. V. Malta, Aline B. de Lima, Joice do P. Silva, Danielle A. G. Zauli, Alessandro C. de S. Ferreira, Ricardo P. Schnekenberg, Daniel J. Laydon, Patrick G. T. Walker, Hannah M. Schlüter, Ana L. P. Dos Santos, Maria S. Vidal, Valentina S. Del Caro, Rosinaldo M. F. Filho, Helem M. Dos Santos, Renato S. Aguiar, José L. Proença-Modena, Bruce Nelson, James A. Hay, Mélodie Monod, Xenia Miscouridou, Helen Coupland, Raphael Sonabend, Michaela Vollmer, Axel

Gandy, Carlos A. Prete, Vitor H. Nascimento, Marc A. Suchard, Thomas A. Bowden, Sergei L. K. Pond, Chieh-Hsi Wu, Oliver Ratmann, Neil M. Ferguson, Christopher Dye, Nick J. Loman, Philippe Lemey, Andrew Rambaut, Nelson A. Fraiji, Maria do P. S. S. Carvalho, Oliver G. Pybus, Seth Flaxman, Samir Bhatt, and Ester C. Sabino. Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. *Science (New York, N.Y.)*, 372(6544):815–821, May 2021. ISSN 1095-9203. doi: 10.1126/science.abh2644.

[36] Luca Ferretti, Chris Wymant, Michelle Kendall, Lele Zhao, Anel Nurtay, Lucie Abeler-Dörner, Michael Parker, David Bonsall, and Christophe Fraser. Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. *Science (New York, N.Y.)*, 368(6491):eabb6936, May 2020. ISSN 1095-9203. doi: 10.1126/science.abb6936.

[37] R. A. Fisher. *The genetical theory of natural selection.* Oxford Press, 1930. ISBN 978-1-176-62502-0.

[38] Patricia L. Foster. Stress responses and genetic variation in bacteria. *Mutation Research*, 569(1-2):3–11, January 2005. ISSN 0027-5107. doi: 10.1016/j.mrfmmm.2004.07.017.

[39] Marcus Frean, Paul B. Rainey, and Arne Traulsen. The effect of population structure on the rate of evolution. *Proceedings of the Royal Society B: Biological Sciences*, 280(1762):20130211, July 2013. ISSN 0962-8452. doi: 10.1098/rspb.2013.0211.

[40] Elisabeth Bautz Freese, Martha I. Chu, and Ernst Freese. Initiation of Yeast Sporulation by Partial Carbon, Nitrogen, or Phosphate Deprivation. *Journal of Bacteriology*, 149(3):840–851, March 1982. ISSN 0021-9193.

[41] Philip J. Gerrish and Richard E. Lenski. The fate of competing beneficial

mutations in an asexual population. *Genetica*, 102(0):127, March 1998. ISSN 1573-6857. doi: 10.1023/A:1017067816551.

[42] Mahan Ghafari, Bardia Hejazi, Arman Karshenas, Stefan Dascalu, Alireza Kadvidar, Mohammad A. Khosravi, Maryam Abbasalipour, Majid Heydari, Sirous Zeinali, Luca Ferretti, Alice Ledda, and Aris Katzourakis. Lessons for preparedness and reasons for concern from the early COVID-19 epidemic in Iran. *Epidemics*, 36:100472, September 2021. ISSN 1878-0067. doi: 10.1016/j.epidem.2021.100472.

[43] Mahan Ghafari, Louis du Plessis, Jayna Raghwani, Samir Bhatt, Bo Xu, Oliver G Pybus, and Aris Katzourakis. Purifying Selection Determines the Short-Term Time Dependency of Evolutionary Rates in SARS-CoV-2 and pH1N1 Influenza. *Molecular Biology and Evolution*, 39(2):msac009, February 2022. ISSN 1537-1719. doi: 10.1093/molbev/msac009.

[44] John H Gillespie. NATURAL SELECTION FOR WITHIN-GENERATION VARIANCE IN OFFSPRING NUMBER. *Genetics*, 76(3):601–606, March 1974. ISSN 1943-2631. doi: 10.1093/genetics/76.3.601.

[45] Ana S. Gonzalez-Reiche, Hala Alshammary, Sarah Schaefer, Gopi Patel, Jose Polanco, Angela A. Amoako, Aria Rooker, Christian Cognigni, Daniel Floda, Adriana van de Guchte, Zain Khalil, Keith Farrugia, Nima Assad, Jian Zhang, Bremy Alburquerque, Levy Sominsky, Komal Srivastava, Robert Sebra, Juan David Ramirez, Radhika Banu, Paras Shrestha, Alberto Paniz-Mondolfi, Emilia Mia Sordillo, Viviana Simon, and Harm van Bakel. Intrahost evolution and forward transmission of a novel SARS-CoV-2 Omicron BA.1 subvariant, May 2022. ISSN 2227-5533.

[46] Tiago Gräf, Gonzalo Bello, Taina Moreira Martins Venas, Elisa Cavalcante

Pereira, Anna Carolina Dias Paixão, Luciana Reis Appolinario, Renata Serrano Lopes, Ana Carolina Da Fonseca Mendonça, Alice Sampaio Barreto da Rocha, Fernando Couto Motta, Tatiana Schäffer Gregianini, Richard Steiner Salvato, Sandra Bianchini Fernandes, Darcita Buerger Rovaris, Andrea Cony Cavalcanti, Anderson Brandão Leite, Irina Riediger, Maria do Carmo Debur, André Felipe Leal Bernardes, Rodrigo Ribeiro-Rodrigues, Beatriz Grinsztejn, Valdinete Alves do Nascimento, Victor Costa de Souza, Luciana Gonçalves, Cristiano Fernandes da Costa, Tirza Mattos, Filipe Zimmer Dezordi, Gabriel Luz Wallau, Felipe Gomes Naveca, Edson Delatorre, Marilda Mendonça Siqueira, Paola Cristina Resende, and Fiocruz COVID-19 Genomic Surveillance Network. Identification of a novel SARS-CoV-2 P.1 sub-lineage in Brazil provides new insights about the mechanisms of emergence of variants of concern. *Virus Evolution*, 7(2):veab091, July 2022. ISSN 2057-1577. doi: 10.1093/ve/veab091.

[47] Rachel L. Graham and Ralph S. Baric. Recombination, Reservoirs, and the Modular Spike: Mechanisms of Coronavirus Cross-Species Transmission. *Journal of Virology*, 84(7):3134–3146, April 2010. doi: 10.1128/JVI.01394-09.

[48] B. Grundy, B. Villanueva, and J. A. Woolliams. Dynamic selection procedures for constrained inbreeding and their consequences for pedigree development. *Genetical Research*, 72(2):159–168, October 1998. ISSN 0016-6723, 1469-5073. doi: 10.1017/S0016672398003474.

[49] Bernardo Gutierrez, Hugo G. Castelán Sánchez, Darlan da Silva Candido, Ben Jackson, Shay Fleishon, Renaud Houzet, Christopher Ruis, Luis Delaye, Nuno R. Faria, Andrew Rambaut, Oliver G. Pybus, and Marina Escalera-Zamudio. Emergence and widespread circulation of a recombinant SARS-CoV-2 lineage in North America. *Cell Host & Microbe*, 30(8):1112–1123.e3, August 2022. ISSN 1931-3128. doi: 10.1016/j.chom.2022.06.010.

[50] James Hadfield, Colin Megill, Sidney M. Bell, John Huddleston, Barney Potter, Charlton Callender, Pavel Sagulenko, Trevor Bedford, and Richard A. Neher. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics (Oxford, England)*, 34(23):4121–4123, December 2018. ISSN 1367-4811. doi: 10.1093/bioinformatics/bty407.

[51] Verity Hill, Louis Du Plessis, Thomas P Peacock, Dinesh Aggarwal, Rachel Colquhoun, Alesandro M Carabelli, Nicholas Ellaby, Eileen Gallagher, Natalie Groves, Ben Jackson, J T McCrone, Áine O'Toole, Anna Price, Theo Sanderson, Emily Scher, Joel Southgate, Erik Volz, Wendy S Barclay, Jeffrey C Barrett, Meera Chand, Thomas Connor, Ian Goodfellow, Ravindra K Gupta, Ewan M Harrison, Nicholas Loman, Richard Myers, David L Robertson, Oliver G Pybus, Andrew Rambaut, and The COVID-19 Genomics UK (COG-UK) Consortium. The origins and molecular evolution of SARS-CoV-2 lineage B.1.1.7 in the UK. *Virus Evolution*, 8(2):veac080, July 2022. ISSN 2057-1577. doi: 10.1093/ve/veac080.

[52] W. G. Hill and Alan Robertson. The effect of linkage on limits to artificial selection. *Genetics Research*, 8(3):269–294, December 1966. ISSN 1469-5073, 0016-6723. doi: 10.1017/S0016672300010156.

[53] William G. Hill. Design and Efficiency of Selection Experiments for Estimating Genetic Parameters. *Biometrics*, 27(2):293–311, 1971. ISSN 0006-341X. doi: 10.2307/2528996.

[54] William G Hill. Maintenance of quantitative genetic variation in animal breeding programmes. *Livestock Production Science*, 63(2):99–109, April 2000. ISSN 03016226. doi: 10.1016/S0301-6226(99)00115-3.

[55] Emma Hodcroft. CoVariants: SARS-CoV-2 Mutations and Variants of Interest. https://covariants.org/, 2021.

[56] Ben Jackson, Maciej F. Boni, Matthew J. Bull, Amy Colleran, Rachel M. Colquhoun, Alistair C. Darby, Sam Haldenby, Verity Hill, Anita Lucaci, John T. McCrone, Samuel M. Nicholls, Áine O'Toole, Nicole Pacchiarini, Radoslaw Poplawski, Emily Scher, Flora Todd, Hermione J. Webster, Mark Whitehead, Claudia Wierzbicki, COVID-19 Genomics UK (COG-UK) Consortium, Nicholas J. Loman, Thomas R. Connor, David L. Robertson, Oliver G. Pybus, and Andrew Rambaut. Generation and transmission of interlineage recombinants in the SARS-CoV-2 pandemic. *Cell*, 184(20):5179–5188.e8, September 2021. ISSN 1097-4172. doi: 10.1016/j.cell.2021.08.014.

[57] Talia Karasov, Philipp W. Messer, and Dmitri A. Petrov. Evidence that Adaptation in Drosophila Is Not Limited by Mutation at Single Sites. *PLoS Genetics*, 6(6):e1000924, June 2010. ISSN 1553-7404. doi: 10.1371/journal.pgen.1000924.

[58] F. Karim, M. Y. S. Moosa, B. I. Gosnell, S. Cele, J. Giandhari, S. Pillay, H. Tegally, E. Wilkinson, J. E. San, N. Msomi, K. Mlisana, K. Khan, M. Bernstein, N. Manickchund, L. Singh, U. Ramphal, Commit-Kzn Team, W. Hanekom, R. J. Lessells, A. Sigal, and T. de Oliveira. Persistent SARS-CoV-2 infection and intra-host evolution in association with advanced HIV infection, June 2021. ISSN 2125-8228.

[59] Ariel Karlinsky and Dmitry Kobak. Tracking excess mortality across countries during the COVID-19 pandemic with the World Mortality Dataset. *eLife*, 10: e69336, June 2021. ISSN 2050-084X. doi: 10.7554/eLife.69336.

[60] Yona Kassir, David Granot, and Giora Simchen. IME1, a positive regulator gene

of meiosis in S. cerevisiae. *Cell*, 52(6):853–862, March 1988. ISSN 0092-8674, 1097-4172. doi: 10.1016/0092-8674(88)90427-8.

[61] Tadeusz J. Kawecki, Richard E. Lenski, Dieter Ebert, Brian Hollis, Isabelle Olivieri, and Michael C. Whitlock. Experimental evolution. *Trends in Ecology & Evolution*, 27(10):547–560, October 2012. ISSN 0169-5347. doi: 10.1016/j.tree.2012.06.001.

[62] Steven A. Kemp, Dami A. Collier, Rawlings P. Datir, Isabella A. T. M. Ferreira, Salma Gayed, Aminu Jahun, Myra Hosmillo, Chloe Rees-Spear, Petra Mlcochova, Ines Ushiro Lumb, David J. Roberts, Anita Chandra, Nigel Temperton, CITIID-NIHR BioResource COVID-19 Collaboration, COVID-19 Genomics UK (COG-UK) Consortium, Katherine Sharrocks, Elizabeth Blane, Yorgo Modis, Kendra E. Leigh, John A. G. Briggs, Marit J. van Gils, Kenneth G. C. Smith, John R. Bradley, Chris Smith, Rainer Doffinger, Lourdes Ceron-Gutierrez, Gabriela Barcenas-Morales, David D. Pollock, Richard A. Goldstein, Anna Smielewska, Jordan P. Skittrall, Theodore Gouliouris, Ian G. Goodfellow, Effrossyni Gkrania-Klotsas, Christopher J. R. Illingworth, Laura E. McCoy, and Ravindra K. Gupta. SARS-CoV-2 evolution during treatment of chronic infection. *Nature*, 592(7853):277–282, April 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03291-y.

[63] Motoo Kimura. On the Probability of Fixation of Mutant Genes in a Population. *Genetics*, 47(6):713–719, June 1962. ISSN 0016-6731.

[64] Hanna Kokko. When Synchrony Makes the Best of Both Worlds Even Better: How Well Do We Really Understand Facultative Sex? *The American Naturalist*, pages 000–000, October 2019. ISSN 0003-0147. doi: 10.1086/706812.

[65] Sergey Kryazhimskiy, Daniel P. Rice, and Michael M. Desai. Population sub-

division and adaptation in asexual populations of saccharomyces cerevisiae: subdivision and adaptation in asexual populations of yeast. *Evolution*, 66(6): 1931–1941, June 2012. ISSN 00143820. doi: 10.1111/j.1558-5646.2011.01569.x.

[66] Suresh V. Kuchipudi, Meera Surendran-Nair, Rachel M. Ruden, Michele Yon, Ruth H. Nissly, Kurt J. Vandegrift, Rahul K. Nelli, Lingling Li, Bhushan M. Jayarao, Costas D. Maranas, Nicole Levine, Katriina Willgert, Andrew J. K. Conlan, Randall J. Olsen, James J. Davis, James M. Musser, Peter J. Hudson, and Vivek Kapur. Multiple spillovers from humans and onward transmission of SARS-CoV-2 in white-tailed deer. *Proceedings of the National Academy of Sciences of the United States of America*, 119(6):e2121644119, February 2022. ISSN 1091-6490. doi: 10.1073/pnas.2121644119.

[67] Richard E Lenski. Experimental evolution and the dynamics of adaptation and genome evolution in microbial populations. *The ISME Journal*, 11(10):2181–2194, October 2017. ISSN 1751-7362, 1751-7370. doi: 10.1038/ismej.2017.69.

[68] Andrew T. Levin, Nana Owusu-Boaitey, Sierra Pugh, Bailey K. Fosdick, Anthony B. Zwi, Anup Malani, Satej Soman, Lonni Besançon, Ilya Kashnitsky, Sachin Ganesh, Aloysius McLaughlin, Gayeong Song, Rine Uhm, Daniel Herrera-Esposito, Gustavo de los Campos, Ana Carolina Peçanha Antonio, Enyew Birru Tadese, and Gideon Meyerowitz-Katz. Assessing the burden of COVID-19 in developing countries: systematic review, meta-analysis and public policy implications. *BMJ Global Health*, 7(5):e008477, May 2022. ISSN 2059-7908. doi: 10.1136/bmjgh-2022-008477.

[69] R C Lewontin. *The Genetic Basis of Evolutionary Change*. Columbia University Press, 1974.

[70] Yang Liu, Jianying Liu, Kenneth S. Plante, Jessica A. Plante, Xuping Xie,

Xianwen Zhang, Zhiqiang Ku, Zhiqiang An, Dionna Scharton, Craig Schindewolf, Steven G. Widen, Vineet D. Menachery, Pei-Yong Shi, and Scott C. Weaver. The N501Y spike substitution enhances SARS-CoV-2 infection and transmission. *Nature*, 602(7896):294–299, February 2022. ISSN 1476-4687. doi: 10.1038/s41586-021-04245-0.

[71] J. O. Lloyd-Smith, S. J. Schreiber, P. E. Kopp, and W. M. Getz. Superspreading and the effect of individual variation on disease emergence. *Nature*, 438(7066): 355–359, November 2005. ISSN 1476-4687. doi: 10.1038/nature04153.

[72] M. G. Lorenz and W. Wackernagel. Bacterial gene transfer by natural genetic transformation in the environment. *Microbiological Reviews*, 58(3):563–602, September 1994. ISSN 0146-0749.

[73] Katrina A. Lythgoe, Matthew Hall, Luca Ferretti, Mariateresa de Cesare, George MacIntyre-Cockett, Amy Trebes, Monique Andersson, Newton Otecko, Emma L. Wise, Nathan Moore, Jessica Lynch, Stephen Kidd, Nicholas Cortes, Matilde Mori, Rebecca Williams, Gabrielle Vernet, Anita Justice, Angie Green, Samuel M. Nicholls, M. Azim Ansari, Lucie Abeler-Dörner, Catrin E. Moore, Timothy E. A. Peto, David W. Eyre, Robert Shaw, Peter Simmonds, David Buck, John A. Todd, on behalf of the Oxford Virus Sequencing Analysis Group (OVSG), Thomas R. Connor, Shirin Ashraf, Ana da Silva Filipe, James Shepherd, Emma C. Thomson, The COVID-19 Genomics UK (COG-UK) Consortium, David Bonsall, Christophe Fraser, and Tanya Golubchik. SARS-CoV-2 within-host diversity and transmission. *Science*, 372(6539):eabg0821, April 2021. doi: 10.1126/science.abg0821.

[74] Katrina A. Lythgoe, Tanya Golubchik, Matthew Hall, Thomas House, George MacIntyre-Cockett, Helen Fryer, Laura Thomson, Anel Nurtay, David Buck, Angie Green, Amy Trebes, Paolo Piazza, Lorne J. Lonie, Ruth Studley, Emma

Rourke, Duncan Cook, Darren Smith, Matthew Bashton, Andrew Nelson, Matthew Crown, Clare McCann, Gregory R. Young, Rui Andre Nunes dos Santos, Zack Richards, Adnan Tariq, Wellcome Sanger Institute COVID-19 Surveillance Team, COVID-19 Infection Survey Group, The COVID-19 Genomics UK (COG-UK) Consortium, Christophe Fraser, Ian Diamond, Jeff Barrett, Sarah Walker, and David Bonsall. Lineage replacement and evolution captured by the United Kingdom Covid Infection Survey, January 2022.

[75] Spyros Lytras, Wei Xia, Joseph Hughes, Xiaowei Jiang, and David L. Robertson. The animal origin of SARS-CoV-2. *Science*, August 2021. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.abh0117.

[76] M. Cyrus Maher, Istvan Bartha, Steven Weaver, Julia di Iulio, Elena Ferri, Leah Soriaga, Florian A. Lempp, Brian L. Hie, Bryan Bryson, Bonnie Berger, David L. Robertson, Gyorgy Snell, Davide Corti, Herbert W. Virgin, Sergei L. Kosakovsky Pond, and Amalio Telenti. Predicting the mutational drivers of future SARS-CoV-2 variants of concern. *Science Translational Medicine*, 14 (633):eabk3445, January 2022. doi: 10.1126/scitranslmed.abk3445.

[77] B. Mai and L. Breeden. CLN1 and its repression by Xbp1 are important for efficient sporulation in budding yeast. *Molecular and Cellular Biology*, 20(2): 478–487, January 2000. ISSN 0270-7306. doi: 10.1128/mcb.20.2.478-487.2000.

[78] Erik A Martens and Oskar Hallatschek. Interfering Waves of Adaptation Promote Spatial Mixing. *Genetics*, 189(3):1045–1060, November 2011. ISSN 1943-2631. doi: 10.1534/genetics.111.130112.

[79] Darren P. Martin, Steven Weaver, Houriiyah Tegally, James Emmanuel San, Stephen D. Shank, Eduan Wilkinson, Alexander G. Lucaci, Jennifer Giandhari, Sureshnee Naidoo, Yeshnee Pillay, Lavanya Singh, Richard J. Lessells, Ravin-

dra K. Gupta, Joel O. Wertheim, Anton Nekturenko, Ben Murrell, Gordon W. Harkins, Philippe Lemey, Oscar A. MacLean, David L. Robertson, Tulio de Oliveira, and Sergei L. Kosakovsky Pond. The emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages. *Cell*, 184(20):5189–5200.e7, September 2021. ISSN 0092-8674. doi: 10.1016/j.cell.2021.09.003.

[80] Michael A. Martin and Katia Koelle. Comment on “Genomic epidemiology of superspreading events in Austria reveals mutational dynamics and transmission properties of SARS-CoV-2”. *Science Translational Medicine*, 13(617):eabh1803, October 2021. doi: 10.1126/scitranslmed.abh1803.

[81] Carlos Martorell and Marcela Martínez-López. Informed dispersal in plants: Heterosperma pinnatum (Asteraceae) adjusts its dispersal mode to escape from competition and water stress. *Oikos*, 123(2):225–231, 2014. ISSN 1600-0706. doi: 10.1111/j.1600-0706.2013.00715.x.

[82] Takeo Maruyama. On the fixation probability of mutant genes in a subdivided population*. *Genetics Research*, 15(2):221–225, April 1970. ISSN 1469-5073, 0016-6723. doi: 10.1017/S0016672300001543.

[83] Michael J. McDonald, Daniel P. Rice, and Michael M. Desai. Sex Speeds Adaptation by Altering the Dynamics of Molecular Evolution. *Nature*, 531(7593): 233–236, March 2016. ISSN 0028-0836. doi: 10.1038/nature17143.

[84] Gil McVean, Philip Awadalla, and Paul Fearnhead. A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics*, 160 (3):1231–1241, March 2002. ISSN 0016-6731.

[85] James G. Mitchell and Kazuhiro Kogure. Bacterial motility: links to the environment and a driving force for microbial physics. *FEMS Microbiology Ecology*,

55(1):3–16, January 2006. ISSN 0168-6496. doi: 10.1111/j.1574-6941.2005.00003.x.

[86] H. J. Muller. Some Genetic Aspects of Sex. *The American Naturalist*, 66(703): 118–138, March 1932. ISSN 0003-0147, 1537-5323. doi: 10.1086/280418.

[87] Joshua R. Nahum, Peter Godfrey-Smith, Brittany N. Harding, Joseph H. Marcus, Jared Carlson-Stevermer, and Benjamin Kerr. A tortoise–hare pattern seen in adapting structured and unstructured populations suggests a rugged fitness landscape in bacteria. *Proceedings of the National Academy of Sciences*, 112(24):7530–7535, June 2015. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1410631112.

[88] R. A. Neher, B. I. Shraiman, and D. S. Fisher. Rate of Adaptation in Large Sexual Populations. *Genetics*, 184(2):467–481, February 2010. ISSN 0016-6731, 1943-2631. doi: 10.1534/genetics.109.109009.

[89] Aaron M. Neiman. Sporulation in the Budding Yeast Saccharomyces cerevisiae. *Genetics*, 189(3):737–765, November 2011. ISSN 0016-6731, 1943-2631. doi: 10.1534/genetics.111.127126.

[90] Gard Nelson, Oleksandr Buzko, Patricia Spilman, Kayvan Niazi, Shahrooz Rabizadeh, and Patrick Soon-Shiong. Molecular dynamic simulation reveals E484K mutation enhances spike RBD-ACE2 affinity and the combination of E484K, K417N and N501Y mutations (501Y.V2 variant) induces conformational change greater than N501Y mutant alone, potentially resulting in an escape mutant, January 2021.

[91] Edward W. Ng and Murray Geller. A table of integrals of the Error functions. *Journal of Research of the National Bureau of Standards, Section*

B: Mathematical Sciences, 73B(1):1, January 1969. ISSN 0098-8979. doi: 10.6028/jres.073B.001.

[92] Megan O'Driscoll, Gabriel Ribeiro Dos Santos, Lin Wang, Derek A. T. Cummings, Andrew S. Azman, Juliette Paireau, Arnaud Fontanet, Simon Cauchemez, and Henrik Salje. Age-specific mortality and immunity patterns of SARS-CoV-2. *Nature*, 590(7844):140–145, February 2021. ISSN 1476-4687. doi: 10.1038/s41586-020-2918-0.

[93] Sarah P. Otto. The Evolutionary Enigma of Sex. *The American Naturalist*, 174 (S1):S1–S14, July 2009. ISSN 0003-0147. doi: 10.1086/599084.

[94] Sarah P. Otto and Thomas Lenormand. Resolving the paradox of sex and recombination. *Nature Reviews Genetics*, 3(4):252–261, April 2002. ISSN 1471-0064. doi: 10.1038/nrg761.

[95] Sarah P. Otto, Troy Day, Julien Arino, Caroline Colijn, Jonathan Dushoff, Michael Li, Samir Mechai, Gary Van Domselaar, Jianhong Wu, David J. D. Earn, and Nicholas H. Ogden. The origins and potential future of SARS-CoV-2 variants of concern in the evolving COVID-19 pandemic. *Current biology: CB*, 31(14):R918–R929, July 2021. ISSN 1879-0445. doi: 10.1016/j.cub.2021.06.049.

[96] Bas B. Oude Munnink, Reina S. Sikkema, David F. Nieuwenhuijse, Robert Jan Molenaar, Emmanuelle Munger, Richard Molenkamp, Arco van der Spek, Paulien Tolsma, Ariene Rietveld, Miranda Brouwer, Noortje Bouwmeester-Vincken, Frank Harders, Renate Hakze-van der Honing, Marjolein C. A. Wegdam-Blans, Ruth J. Bouwstra, Corine GeurtsvanKessel, Annemiek A. van der Eijk, Francisca C. Velkers, Lidwien A. M. Smit, Arjan Stegeman, Wim H. M. van der Poel, and Marion P. G. Koopmans. Transmission of SARS-CoV-2 on mink farms between humans and mink and back to humans. *Science*

(New York, N.Y.), 371(6525):172–177, January 2021. ISSN 1095-9203. doi: 10.1126/science.abe5901.

[97] Su-Chan Park and Joachim Krug. Clonal interference in large populations. *Proceedings of the National Academy of Sciences of the United States of America*, 104(46):18135–18140, November 2007. ISSN 1091-6490. doi: 10.1073/pnas.0705778104.

[98] Z Patwa and L.M Wahl. The fixation probability of beneficial mutations. *Journal of The Royal Society Interface*, 5(28):1279–1289, November 2008. ISSN 1742-5689, 1742-5662. doi: 10.1098/rsif.2008.0248.

[99] Thomas P. Peacock, Rebekah Penrice-Randal, Julian A. Hiscox, and Wendy S. Barclay. SARS-CoV-2 one year on: evidence for ongoing viral adaptation. *The Journal of General Virology*, 102(4):001584, April 2021. ISSN 1465-2099. doi: 10.1099/jgv.0.001584.

[100] Michael T. Pearce and Daniel S. Fisher. Rapid adaptation in large populations with very rare sex: Scalings and spontaneous oscillations. *Theoretical Population Biology*, 129:18–40, October 2019. ISSN 00405809. doi: 10.1016/j.tpb.2017.11.005.

[101] Lilia Perfeito, M. Inês Pereira, Paulo R.A Campos, and Isabel Gordo. The effect of spatial structure on adaptation in Escherichia coli. *Biology Letters*, 4 (1):57–59, November 2007. doi: 10.1098/rsbl.2007.0481.

[102] Bradley Pickering, Oliver Lung, Finlay Maguire, Peter Kruczkiewicz, Jonathon D. Kotwa, Tore Buchanan, Marianne Gagnier, Jennifer L. Guthrie, Claire M. Jardine, Alex Marchand-Austin, Ariane Massé, Heather McClinchey, Kuganya Nirmalarajah, Patryk Aftanas, Juliette Blais-Savoie, Hsien-Yao Chee, Emily Chien, Winfield Yim, Andra Banete, Bryan D. Griffin, Lily Yip, Melissa

Goolia, Matthew Suderman, Mathieu Pinette, Greg Smith, Daniel Sullivan, Josip Rudar, Oksana Vernygora, Elizabeth Adey, Michelle Nebroski, Guillaume Goyette, Andrés Finzi, Geneviève Laroche, Ardeshir Ariana, Brett Vahkal, Marceline Côté, Allison J. McGeer, Larissa Nituch, Samira Mubareka, and Jeff Bowman. Divergent SARS-CoV-2 variant emerges in white-tailed deer with deer-to-human transmission. *Nature Microbiology*, 7(12):2011–2024, December 2022. ISSN 2058-5276. doi: 10.1038/s41564-022-01268-9.

[103] Delphine Planas, Nell Saunders, Piet Maes, Florence Guivel-Benhassine, Cyril Planchais, Julian Buchrieser, William-Henry Bolland, Françoise Porrot, Isabelle Staropoli, Frederic Lemoine, Hélène Péré, David Veyer, Julien Puech, Julien Rodary, Guy Baele, Simon Dellicour, Joren Raymenants, Sarah Gorissen, Caspar Geenen, Bert Vanmechelen, Tony Wawina-Bokalanga, Joan Martí-Carreras, Lize Cuypers, Aymeric Sève, Laurent Hocqueloux, Thierry Prazuck, Félix A. Rey, Etienne Simon-Loriere, Timothée Bruel, Hugo Mouquet, Emmanuel André, and Olivier Schwartz. Considerable escape of SARS-CoV-2 Omicron to antibody neutralization. *Nature*, 602(7898):671–675, February 2022. ISSN 1476-4687. doi: 10.1038/s41586-021-04389-z.

[104] R. J. Redfield. Genes for Breakfast: The Have-Your-Cake and-Eat-lt-Too of Bacterial Transformation. *Journal of Heredity*, 84(5):400–404, September 1993. ISSN 0022-1503. doi: 10.1093/oxfordjournals.jhered.a111361.

[105] Alan Robertson. A theory of limits in artificial selection. *Proceedings of the Royal Society of London*, 153(951):234–249, 1960. doi: 10.1098/rspb.1960.0099.

[106] Alan Robertson. Inbreeding in artificial selection programmes. *Genetical Research*, 2(2):189–194, July 1961. ISSN 0016-6723, 1469-5073. doi: 10.1017/S0016672300000690.

[107] Alan Robertson. Some optimum problems in individual selection. *Theoretical Population Biology*, 1(1):120–127, May 1970. ISSN 0040-5809. doi: 10.1016/0040-5809(70)90045-6.

[108] Denis Roze and Nick H. Barton. The Hill–Robertson Effect and the Evolution of Recombination. *Genetics*, 173(3):1793–1811, July 2006. ISSN 0016-6731. doi: 10.1534/genetics.106.058586.

[109] J. Maynard Smith. Optimization Theory in Evolution. *Annual Review of Ecology and Systematics*, 9(1):31–56, November 1978. ISSN 0066-4162. doi: 10.1146/annurev.es.09.110178.000335.

[110] Kwangmin Son, Douglas R. Brumley, and Roman Stocker. Live from under the lens: exploring microbial motility with dynamic imaging and microfluidics. *Nature Reviews Microbiology*, 13(12):761–775, December 2015. ISSN 1740-1534. doi: 10.1038/nrmicro3567.

[111] Tyler N. Starr, Allison J. Greaney, Sarah K. Hilton, Daniel Ellis, Katharine H. D. Crawford, Adam S. Dingens, Mary Jane Navarro, John E. Bowen, M. Alejandra Tortorici, Alexandra C. Walls, Neil P. King, David Veesler, and Jesse D. Bloom. Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell*, 182(5):1295–1310.e20, September 2020. ISSN 1097-4172. doi: 10.1016/j.cell.2020.08.012.

[112] Kaiming Tao, Philip L. Tzou, Janin Nouhin, Ravindra K. Gupta, Tulio de Oliveira, Sergei L. Kosakovsky Pond, Daniela Fera, and Robert W. Shafer. The biological and clinical significance of emerging SARS-CoV-2 variants. *Nature Reviews Genetics*, 22(12):757–773, December 2021. ISSN 1471-0064. doi: 10.1038/s41576-021-00408-x.

[113] Houriiyah Tegally, Eduan Wilkinson, Marta Giovanetti, Arash Iranzadeh,

Vagner Fonseca, Jennifer Giandhari, Deelan Doolabh, Sureshnee Pillay, Emmanuel James San, Nokukhanya Msomi, Koleka Mlisana, Anne von Gottberg, Sibongile Walaza, Mushal Allam, Arshad Ismail, Thabo Mohale, Allison J. Glass, Susan Engelbrecht, Gert Van Zyl, Wolfgang Preiser, Francesco Petruccione, Alex Sigal, Diana Hardie, Gert Marais, Nei-yuan Hsiao, Stephen Korsman, Mary-Ann Davies, Lynn Tyers, Innocent Mudau, Denis York, Caroline Maslo, Dominique Goedhals, Shareef Abrahams, Oluwakemi Laguda-Akingba, Arghavan Alisoltani-Dehkordi, Adam Godzik, Constantinos Kurt Wibmer, Bryan Trevor Sewell, José Lourenço, Luiz Carlos Junior Alcantara, Sergei L. Kosakovsky Pond, Steven Weaver, Darren Martin, Richard J. Lessells, Jinal N. Bhiman, Carolyn Williamson, and Tulio de Oliveira. Detection of a SARS-CoV-2 variant of concern in South Africa. *Nature*, 592(7854):438–443, April 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03402-9.

[114] Houriiyah Tegally, Monika Moir, Josie Everatt, Marta Giovanetti, Cathrine Scheepers, Eduan Wilkinson, Kathleen Subramoney, Zinhle Makatini, Sikhulile Moyo, Daniel G. Amoako, Cheryl Baxter, Christian L. Althaus, Ugochukwu J. Anyaneji, Dikeledi Kekana, Raquel Viana, Jennifer Giandhari, Richard J. Lessells, Tongai Maponga, Dorcas Maruapula, Wonderful Choga, Mogomotsi Matshaba, Mpaphi B. Mbulawa, Nokukhanya Msomi, Yeshnee Naidoo, Sureshnee Pillay, Tomasz Janusz Sanko, James E. San, Lesley Scott, Lavanya Singh, Nonkululeko A. Magini, Pamela Smith-Lawrence, Wendy Stevens, Graeme Dor, Derek Tshiabuila, Nicole Wolter, Wolfgang Preiser, Florette K. Treurnicht, Marietjie Venter, Georginah Chiloane, Caitlyn McIntyre, Aine O'Toole, Christopher Ruis, Thomas P. Peacock, Cornelius Roemer, Sergei L. Kosakovsky Pond, Carolyn Williamson, Oliver G. Pybus, Jinal N. Bhiman, Allison Glass, Darren P. Martin, Ben Jackson, Andrew Rambaut, Oluwakemi Laguda-Akingba, Simani Gaseitsiwe, Anne von Gottberg, and Tulio de Oliveira. Emergence

of SARS-CoV-2 Omicron lineages BA.4 and BA.5 in South Africa. *Nature Medicine*, 28(9):1785–1790, September 2022. ISSN 1546-170X. doi: 10.1038/s41591-022-01911-2.

[115] Masahiko Ueda, Nobuto Takeuchi, and Kunihiko Kaneko. Stronger selection can slow down evolution driven by recombination on a smooth fitness landscape. *PLOS ONE*, 12(8):e0183120, August 2017. ISSN 1932-6203. doi: 10.1371/journal.pone.0183120.

[116] Jeremy Van Cleve and Daniel B. Weissman. Measuring ruggedness in fitness landscapes. *Proceedings of the National Academy of Sciences*, 112(24):7345–7346, June 2015. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1507916112.

[117] Raquel Viana, Sikhulile Moyo, Daniel G. Amoako, Houriiyah Tegally, Cathrine Scheepers, Christian L. Althaus, Ugochukwu J. Anyaneji, Phillip A. Bester, Maciej F. Boni, Mohammed Chand, Wonderful T. Choga, Rachel Colquhoun, Michaela Davids, Koen Deforche, Deelan Doolabh, Louis du Plessis, Susan Engelbrecht, Josie Everatt, Jennifer Giandhari, Marta Giovanetti, Diana Hardie, Verity Hill, Nei-Yuan Hsiao, Arash Iranzadeh, Arshad Ismail, Charity Joseph, Rageema Joseph, Legodile Koopile, Sergei L. Kosakovsky Pond, Moritz U. G. Kraemer, Lesego Kuate-Lere, Oluwakemi Laguda-Akingba, Onalethatha Lesetedi-Mafoko, Richard J. Lessells, Shahin Lockman, Alexander G. Lucaci, Arisha Maharaj, Boitshoko Mahlangu, Tongai Maponga, Kamela Mahlakwane, Zinhle Makatini, Gert Marais, Dorcas Maruapula, Kereng Masupu, Mogomotsi Matshaba, Simnikiwe Mayaphi, Nokuzola Mbhele, Mpaphi B. Mbulawa, Adriano Mendes, Koleka Mlisana, Anele Mnguni, Thabo Mohale, Monika Moir, Kgomotso Moruisi, Mosepele Mosepele, Gerald Motsatsi, Modisa S. Motswaledi, Thongbotho Mphoyakgosi, Nokukhanya Msomi, Peter N. Mwangi, Yeshnee Naidoo, Noxolo Ntuli, Martin Nyaga, Lucier Olubayo, Sureshnee Pillay, Bot-

shelo Radibe, Yajna Ramphal, Upasana Ramphal, James E. San, Lesley Scott, Roger Shapiro, Lavanya Singh, Pamela Smith-Lawrence, Wendy Stevens, Amy Strydom, Kathleen Subramoney, Naume Tebeila, Derek Tshiabuila, Joseph Tsui, Stephanie van Wyk, Steven Weaver, Constantinos K. Wibmer, Eduan Wilkinson, Nicole Wolter, Alexander E. Zarebski, Boitumelo Zuze, Dominique Goedhals, Wolfgang Preiser, Florette Treurnicht, Marietje Venter, Carolyn Williamson, Oliver G. Pybus, Jinal Bhiman, Allison Glass, Darren P. Martin, Andrew Rambaut, Simani Gaseitsiwe, Anne von Gottberg, and Tulio de Oliveira. Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa. *Nature*, 603(7902):679–686, March 2022. ISSN 1476-4687. doi: 10.1038/s41586-022-04411-y.

[118] Changshuo Wei, Ke-Jia Shan, Weiguang Wang, Shuya Zhang, Qing Huan, and Wenfeng Qian. Evidence for a mouse origin of the SARS-CoV-2 Omicron variant. *Journal of Genetics and Genomics*, 48(12):1111–1121, December 2021. ISSN 1673-8527. doi: 10.1016/j.jgg.2021.12.003.

[119] Daniel Weissman. Stress-Induced Variation Can Cause Average Mutation and Recombination Rates to Be Positively Correlated with Fitness. *The 2019 Conference on Artificial Life*, 26:43–44, July 2014. doi: 10.1162/978-0-262-32621-6-ch008.

[120] Daniel B. Weissman and Nicholas H. Barton. Limits to the Rate of Adaptive Substitution in Sexual Populations. *PLOS Genetics*, 8(6):e1002740, June 2012. ISSN 1553-7404. doi: 10.1371/journal.pgen.1002740.

[121] Daniel B Weissman and Oskar Hallatschek. The Rate of Adaptation in Large Sexual Populations with Linear Chromosomes. *Genetics*, 196(4):1167–1183, April 2014. ISSN 1943-2631. doi: 10.1534/genetics.113.160705.

[122] Daniel B. Weissman, Michael M. Desai, Daniel S. Fisher, and Marcus W. Feldman. The rate at which asexual populations cross fitness valleys. *Theoretical Population Biology*, 75(4):286–300, June 2009. ISSN 00405809. doi: 10.1016/j.tpb.2009.02.006.

[123] Michael C Whitlock. Fixation Probability and Time in Subdivided Populations. *Genetics*, 164(2):767–779, June 2003. ISSN 1943-2631. doi: 10.1093/genetics/164.2.767.

[124] S A J Wilkinson, Alex Richter, Anna Casey, Husam Osman, Jeremy D Mirza, Joanne Stockton, Josh Quick, Liz Ratcliffe, Natalie Sparks, Nicola Cumley, Radoslaw Poplawski, Samuel N Nicholls, Beatrix Kele, Kathryn Harris, Thomas P Peacock, Nicholas J Loman, and The COVID-19 Genomics UK (COG-UK) consortium. Recurrent SARS-CoV-2 mutations in immunodeficient patients. *Virus Evolution*, 8(2):veac050, July 2022. ISSN 2057-1577. doi: 10.1093/ve/veac050.

[125] Maia Kavanagh Williamson, Fergus Hamilton, Stephanie Hutchings, Hannah M. Pymont, Mark Hackett, David Arnold, Nick A. Maskell, Alasdair MacGowan, Mahableshwar Albur, Megan Jenkins, Izak Heys, Francesca Knapper, Mustafa Elsayed, Rachel Milligan, The COVID-19 Genomics UK (COG-UK) Consortium, Peter Muir, Barry Vipond, David A. Matthews, Ed Moran, and Andrew D. Davidson. Chronic SARS-CoV-2 infection and viral evolution in a hypogammaglobulinaemic individual, June 2021.

[126] Wolfram Research, Inc. Mathematica (Version 13.1). https://www.wolfram.com/, 2022.

[127] World Health Organization. WHO Coronavirus (COVID-19) Dashboard. https://covid19.who.int, 2021.

[128] Sewall Wright. *The roles of mutation, inbreeding, crossbreeding and selection*

in evolution, volume 1. Proceedings of the Sixth International Congress of Genetics, 1932.

[129] Joseph T. Wu, Kathy Leung, Mary Bushman, Nishant Kishore, Rene Niehus, Pablo M. de Salazar, Benjamin J. Cowling, Marc Lipsitch, and Gabriel M. Leung. Estimating clinical severity of COVID-19 from the transmission dynamics in Wuhan, China. *Nature Medicine*, 26(4):506–510, April 2020. ISSN 1546-170X. doi: 10.1038/s41591-020-0822-7.

[130] Hui-Ling Yen, Thomas H. C. Sit, Christopher J. Brackman, Shirley S. Y. Chuk, Haogao Gu, Karina W. S. Tam, Pierra Y. T. Law, Gabriel M. Leung, Malik Peiris, Leo L. M. Poon, Samuel M. S. Cheng, Lydia D. J. Chang, Pavithra Krishnan, Daisy Y. M. Ng, Gigi Y. Z. Liu, Mani M. Y. Hui, Sin Ying Ho, Wen Su, Sin Fun Sia, Ka-Tim Choy, Sammi S. Y. Cheuk, Sylvia P. N. Lau, Amy W. Y. Tang, Joe C. T. Koo, and Louise Yung. Transmission of SARS-CoV-2 delta variant (AY.127) from pet hamsters to humans, leading to onward human-to-human transmission: a case study. *The Lancet*, 399(10329):1070–1078, March 2022. ISSN 0140-6736, 1474-547X. doi: 10.1016/S0140-6736(22)00326-9.

[131] Jiří Zahradník, Shir Marciano, Maya Shemesh, Eyal Zoler, Daniel Harari, Jeanne Chiaravalli, Björn Meyer, Yinon Rudich, Chunlin Li, Ira Marton, Orly Dym, Nadav Elad, Mark G. Lewis, Hanne Andersen, Matthew Gagne, Robert A. Seder, Daniel C. Douek, and Gideon Schreiber. SARS-CoV-2 variant prediction and antiviral drug design are enabled by RBD in vitro evolution. *Nature Microbiology*, 6(9):1188–1198, September 2021. ISSN 2058-5276. doi: 10.1038/s41564-021-00954-4.

[132] Lizhou Zhang, Cody B. Jackson, Huihui Mou, Amrita Ojha, Haiyong Peng, Brian D. Quinlan, Erumbi S. Rangarajan, Andi Pan, Abigail Vanderheiden, Mehul S. Suthar, Wenhui Li, Tina Izard, Christoph Rader, Michael Farzan, and

Hyeryun Choe. SARS-CoV-2 spike-protein D614G mutation increases virion spike density and infectivity. *Nature Communications*, 11(1):6013, November 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-19808-4.

[133] Xu-Sheng Zhang and William G. Hill. Predictions of Patterns of Response to Artificial Selection in Lines Derived From Natural Populations. *Genetics*, 169 (1):411–425, January 2005. ISSN 0016-6731, 1943-2631. doi: 10.1534/genetics.104.032573.