Genetic diversity during selective sweeps in non-recombining populations

Sachin Kaushik[1,2,†,*], Kavita Jain[3,†], and Parul Johri[1,2,4,*]

[1]Department of Biology, University of North Carolina at Chapel Hill, NC 27599, USA

[2]Integrative Program for Biological & Genome Sciences, University of North Carolina at Chapel Hill, NC 27599, USA

[3]Theoretical Sciences Unit, Jawaharlal Nehru Centre for Advanced Scientific Research, Bengaluru, 560064, India

[4]Department of Genetics, University of North Carolina at Chapel Hill, NC 27599, USA

[†] These authors contributed equally

[*]Corresponding authors: sackau@ad.unc.edu, pjohri@unc.edu

Abstract

Selective sweeps, resulting from the spread of beneficial, neutral, or deleterious mutations through a population, shape patterns of genetic variation at linked neutral sites. While many theoretical, computational, and statistical advances have been made in understanding the genomic signatures of selective sweeps in recombining populations, substantially less is understood in populations with little/no recombination. We present a mathematical framework based on diffusion theory for obtaining the site frequency spectrum (SFS) at linked neutral sites immediately post and during the fixation of moderately or strongly beneficial mutations. We find that when a single hard sweep occurs, the SFS decays as $1/x$ for low derived allele frequencies (x), similar to the neutral SFS at equilibrium, whereas at higher derived allele frequencies, it follows a $1/x^2$ power law. These power laws are universal in the sense that they are independent of the dominance and inbreeding coefficient, and also characterize the SFS during the sweep. Additionally, we find that the derived allele frequency where the SFS shifts from the $1/x$ to $1/x^2$ law, is inversely proportional to the selection strength: thus under strong selection, the SFS follows the $1/x^2$ dependence for most allele frequencies, resembling a rapidly expanding neutral population. When clonal interference is pervasive, the SFS immediately post-fixation becomes U-shaped and is better explained by the equilibrium SFS of selected sites. Our results will be important in developing statistical methods to infer the timing and strength of recent selective sweeps in asexual populations, genomic regions that lack recombination, and clonally propagating tumor populations.

Keywords: selective sweeps, hitchhiking, site frequency spectrum, diffusion theory, no recombination, dominance, inbreeding, clonal interference

Introduction

Genetic variation is determined by multiple evolutionary processes such as mutation, recombination, population size changes, and selection. While selection against deleterious mutations can affect genome-wide patterns of variation (Charlesworth, 2013), local effects of fixation of beneficial mutations along the genome, also known as selective sweeps, can be used to identify loci involved in recent positive selection. The effect of fixation of a new beneficial mutation at linked sites was first described by Maynard Smith and Haigh (1974) who obtained analytical expressions for the reduction in expected heterozygosity as a function of the distance from the beneficial mutation. Although Maynard Smith and Haigh restricted their study to describing patterns of diversity due to the fixation of a single allele, numerous studies that followed developed the relevant theory to include the effects of recurrent hitchhiking, wherein new beneficial mutations arise continually at the relevant mutation rate and could occur at any position along the chromosome. Both coalescence theory (Kaplan et al., 1989) and diffusion equations (Stephan et al., 1992) were used to obtain the expected number of segregating sites and heterozygosity in the presence of recurrent hitchhiking, respectively. Note that all these analyses assumed that only a single beneficial mutation was on its path to fixation at a time, which is a reasonable assumption in highly recombining species such as *Drosophila melanogaster*.

An insight into the skew in the site frequency spectrum (SFS) of neutral alleles as a result of genetic hitchhiking due to a favorable allele was first investigated using coalescent simulations by Braverman et al. (1995) and subsequently, analytical expressions were obtained by Fay and Wu (2000). It was shown that immediately post-fixation of a strongly beneficial mutation, the SFS became U-shaped, with a larger proportion of high frequency and low frequency variants, while a lower proportion of intermediate frequency variants than expected under strict neutrality. As neutral diversity recovered post-fixation due to new mutations, the skew towards the high-frequency alleles returned to equilibrium levels faster than the skew towards low frequency alleles. Thus while the SFS is U-shaped after a single fixation event, it becomes predominantly left-skewed when recurrent

hitchhiking is considered, as shown by the exact expressions for the neutral SFS using diffusion theory (Kim, 2006).

Although much work has been done to model the effects of selective sweeps in recombining populations, including incorporating background selection as the null model (Kim and Stephan, 2000), accounting for interference between beneficial mutations (Kim and Stephan, 2003), and the development of maximum likelihood methods to detect selective sweeps by studying the spatial pattern of diversity and linkage disequilibria that are characteristics of selective sweeps (Kim and Stephan, 2002; Kim and Nielsen, 2004), not much is known about sweep signatures in non-recombining populations. Arguments based on the coalescence theory suggest that in the absence of recombination, a star genealogy will be produced, with new mutations contributing to the low frequency alleles (Figure 2 of Fay and Wu (2000)). Concordantly, coalescent-based simulations with no recombination indicated a prevalence of only low frequency alleles due to hitchhiking (Simonsen et al., 1995; Fu, 1997) with the shape of the SFS resembling that of a neutral logistically growing population (Fu, 1997). However, we lack a detailed theory that describes the reduction in diversity and the skew in the SFS due to selective sweeps in fully non-recombining populations. We here use a diffusion theory approach to model selective sweeps in a non-recombining population and provide expressions for the SFS at linked neutral sites for arbitrary dominance and probability of inbreeding. Using extensive forward simulations, we test the accuracy of our theory and show how the shape of the SFS depends on the strength of selection, dominance and inbreeding coefficient, and the frequency of beneficial mutations in the population. Our model is most suitable for species or genomic regions with little or no recombination, such as sex chromosomes (Charlesworth, 2017), and clonally reproducing organisms including tumor populations (Durrett, 2015).

Model

We first consider an asexual (i.e., fully non-recombining) population of constant size N with L biallelic loci in which all but one locus are neutral, and the selected locus is haploid. The wildtype

and mutant allele at the neutral locus are denoted, respectively, by a and A, and at the selected locus, by a^* and A^*, with respective fitnesses, 1 and $1+s, s>0$. It is clear that if only one neutral locus is fully linked to the selected locus and the mutant allele at the selected locus eventually fixes, all the individuals will carry the neutral allele that was initially linked to the beneficial mutation thus reducing the neutral diversity to zero. For this reason, in the classic literature on selective sweeps in two-locus models (Maynard Smith and Haigh, 1974), one considers a recombining population. Here, in contrast, we consider an infinite-loci model for asexuals in which, as in Kimura's infinite-sites model (Kimura, 1969, 1971) for fully recombining population, all the $L \to \infty$ neutral loci linked to the selected locus are initially fixed for the wildtype allele and an irreversible mutation occurs at a neutral locus with probability μ per locus per individual, see Fig. 1. Thus, while diversity is continually generated at the neutral loci due to mutations, it can be lost due to random genetic drift or due to linked effects of selection at the selected site, resulting in nontrivial genetic diversity in an asexual population, as illustrated in Fig. 1.

The model described above can be extended in a straightforward manner to include dominance and inbreeding. At the selected locus, we assume that the three genotypes A^*A^*, a^*A^* and a^*a^* have fitnesses $1+s$, $1+hs$ and 1, respectively, where $0<h<1$ is the dominance coefficient. Then if p and $q=1-p$ denote the deterministic frequency of the A^* and a^* allele, respectively, and $0<F<1$ is the inbreeding coefficient, in an inbreeding population, these genotypes occur with frequencies p^2+Fpq, $2pq(1-F)$, and q^2+Fpq. At time $t=0$, a single copy of the beneficial allele, A^* is introduced at the selected site, and then the population evolves following the standard Wright-Fisher dynamics in forward time using the genotypic frequencies and fitnesses mentioned above. These dynamics are implemented in numerical simulations as described below in detail in Methods.

In the following, we numerically study how neutral diversity is affected due to a single hard sweep at the selected locus during and at the end of the selective sweep when all the loci are fully linked; although we are mainly interested in fully non-recombining population, we also perform some simulations for weak recombination. As a measure of diversity, we consider the (unfolded)

SFS $f(x,t)$ of the neutral mutant allele with frequency x at time t in the *subpopulation* carrying the selected allele A^* as shown in Fig. 1. Since we are interested in the dynamics of the asexual subpopulation carrying the selected allele, this means one needs to consider only the asexual neutral population whose size is changing in time. Now assuming that the allele at the neutral loci evolve independently (as in Kimura's infinite-sites model) but in an expanding population with instantaneous size $N(t) = \frac{Np_t}{1+F}$, we formulate a diffusion theory for the SFS which is described in Appendix A and Appendix B, and study it in detail in Appendix C and Appendix D. These results are compared with those obtained from extensive numerical simulations of the asexual and weakly recombining populations, and are discussed in Results. Some important extensions of the model described above, *viz.*, the effect of clonal interference and deleterious sweeps are also discussed.

Methods

Forward time simulations

All forward time simulations are conducted using SLiM version 4.0.1 (Haller and Messer, 2023).

Single sweep

We simulate a Wright-Fisher population consisting of N diploid individuals with no recombination and a mutation rate, $\mu = 10^{-7}$ per site per generation. Mutations are assumed to be semidominant ($h = 0.5$), recessive ($h = 0.2$), or dominant ($h = 0.8$), as specified. The simulated genomic region of length $L = 10^4$ consists of a single selected site where beneficial mutations occur, while the rest of the genome is assumed to be completely neutral. For a single selective sweep, a beneficial mutation with selection coefficient, $s > 0$ is introduced at the selected site. The number of loci simulated is chosen such that there is sufficient neutral diversity at the end of the selective sweep. Since the conditional mean fixation time, t^*_{fix} of the mutant allele at the selected locus scales as $\frac{\ln(Ns)}{s}$ and the mean absorption time of a single neutral mutation, $T_{abs} \sim \ln N$ (Ewens, 2004), neutral diversity

at all the L loci where a mutant appears at rate $N\mu$ per locus vanishes if the total absorption time $\frac{\ln N}{LN\mu} \gg \frac{\ln(Ns)}{s}$. We therefore require $L \gg \frac{s}{N\mu}$ which, for the typical parameters used here, is satisfied.

The site frequency spectrum (SFS) of alleles at neutral loci is recorded immediately post-fixation of the selected allele and the data shown here are averaged over 10^6 fixation events. The neutral SFS is also obtained when the frequency of the beneficial mutation in the selected allele's trajectory (conditional on the eventual fixation of the beneficial mutation) reaches 0.5 and 0.7 for the first time. Note that the SFS obtained during the fixation of the beneficial mutation refers to the SFS of the neutral alleles only from the subpopulation that carries the beneficial mutation, which becomes equivalent to the entire population after the fixation of the beneficial allele. The expected heterozygosity is also calculated for the subpopulation with the beneficial allele at the selected site, at various time points during its path to fixation.

To investigate the effect of inbreeding on the SFS, we extend our simulations to include varying rates of selfing with selfing probabilities σ=0 and 0.9. The inbreeding coefficient F is related to the selfing probability by the relation $F = \frac{\sigma}{2-\sigma}$ (Wright, 1969) so that $F = 0$ and 0.81, respectively. Simulations of deleterious sweeps (Johri et al., 2021; Kaushik and Jain, 2021) are conducted in a similar fashion but limited to weak selection ($|Ns| = 2$) as the probability of fixation of deleterious mutations decreases exponentially with increasing strength of selection.

Recurrent sweeps with clonal interference

To understand the effect of clonal interference on the SFS immediately post-sweep, we consider a more complex genomic architecture that includes multiple beneficial mutations arising at selected sites. Two different types of beneficial mutations are simulated: the first type of beneficial mutations (referred to as the m_1 class) have selective effects that follow an exponential distribution and a high beneficial mutation rate (denoted by μ_{b1}), and are expected to cause interference. The second type of beneficial mutations (m_2 class) have fixed selective effects and very low mutation rates ($\mu_{b2} \to 0$)

thus minimizing any possibility of interference amongst them; the fixation of beneficial mutations in this class are of our interest.

Specifically, in our simulations, the genome architecture consists of a total length of $L = 11,000$ base pairs with the first $L_{sel} = 10^3$ base pairs consisting of sites where either a neutral or a beneficial mutation from either the first or the second class can occur with mutation rates per site/generation denoted by μ_n, μ_{b1}, and, μ_{b2} respectively. The rest of the genome consists of neutral sites where, as before, the mutation rate per generation is μ and at which the SFS is computed. We assume that the total mutation rate at the selected part of the genome is also μ with $\mu_n + \mu_{b1} + \mu_{b2} = \mu$ and $\mu_n = 0.895\mu, \mu_{b1} = 0.1\mu, \mu_{b2} = 0.005\mu$. The selective effect of mutations in class m_1 are chosen from an exponential distribution with mean $Ns_1 = 100$ while that of class m_2 mutations is fixed with scaled selection strength $Ns_2 = 50$. As the mutation rate of m_2 is sufficiently low so that only a single m_2 mutation ("focal mutation") arises in the population until its fixation, the SFS of linked neutral alleles is obtained immediately after the fixation of the focal beneficial mutation following a burn-in period of $10N$ generations to allow for the background (m_1 class) beneficial mutations to reach equilibrium frequencies.

Effect of recombination

The simulation framework is then extended to include recombination among genomes to obtain the SFS immediately post-fixation. Two very different scenarios with recombination are evaluated to test the limitations of our model:

Scenario 1: Our model assumes no and very low rates of recombination such that it is only applicable to regions with recombination rates much lower than the neutral mutation rate and the strength of selection of the beneficial allele. More specifically, the model presented here is appropriate when the force of recombination is much weaker than that of mutation and selection such that $Nr \ll N\mu$ and Ns, where N is the effective population size, μ and r are the mutation and recombination rate per site/generation respectively, and s is the strength of the beneficial mutation.

In order to investigate whether our model would remain accurate within these limits, we perform simulations with an extremely low rate of recombination ($Nr = 10^{-6}$) and a significantly higher one ($Nr = 10^{-3}$) such that $Nr \ll N\mu$ and $Nr \ll Ns$ in the former while Nr is comparable to $N\mu$ in the latter scenario.

Scenario 2: Because our model is a single-locus model, we also explore the effect of linkage under complete neutrality. We simulate a neutral population whose size evolves according to the mean frequency of a beneficial mutation ($s = 0.1$) conditional on its eventual fixation, with a nonzero rate of recombination ($r = 0.01$) and no recombination ($r = 0$). Note that unlike in the theory discussed later where we use the deterministic equation [see (4) below] for the growth of the selected allele, here the frequency of the selected allele in the conditional process was obtained in the numerical simulations.

Obtaining the SFS and nucleotide diversity

At the end of each fixation event, all or a sample of n=10 diploid individuals are sampled from the population. At each polymorphic site and for each individual, we record the frequency of neutral mutations and store this information in a vector v_1. We then generate a histogram of the v_1 vector converting allele counts into allele frequencies while maintaining a constant area under the histogram. This procedure is repeated for 10^6 fixation events, and the averaged histogram is obtained. The resulting averaged data is then binned with a bin size of 100 to obtain the final data which is the SFS shown in the figures. The expected heterozygosity is calculated by averaging $2x(1-x)$ across all genomes, where x represents the frequency of the neutral mutation at each neutral site (also see (9) below). The SFS is obtained both post-fixation and during the selective sweep. While obtaining the SFS during the fixation, i.e., at the intermediate stages, the population consists of two subpopulations: one with the beneficial allele at the selected site and the other with a neutral allele at the selected site. Only the subpopulation that carries the beneficial allele was used to make the SFS.

Python scripts for generating the SFS described above are provided at `https://github.com/JohriLab/sweeps_in_nonrecombining_populations`. Simulation parameters such as genome length, scaled selection coefficient, recombination rates, inbreeding coefficient, population size, and mutation rates of both selected and neutral sites are detailed in figure legends or specified as needed.

Theoretical expectations from infinite-sites model

To obtain an insight into the numerical results, we also studied the SFS in the framework of diffusion theory.

Single sweep

As described in Appendix A, the neutral SFS $f(x,t)$ in a temporally changing population of effective size $N_e(t)$ can be described by the following partial differential equation (Evans et al., 2007):

$$\frac{\partial f(x,t)}{\partial t} = \frac{1}{4N_e(t)}\frac{\partial^2}{\partial x^2}[x(1-x)f(x,t)] \tag{1}$$

where, as detailed below, $N_e(t) = \hat{N}p_t$ due to changing size of the subpopulation with selected allele and $\hat{N} = \frac{N}{1+F}$ is the effective population size due to inbreeding alone (Pollak, 1987; Glémin, 2012). The partial differential equation (1) is subject to boundary conditions (Evans et al., 2007)

$$\lim_{x\to 0} xf(x,t) = \frac{N_e(t)}{\hat{N}}\theta, \quad f(1,t) = \text{finite} \tag{2}$$

where,

$$\theta = 4\hat{N}L\mu \tag{3}$$

and to ensure equivalence between the single-locus diffusion theory and forward time simulations for L loci, the mutation rate in the diffusion model is scaled to $L\mu$. At low allele frequency, (2) models the balance between the loss of polymorphism due to the absorption of the mutant allele by genetic drift and the gain by new mutations. Furthermore, as there are no polymorphic neutral sites when the selected allele arises, the initial frequency $f(x,0) = 0$.

As discussed in Appendix B, for strong selection ($Ns \gg 1$), the effective population size can be approximated by $N_e(t) \approx \hat{N}p_t$ where the mutant allele frequency p_t at the selected locus evolves deterministically according to

$$\frac{dp}{dt} = sp(1-p)[(1-F)(1-2h)p + F + h - Fh] \tag{4}$$

with initial frequency (Maynard Smith, 1976),

$$p_0 = (2\hat{N}p_{fix})^{-1} \tag{5}$$

where, the fixation probability of a single beneficial mutant is given by

$$p_{fix} = 2s(F + h(1-F)) \tag{6}$$

Thus in the framework of diffusion theory, (1)-(6) yield the SFS $f(x,t)$ at a given time slice.

To find the SFS immediately post-fixation, we require the distribution of time to fixation conditional on fixation, which have been derived for randomly mating population (Martin and Lambert, 2015; Kaushik and Jain, 2021; Götsch and Bürger, 2024), and for an inbred population (Kaushik, 2023) using a semi-deterministic theory. But as the expressions for these distributions are approximate and not explicit (barring the case of $h = 1/2, F = 0$), and as shown in Fig. S1, the SFS calculated by averaging it over the conditional distribution of fixation time is not significantly different from that obtained by using the conditional mean fixation time except when selection is weak and/or the selected allele is recessive, here we find the SFS immediately post-fixation at the conditional mean fixation time, t^*_{fix} of the selected allele. For a large randomly mating population, t^*_{fix} is given by (Ewing et al., 2011)

$$t^*_{fix} \approx \frac{\ln(4Ns)}{sh(1-h)} + \frac{\gamma + (2-3h)\ln h + (3h-1)\ln(1-h)}{sh(1-h)} \tag{7}$$

where γ is the Euler constant, whereas for an inbred population, it is obtained from equation (17) of Glémin (2012). The time-dependent SFS is, however, calculated at the time when a given value of deterministic allele frequency p_t is reached and which is found from (4).

The normalized sample SFS, $f_{n,i}(t)$ in a sample of size $n \ll N$ can be obtained from $f(x,t)$ through

$$f_{n,i}(t) = \frac{\hat{f}_{n,i}(t)}{S_n(t)} \ , \ 1 \leq i \leq n-1 \tag{8}$$

where the unnormalized sample SFS, $\hat{f}_{n,i}(t) = \binom{n}{i} \int_0^1 dx x^i (1-x)^{n-i} f(x,t)$ and $S_n(t) = \sum_{i=1}^{n-1} \hat{f}_{n,i}(t)$ is the mean number of segregating loci in the sample of size n at time t. The mean heterozygosity is then obtained from the unnormalized sample SFS as

$$H(t) = \hat{f}_{2,1} = 2 \int_0^1 dx x(1-x) f(x,t) \tag{9}$$

Although we do not provide analytical expressions for these reduced quantities, they are obtained using the numerical solution of (1) in the above definitions.

In Appendix C, the exact solution of (1) is described and an approximate analytical expression for the time-dependent SFS is obtained in Appendix D (see (13) and (14) below). These analytical results as well as the exact numerical solution of (1) is compared with the simulations in the Results section. As it is easier to integrate (1) numerically instead of computing the sum in its exact solution given by (C.2), the numerical solution of the diffusion equation is used throughout the article which is obtained using Mathematica 13.1.1, and is available in a Mathematica file at `https://github.com/JohriLab/sweeps_in_nonrecombining_populations`.

Clonal interference and multiple mutations

When multiple beneficial mutations arise in a non-recombining population, competition between beneficial mutations can result in a decrease in their probability of fixation and is referred to as clonal interference (Gerrish and Lenski, 1998). We assume that a focal beneficial mutation with selective advantage s_2 and dominance coefficient h_2 experiences clonal interference from other background beneficial mutations with selective advantage s_1 and dominance coefficient h_1. Then the reduced fixation probability of the beneficial mutation at the focal site is given by (Gerrish and Lenski, 1998),

$$p_{fix,I} = p_{fix,2} \exp(-P_I) \tag{10}$$

where the expected number of interfering mutations are

$$P_I = 2N\mu_{b1}L \times t^*_{fix,2} \times p_{fix,1} \tag{11}$$

In the above expressions, $p_{fix,i}$ and $t^*_{fix,i}$ for $i = 1, 2$ are obtained by replacing s and h in (6) and (7) by s_i and h_i, respectively. The SFS at the neutral sites is then obtained using $p_{fix,I}$ instead of p_{fix} in (5).

As we discuss later, when the rate of new beneficial mutations is high enough so that $P_I > 1$, the SFS at neutral sites immediately post fixation of the beneficial mutation at the focal site can be explained by the equilibrium SFS in the infinite-sites model under *selection* with selection coefficient s_1 and dominance coefficient h_1 which, in the absence of inbreeding, is given by (Jain and Kaushik, 2022)

$$f_{sel}(x) = \frac{\theta_1 e^{4Ns_1((1-2h_1)x^2+2h_1x)} \int_x^1 e^{-4Ns_1((1-2h_1)y^2+2h_1y)}\, dy}{x(1-x)\int_0^1 e^{-4Ns_1((1-2h_1)y^2+2h_1y)}\, dy} \tag{12}$$

where $\theta_1 = 4N\mu_{b1}L_{sel}$. As the selective effect of mutations in the class m_1 is chosen from an exponential distribution, the above SFS is averaged over this distribution to obtain the final equilibrium SFS, $\langle f_{sel}(x)\rangle$.

Data availability

All scripts to perform simulations, numerical computations, and calculation of statistics have been made available at `https://github.com/JohriLab/sweeps_in_nonrecombining_populations`.

Results

Single hard sweep

The site frequency spectrum (SFS) of linked neutral alleles when a single hard sweep occurs at the selected locus is shown for various parameters in Figs. 2-5, and below, we discuss how the SFS depends on them.

Crossover in SFS

In Fig. 2, the simulation data for the SFS at the completion of the sweep shows that below a crossover allele frequency x_c, the SFS decreases as x^{-1} - this is expected as the loci at low allele frequencies behave in a neutral fashion and the equilibrium SFS in a large neutral population, $f_{neu}(x) = \frac{\theta}{x}$ (Wright, 1938; Kimura, 1968). But for higher frequencies, the SFS decays as x^{-2} which has been observed in exponentially growing populations (Durrett, 2015). Figs. 3-5 show the simulation data for the ratio $\frac{f(x,t)}{f_{neu}(x)}$: for $x \ll x_c$, we find that this ratio is constant with respect to the allele frequency which means that the SFS decays as x^{-1}, at the same rate as the neutral expectation, but for $x \gg x_c$, it decreases with the allele frequency so that the SFS $f(x,t)$ observed during and post-sweep, decreases faster than the neutral expectation. The inset of Fig. 2 shows the normalized sample SFS, which is defined in (8), and reinforces the conclusion that the fraction of polymorphic sites at the end of selective sweep relative to that for neutral equilibrium SFS is negligible for large i, where i is the derived allele count in a sample SFS (Braverman et al., 1995; Fay and Wu, 2000).

To understand these results quantitatively, in Appendix D, we show that when the selective sweep is nearly complete, the time-dependent SFS can be approximated by

$$f(x,p_t) \approx \frac{\theta p_t}{x}\left[\frac{e^{-\frac{w}{4v}} w\left(\mathrm{Ei}\left(\frac{w}{4v}\right) + \ln\left(\frac{4v^2}{w}\right)\right)}{4v^2} + e^{-\frac{w}{4v}}\left(1+\frac{1}{v}\right) - \frac{1}{v}\right] \tag{13}$$

where $w(x) = 16\hat{N}s(F + h(1-F))x, v = \frac{F+h(1-F)}{1-h(1-F)}|\ln(1-p_t)|$, and $\mathrm{Ei}(z)$ is the exponential integral function. On expanding the RHS of (13) when $\frac{w}{4v}$ is smaller and larger than one, we find that

$$f(x,p_t) \approx \begin{cases} \dfrac{\theta p_t}{x} & , 0 < x \ll x_c \quad (14a) \\ \dfrac{\theta p_t p_0}{x^2} & , x_c \ll x < 1 \quad (14b) \end{cases}$$

where the scaled mutation rate θ is given by (3), the frequency p_t of the selected allele at time t is determined by (4), the initial frequency p_0 is given by (5) and the crossover from the x^{-1} to x^{-2}

behavior occurs at an allele frequency,

$$x_c = \frac{F + h(1-F)}{1 - h(1-F)} \times p_0 |\ln(1-p_t)| = \frac{1}{1 - h(1-F)} \times \frac{|\ln(1-p_t)|}{4\hat{N}s} \tag{15}$$

Thus (14b) shows that the decay exponent two of the SFS at large frequencies is *universal* in the sense that it is independent of the dominance and inbreeding coefficients, and holds even before the completion of the sweep. But the crossover allele frequency x_c given by (15) depends on various parameters of the model.

Fig. 2 show a comparison between the exact numerical solution of (1), its approximate solution (13) and the asymptotic results (14) when they are evaluated at the mean fixation time t^*_{fix} given by (7); we find that they agree well with each other and also with those obtained from the numerical simulations for the asexual population.

Strength of selection

Fig. 3 shows the SFS immediately post-fixation for three values of the scaled selection parameter Ns (keeping N fixed); these results for the asexual population are also compared with those obtained from diffusion theory by numerically integrating (1). We find that the proportion of moderate-to-high frequency alleles is consistently smaller in simulations than that obtained from theory, and the quantitative match between the results for the two models is relatively poor for $0.01 < x < 0.1$; presently we do not have an understanding of this discrepancy.

However, the qualitative conclusions drawn from our diffusion theory on the dependence of the SFS on the strength of selection are in agreement with the simulation results. We find that as expected, selection strength weakly affects the SFS at small allele frequencies, but the expected number of polymorphic loci at high allele frequencies is greater for smaller selection coefficients. This observation is consistent with the theoretical prediction in (14b) which shows that the prevalence of high-frequency derived alleles is inversely proportional to the initial frequency, $p_0 \propto (4Ns)^{-1}$. Note that because the population is fully non-recombining and the SFS is from neutral alleles in the *subpopulation*, high-frequency derived alleles comprise only of mutations that occurred in the

early stages of the selective sweep, while low-frequency derived alleles comprise of mutations that occurred during the later stages of the sweep, i.e. closer to the time near fixation. When selection on the selected allele is stronger, the initial frequency of the selected allele is small but the deterministic growth in its frequency occurs soon after it arises and it fixes quickly in the population, therefore there are fewer derived alleles that hitch-hike with it to high frequencies. On the other hand, when selection is weaker, the frequency of the beneficial allele after which it is likely to increase deterministically is much higher (Kaplan et al., 1989; Kim and Stephan, 2002), and as the fixation time of the selected allele is longer, more haplotypes will go to fixation, and there will be more high-frequency derived alleles.

Equation (15) suggests that the crossover allele frequency x_c decreases with increasing Ns. This is because x_c is inversely proportional to the factor Ns, while directly proportional to the logarithmic factor $|\ln(1 - p_{t^*_{fix}})| \sim \ln(Ns)$ since $1 - p_{t^*_{fix}} \sim (Ns)^{-1}$ when selection is strong and initial frequency is small. The decrease in x_c with Ns is consistent with the simulation data shown in Fig. 3 as the ratio $\frac{f(x,t^*_{fix})}{f_{neu}}$ remains close to one for larger range of allele frequencies when selection is weaker. This also implies that under very strong selection, for most of the allele frequency range, the SFS follows a x^{-2} behavior, with the x^{-1}-dependence window being significantly reduced.

In Fig. 3, the selection coefficient s is varied keeping the population size constant. From (14b), for a given selection coefficient, the product θp_0 is independent of N, and thus the tail of the SFS is independent of the population size. Indeed, the simulation data in Fig. S2 for the SFS immediately post-fixation supports this theoretical expectation for allele frequency $x > 0.5$.

Inbreeding

Here, we relax the assumption of random mating and investigate the effect of inbreeding on the SFS immediately post-fixation. For various inbreeding coefficients under codominance, the data in Fig. S3 indicates that for the entire range of derived allele frequency, the expected number of polymorphic loci decreases with increasing inbreeding coefficient, but the ratio $\frac{f(x,t^*_{fix})}{f_{neu}(x)}$ shown in the

inset of Fig. 3 is, in fact, independent of the inbreeding coefficient. These trends are well captured by diffusion theory: for codominance, as $w(x)$ and v defined below (13) are independent of F while the scaled mutation rate θ in (3) is proportional to $(1+F)^{-1}$, it follows from (13) that $f(x,t)$ decreases with increasing F but the ratio $\frac{f(x,t^*_{fix})}{f_{neu}(x)}$ is unaffected by inbreeding. Moreover, (5) suggests that for arbitrary dominance coefficients, p_0 is only mildly different when inbreeding is high. As a result, the neutral SFS at higher derived allele frequencies immediately post-sweep is less impacted by the dominance coefficient for a highly inbred population (data not shown).

Dominance

In Fig. 4, we compare the SFS at the end of the selective sweep when the selected allele is under-, co- and over-dominant for a fixed selection strength and no inbreeding. The simulation data shows that when the derived allele frequencies are small, the SFS is independent of the dominance coefficient; for moderately large x, the expected number of polymorphic loci are higher for higher dominance coefficient but this trend reverses at higher derived allele frequencies. We also note that the number of polymorphic sites observed in the simulated asexual population is consistently smaller than the corresponding results from the diffusion theory, and the discrepancy is most prominent for the underdominant allele and for moderate allele frequencies in the range $10^{-2} - 10^{-1}$. Due to this mismatch between the diffusion theory and the simulations, the dependence of the crossover frequency on the dominance coefficient as predicted by (15) also does not align with the trends observed from simulations.

However, the dominance patterns in the tail of the SFS are captured by (14b) which shows that for $F = 0$, the SFS for $x \gg x_c$ is inversely proportional to the dominance coefficient. This is consistent with the simulation results in Fig. 4 where an increase (decrease) in the proportion of derived high frequency alleles is observed when the beneficial mutation is recessive (dominant) relative to the semi-dominant case. This result can be understood as follows: for strong selection, as the frequency of the selected allele grows as e^{hst} during the early phase, the frequency of the

beneficial mutation will remain quite small for a longer time period when the mutation is recessive ($h < 0.5$) vs when it is dominant ($h > 0.5$). Consequently, derived neutral alleles are more likely to hitchhike to higher frequencies with the recessive beneficial allele compared to the dominant one, resulting in higher expected number of polymorphic loci for $h = 0.2$ relative to that for $h = 0.8$ in Fig. 4. Whereas for low derived allele frequencies, which are largely produced by mutations that occur close to the completion of the sweep, there are more such alleles when the selected allele is dominant vs recessive. This is due to the longer time to fixation of the dominant allele after reaching a high frequency in the population compared to the recessive one near the end of the sweep. However, note that this effect completely disappears in the sample SFS (refer to the inset in Fig. 4), as it is difficult to observe alleles with frequencies as low as 0.005 in a sample of 20 genomes.

Time-dependence of SFS

In the preceding subsections, we discussed the SFS immediately post selective sweep demonstrating that the SFS has a $1/x^2$ dependence on higher derived allele frequencies. We now extend our investigation to the neutral SFS *during* the fixation of a beneficial mutation. When the sweep at the selected locus is incomplete, the population is comprised of two subpopulations: (i) individuals with the beneficial mutation at the selected locus, and (ii) individuals with the wildtype or neutral mutation at the selected locus; here, we focus on the former subpopulation. Fig. 5 shows the simulation data for time-dependent SFS $f(x, t)$ at two time slices where the frequency of the selected allele is far smaller than one. A comparison with the diffusion theory shows that it is not quantitatively accurate when the derived allele frequency is small. Also, as our approximate expression (14) is not valid for a low frequency of the selected allele, here we discuss the results obtained from simulations and numerically integrating the diffusion equation (1).

We find that the SFSs obtained from the subpopulation at different time points differ mostly in the number of low-frequency derived alleles with almost no difference in the number of high-frequency derived alleles, and there are a larger number of low-frequency derived alleles when

the sweep is closer to completion. Intuitively, if one samples a sweep closer to fixation, there are many more genomes sampled in the subpopulation (mimicking a scenario where a population has undergone more generations of exponential growth) and thus there are more new mutations that could have occurred recently; this scenario therefore yields many more singletons or low-frequency derived alleles. On the other hand, most high-frequency derived alleles arose very early on, during the beginning of the sweep, and their frequency is independent of the time of sampling the subpopulation (as $r = 0$). Note that this result is specific to the way we are calculating the SFS only in the subpopulation. For the entire population, the proportion of high-frequency derived alleles would likely not be independent of the time of sampling the sweep. In summary, the contribution of time since the beneficial mutation arose is restricted to low-frequency derived alleles and remains visible in the unnormalized sample SFS although this effect disappears in the normalized sample SFS, as illustrated in Fig. S4.

The dynamics of heterozygosity shown in the inset of Fig. 5 during the course of the sweep in the subpopulation agrees well with those obtained by solving (1) numerically except in the early phases of the sweep. This suggests that at least for some statistics (such as heterozygosity), diffusion theory is a good approximation even when the selective sweep is far from complete.

Recombination

So far, we have discussed the results for a fully non-recombining population where all the neutral loci and the selected locus always remain completely linked, and found that the diffusion theory for unlinked loci matches well with the simulation results for the non-recombining population Our approach is also applicable to regions with recombination rates much lower than the neutral mutation rate and the strength of selection of the beneficial allele. More specifically, the model presented here is appropriate when the force of recombination is much weaker than that of mutation and selection such that $Nr \ll N\mu$ and $Nr \ll Ns$.

These expectations are tested in Fig. 6 which shows that when the recombination rate is much

smaller than the mutation rate and the selection coefficient, the SFS matches the results obtained from the diffusion theory for unlinked loci. However, when the recombination rate is comparable to the mutation rate and the selection coefficient, neutral alleles linked to the selected site can escape fixation by recombining with other backgrounds (individuals with neutral alleles at the selected site), weakening the hitchhiking effect and the frequency spectrum approaches that predicted for recombining populations (Fay and Wu, 2000).

To explore the effect of recombination between neutral loci only, as explained in Methods, we simulate a neutral population whose size is growing and compute the SFS when the population size is $N(t^*_{fix})$, where t^*_{fix} is the conditional mean fixation time given by (7). The inset of Fig. 6 demonstrates that the SFS for a neutral population with logistic growth is unchanged in the presence of recombination, confirming that the linkage between neutral sites does not influence the SFS post-fixation.

Clonal interference and multiple mutations

In the previous sections, we have considered a single hard sweep affecting linked neutral genetic variation. However, when multiple beneficial mutations at selected loci occur during the fixation of a single beneficial mutation (here referred to as the "focal" mutation) (Kim and Stephan, 2003; Chevin et al., 2008), the dynamics are more complex, and interference between selected alleles is highly likely in non-recombining populations. Below we explore three different evolutionary regimes with varying degrees of interference amongst selected sites. Note that all our considered scenarios are characterized by strong selection but the extent of interference is modulated by the rate of new recurrent beneficial mutations entering the population.

In Fig. 7, the simulation data depicts the SFS obtained immediately after the fixation of the focal beneficial mutation (denoted by m_2) which experiences interference from other beneficial mutations (denoted by m_1) which have an exponential distribution of fitness effects. The degree of interference is determined by the expected number of interfering beneficial mutations P_I while the focal mutation

is undergoing fixation and is given by (11). A value of P_I close to or greater than one is likely to result in substantial interference, while a value much smaller than one might indicate almost no interference. In all the cases considered below, the probability of a second m_2 mutation occurring in the population before the fixation of the former m_2 mutation is very low as the expected number of interfering beneficial mutations of type m_2 is 0.07 on using (11), which is quite low.

In the first regime, which we refer to as the "very weak mutation" regime, the rate of new beneficial mutations in m_1 class is kept sufficiently low so that interference between selected alleles is unlikely with $P_I \approx 0.00184$. Consequently, as expected, the simulation data for the neutral SFS post-fixation matches our theory for a single selective sweep quite well (shown in Fig. 7) except for derived allele frequencies close to one. In the second "weak mutation" regime, the rate of new beneficial mutations was moderately high such that $P_I \approx 0.184$. Here, while high frequency derived alleles continue to converge to $1/x^2$ for frequencies in the range $0.7 - 1$, the neutral SFS shows an uptick, thus forming a U-shaped SFS. In the third "strong mutation" regime characterized by a high rate of new beneficial mutations, interference between beneficial mutations was highly likely ($P_I \approx 1.84$) and the effect of selective sweeps at neutral sites is dominated by the presence of multiple beneficial mutations at the selected sites. This intensifies the U-shaped pattern in the SFS, and as shown in Fig. 7, the SFS for high frequency alleles is no longer explained by $1/x^2$-law.

When the rate of beneficial mutations is sufficiently high and segregating beneficial mutations can interfere with each other, there are two events that determine their dynamics in the population: if new beneficial mutations arise on different haplotypes, they will compete with each other, thus reducing each other's probability of fixation. On the other hand, if multiple beneficial mutations arise on the same haplotype, then their probability of fixation might be increased. While the former event is likely with moderate levels of interference, the latter becomes more likely as the rate of new beneficial mutations increases and thus the dynamics become dominated by multiple mutations. Frequent hitchhiking due to multiple beneficial mutations has been shown to generate a U-shaped SFS (Kim, 2006) because beneficial alleles at selected sites reach high frequencies and linked neutral alleles hitchhike along with them leading to more number of expected polymorphic

sites. Concordantly, we observe a more exaggerated U-shape as the rate of beneficial mutations increases in the population. Therefore using the single-locus theory (12) for direct selection and assuming that selected alleles follow an exponential distribution of fitness effects, we calculate the expected SFS at equilibrium, $\langle f_{sel}(x) \rangle$ (shown by magenta line in Fig. 7). Interestingly, we find that the expected allele frequency distribution matches our observed SFS from simulations quite well, suggesting that when the extent of clonal interference is high, the SFS at neutral sites is well explained by those at selected sites.

Finally, although the uptick in the high frequency derived alleles is explained well by (12), we asked if accounting for the probability of fixation with interference using $p_{fix,I}$ given by (10) in the diffusion theory would help explain the shape of the observed SFS at the intermediate frequencies. However it does not capture the qualitative shape of the SFS in the intermediate frequencies as shown in Fig. 7 which is expected because the theory does not account for the effect of multiple mutations as this scenario involves multiple subpopulations with varying numbers of beneficial mutations, making it challenging to incorporate into the existing framework. In summary, while parts of the SFS post-selective sweep when interference levels are moderate can be captured by the theory presented here, our theory cannot explain the shape of the SFS when interference is common.

Deleterious sweep

In this section, we briefly investigate the SFS at linked neutral sites when a single mildly deleterious mutation goes to fixation. In a constant environment, the conditional mean fixation time for a co-dominant allele is the same for a beneficial and deleterious mutation of same magnitude of selection coefficient (Maruyama and Kimura, 1974; Johri et al., 2021; Kaushik and Jain, 2021) although the sojourn time or the equilibrium SFS for a selected allele is not the same for positively or negatively selected mutant allele (Jain and Kaushik, 2022). The simulation data in Fig. S5 suggests that for co-dominance, the neutral SFS post-fixation of a beneficial or deleterious

allele with the same magnitude of selection coefficient is extremely similar. Our results are consistent with the observation that the skew in the SFS is very similar after the fixation of a mildly deleterious and beneficial semidominant mutation with the same magnitude of selection coefficient in recombining populations (Johri et al., 2021). Although the simulation data aligns well with the Maruyama-Kimura symmetry (Maruyama and Kimura, 1974), a full theoretical treatment for a deleterious sweep, including an analytical proof for this observed symmetry of the SFS will be discussed elsewhere.

Discussion

In the simplest setting of a panmictic, randomly mating, neutral population of constant size N in which a wild type allele mutates at rate μ per generation to a mutant allele, the unfolded equilibrium site frequency spectrum (SFS) decays algebraically, as $4N\mu/x$ (Wright, 1938; Kimura, 1968). But this equilibrium result is known to be affected by several factors, such as background selection, selective sweeps, and fluctuations in population size (Fay and Wu, 2000; Przeworski, 2002; Kim and Stephan, 2003; Kim, 2006; Evans et al., 2007; Chevin and Hospital, 2008b; Coop and Ralph, 2012; Kosheleva and Desai, 2013; Charlesworth and Jain, 2014; Živković et al., 2015; Cvijović et al., 2018; Johri et al., 2020; Charlesworth and Jensen, 2021; Jain and Kaushik, 2022; Min et al., 2022; Kaushik, 2023; Balick, 2023). In this study, we focus on the effect of selective sweeps on the SFS at linked neutral sites in species or genomes with little or no recombination. Using extensive numerical simulations and diffusion theory, we study the time-dependent and stationary state SFS, and find that a crossover occurs from the standard neutral theory result of $1/x$ at low allele frequencies to $1/x^2$ at high allele frequencies. The crossover derived allele frequency (15) depends on the population-genetic parameters in the model; in particular, the frequency at which the crossover occurs is inversely proportional to $4Ns$ suggesting that larger the difference in fitness between the favorable and wildtype genotype, the smaller the frequency at which this crossover occurs. This means that if selection is strong, the majority of the SFS post-fixation

will decay as $1/x^2$ whereas the transition between the two power laws will only be observed during the fixation of a weakly selected mutation. In fact, the smaller the sample size, the harder it will be to observe the crossover and thus only $1/x^2$ is likely to be detected in data.

Comparison to a growing tumor or an expanding neutral population

Motivated by the growth of cancerous cells (Durrett, 2015; Tung and Durrett, 2021; Gunnarsson et al., 2021) and bacterial growth in Luria-Delbrück's experiments (Cheek and Antal, 2018), recent studies have focused on understanding how neutral diversity is affected in an exponentially growing population using a branching process and shown a similar $1/x^2$ decay of the SFS. This behavior can be understood using a simple, purely deterministic argument as follows (Williams *et al.*, 2016; Tung and Durrett, 2021): suppose a tumor cell divides at a finite rate λ but there are no death events, and every mutation occurring at a rate μ per cell per generation results in a new type (or allele). Then, starting from a single cell, the size of the population (irrespective of its type) grows exponentially fast as $N(t) = e^{\lambda t}$, and the number of new types created at time t' is the mutation rate multiplied by the population size at that instant, that is, $\mu e^{\lambda t'}$. The population size of a type created at time t' will grow exponentially as $e^{\lambda(t-t')}$ and will reach frequency $x = \frac{e^{\lambda(T-t')}}{e^{\lambda T}} = e^{-\lambda t'}$ at time T. Thus, t' is the latest time at which new types can be created to reach frequency x, and all the types created before t' have a frequency more than x. Therefore, at large times, the number of types of mutants with frequency $\geq x$ is $\mu e^{\lambda t'} = \mu/x$, or the number with frequency x is $1/x^2$.

If now death events are also allowed so that the population can go extinct with a finite probability, a change from $1/x$ to $1/x^2$ behavior occurs at a crossover frequency that increases with the extinction probability. This behavior is analogous to the crossover seen here as the extinction probability in cancer modeling is analogous to the population size of the Wright-Fisher process and one may interpret the low (high) cell viability as small (large) population size. Gunnarsson et al. (2021) suggest that this transition represents the transition from a purely birth-death process

early on during tumor growth to almost a stable population size in the later stages. This would also be consistent with what we observe in our model of logistic growth, where early on there is exponential growth and later stages mimic a nearly constant-sized population. However, in our study we observe a transition between the two power-laws even when the beneficial mutation has not reached fixation, which would be closer to an exponential and not a logistic growth. So it is unclear why we see the transition of power laws in our subpopulation before the event of fixation. Note though that the smaller the frequency of the beneficial mutation at the time of sampling (i.e., the further it is from fixation), the smaller the population frequency x_c at which crossover occurs (see 15), and closer is the SFS to following mostly a $1/x^2$ power law. This is thus consistent with the argument presented in Gunnarsson et al. (2021). Moreover, this behavior is also consistent with the fact that studies that focussed on modeling only an exponential growth under neutrality (Durrett, 2015) do not find the crossover observed here, and the stationary state SFS is found to decay as x^{-2} for all mutant frequencies in these studies.

Note that we and many previous studies assume a lack of recombination in our models. However, the $1/x^2$ power law should remain applicable to an expanding population under strict neutrality even in the presence of recombination. This is because the SFS measures the *mean* number of polymorphic sites with frequency in the range $(x, x + dx)$. In fact, it has been shown that the mean number of segregating sites without and with recombination agree when the scaled mutation rate is small enough. But other quantities such as the probability that the population is monomorphic can be substantially different especially for high mutation rates (Watterson, 1975).

Comparison to selective sweeps in a recombining population

In contrast to the $1/x^2$ power law followed in a non-recombining population, the SFS post-fixation of a beneficial mutation in a recombining population follows a $[x(1-x)]^{-1}$ power law and thus exhibits a U-shape (Kim, 2006). Thus while the SFS post selective sweep in a recombining population has high frequency alleles, the SFS in a non-recombining population is only skewed towards low frequency

alleles, consistent with Braverman et al. (1995). Although here we have mostly focused on the SFS before and at the completion of the sweep, we expect that it would recover over time scales of order N. Assuming that no more beneficial mutations occur on this time scale, the SFS immediately post-fixation will equilibrate to $f_{neu}(x) = \theta/x$, and we wish to find if some range of frequencies take shorter or longer to reach the neutral SFS value (Przeworski, 2002). Our numerical analysis of (1) shown in Fig. S6 suggests that high frequencies take longer than low and intermediate frequencies to recover; this may be expected as the SFS immediately post-fixation at low frequencies is already close to the neutral equilibrium result. This is also because in a non-recombining population there are no haplotypes that escape during the sweep via recombination (which are responsible for the high frequency alleles observed with recombination). Thus the recovery of high frequency alleles in a non-recombining population must occur gradually as new neutral mutations drift to higher frequencies. Note that other constantly operating evolutionary processes such as background selection will also cause the equilibrium SFS to be strongly left-skewed in a non-recombining population (Cvijović et al., 2018), making the pattern generated by a recent selective sweep quite difficult to distinguish from that of background selection.

Further work in understanding the recovery of the SFS as well as obtaining the SFS of the entire population is needed, so that expectations under recurrent hitchhiking due to beneficial mutations in non-recombining populations can be obtained. Such efforts can help predict neutral genomic diversity in species with little/no recombination and perhaps better resolve Lewontin's paradox of variation (Charlesworth and Jensen, 2022).

The feasibility of our approach with pervasive clonal interference

In this study, we find that when the rate of new beneficial mutations is high, i.e., there are multiple beneficial mutations interfering with each other in a nonrecombining population, the equilibrium SFS of neutral linked alleles in the population is U-shaped and likely dominated by recurrent

hitchiking effect caused by the favorable mutations. In this regime, the SFS post-fixation of beneficial mutations is also U-shaped and no longer follows the $1/x^2$ power law. As expected, accounting for the altered fixation probability of beneficial mutations in this regime does not lead to explaining the SFS accurately. This suggests that when clonal interference is pervasive due to a high rate of input of beneficial mutations, the signatures of selective sweeps are not captured by the theory presented here and should be investigated further. While there are a few studies that have modeled the decrease in the probability of fixation of mutations with pervasive clonal interference (Gerrish and Lenski, 1998; Sniegowski and Gerrish, 2010; Jain, 2019), much more work is needed to understand the SFS at linked neutral sites in this regime. We also mention that the U-shaped SFS has also been observed when there is background selection (Cvijović et al., 2018); a comprehensive study incorporating the effect of both positive and negative selection on linked neutral sites is highly desirable.

The usefulness of our approach to perform inference from sequence data

The theoretical development presented here is restricted to modeling the subpopulation of haplotypes that carry the beneficial allele. Thus, while we have a general expression for the distribution of allele frequencies at any time, during the progression and post-fixation of the selective sweep, only the scenario post-fixation is comparable to previous theory. It remains a challenge to obtain the SFS of alleles in the entire population, during the progression of the sweep. An expression of the SFS of the entire population would allow us to accurately model partial sweeps and obtain the expected distribution of allele frequencies due to recurrent hitchhiking in completely non-recombining populations. Further work in this direction is needed.

Despite the above caveat, the theory presented here could be useful for estimating the strength of selection and/or time to the origin of a beneficial allele in non-recombining populations, especially in cases where the beneficial variant is known. This will be particularly applicable to sex chromosomes that lack recombination like the human Y chromosome, the 4th or dot chromosome in *Drosophila*

melanogaster, asexual organisms or viruses that do not recombine, and in tumor populations. As propagation of cancerous cells in a tumor involves clonal reproduction and no recombination, the theory presented here could be used to estimate the strength of selection acting on and/or the time of origin of driver mutations (if the driver mutations are known). In the case of viruses or other pathogenic organisms that reproduce asexually, the drug-resistant allele is known in many cases, e.g., in SARS CoV-2 (Moghadasi et al., 2023) and influenza (Behillil et al., 2020; Graitcer et al., 2011). Thus obtaining the SFS in the subpopulation carrying the drug-resistant allele can be informative of the strength of selection and timing of the selective sweep.

Presently, there are a few caveats of our study that would need to be addressed in order to apply them to natural or experimental populations. Firstly, we assume that the entire population has a constant size. However, it is possible to incorporate a change in population size in this model in the future, allowing for a more general applicability of this model. This will be especially important in a tumor population, which itself is likely to be expanding. Secondly, we would likely require a large sample size to perform inference so that the crossover frequency can be observed in the sample SFS. Thirdly, our theory works well when the beneficial mutation rate is not too high, i.e., when the dynamics of the population are not determined completely by multiple beneficial mutations (Gerrish and Lenski, 1998). The presence of pervasive clonal interference is difficult to determine empirically and might limit the application of the theory presented here to certain scenarios.

Acknowledgements

This research was conducted using computational resources provided by ITS Research Computing at the University of North Carolina at Chapel Hill. This work was funded by the National Institutes of Health grant 1R35GM154969-01 to PJ.

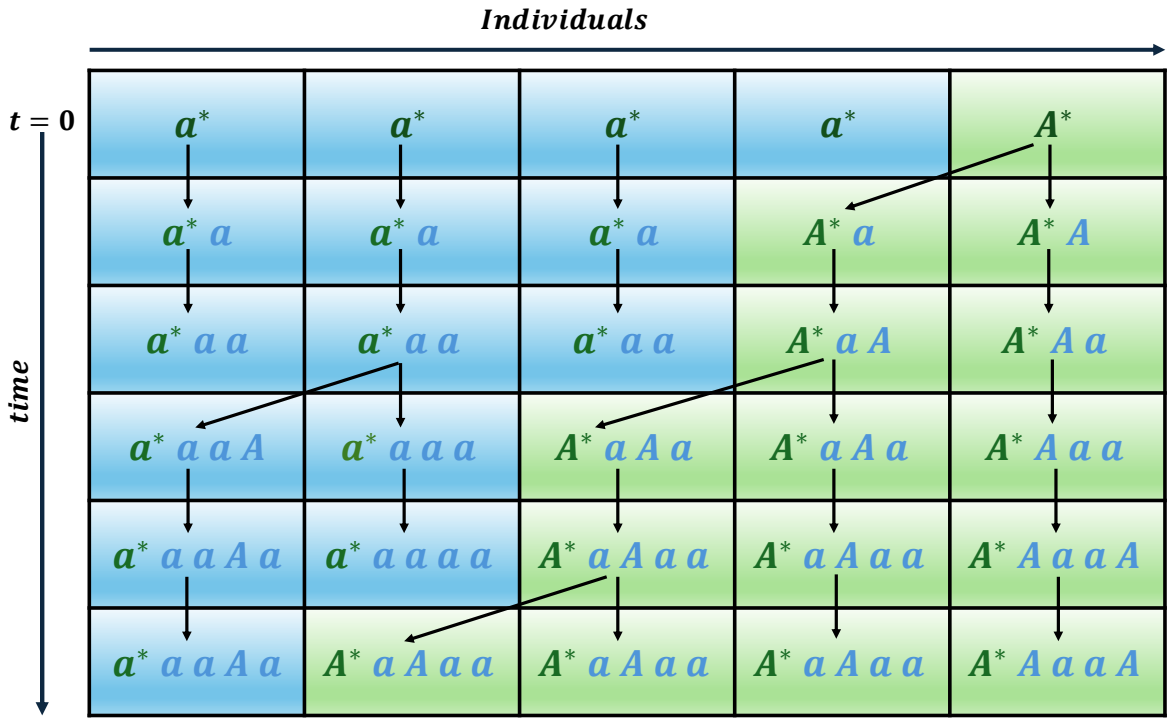

Figure 1: Schematic representation of the model for a haploid asexual population of size $N = 5$. The first locus (where the alleles are marked with a star) represents the selected site, while the remaining loci are neutral. At $t = 0$, all the neutral sites are fixed for the wildtype allele a and a single beneficial mutation A^* arises at the selected site. New neutral mutations, denoted by A, arise in subsequent generations according to the infinite-sites model. Here, the selected site and only those neutral sites that are polymorphic are depicted. The population evolves in time following Wright-Fisher dynamics as indicated by the arrows. At the time of sampling, the SFS at the neutral sites is derived from the subpopulation (depicted with a green background) that contains the beneficial mutation at the selected locus. In the example provided, at $t = 5$, the neutral mutant counts across the four neutral loci are $\{1, 3, 0, 1\}$, corresponding to frequencies of $\{1/4, 3/4, 0/4, 1/4\}$, with 4 being the size of the subpopulation. The SFS is obtained by creating a histogram from these frequencies. The above model is generalized to include dominance, inbreeding and recurrent beneficial mutations, see text for details.

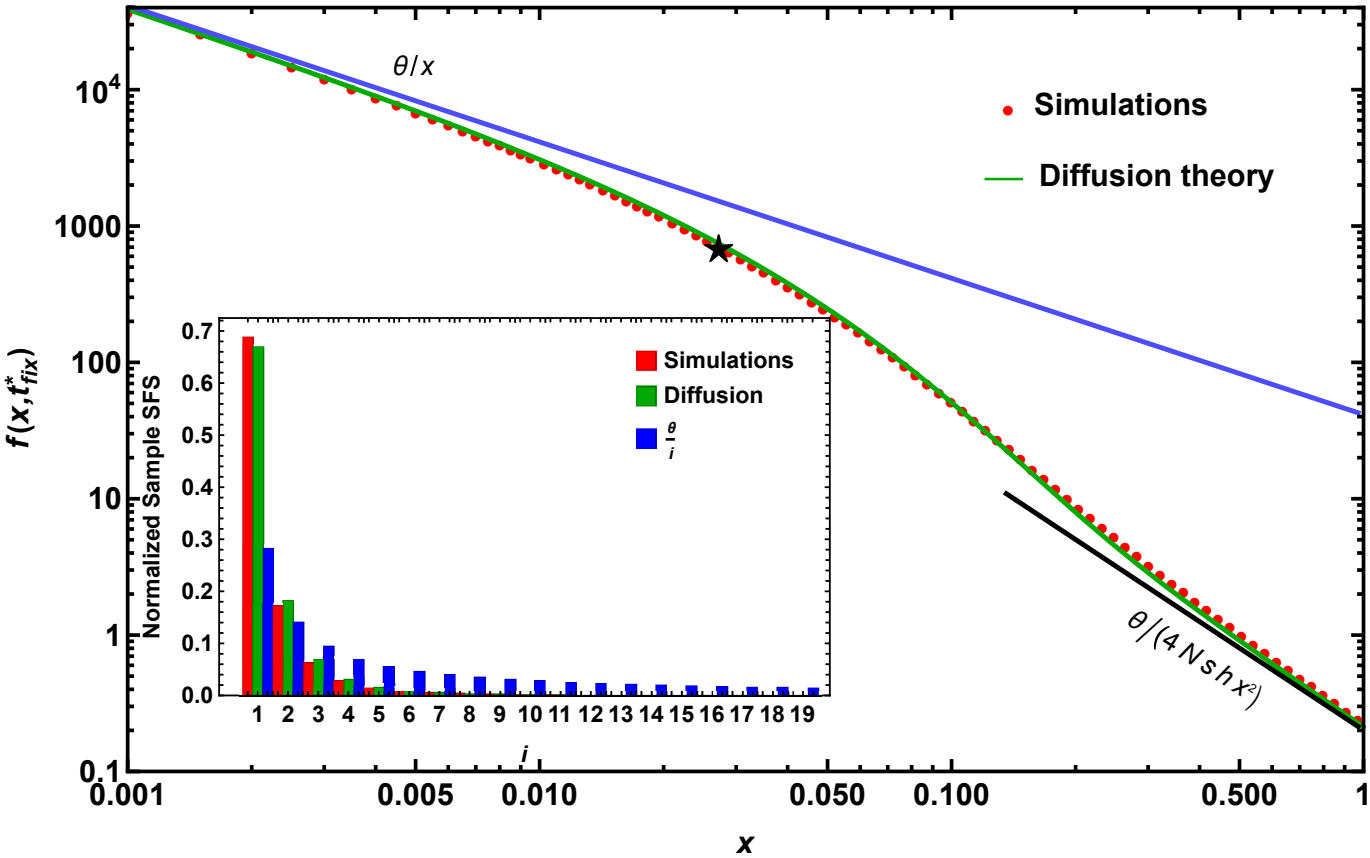

Figure 2: The site frequency spectrum (SFS), i.e., the expected distribution of allele frequencies at neutral sites fully linked to a single new beneficial mutation immediately post-fixation. Here, the red points represent results obtained from forward time simulations, while the green line corresponds to the analytical expression (13) which is obtained using diffusion theory for unlinked neutral loci. The blue and black solid lines, respectively, depict the result (14a) for the SFS at neutral equilibrium and (14b) for higher derived allele frequencies on using $p_t = 0.996, h = 1/2, F = 0$ in these expressions. The crossover frequency x_c is obtained from (15) and is indicated by a black star. The inset shows the normalized SFS from a sample of 20 haploid genomes obtained from simulations and diffusion theory, with the neutral equilibrium sample SFS shown for comparison. Here, the parameters used are $N = 10^3$, $\mu = 10^{-7}$, $L = 10^5, s = 0.1$, $h = 0.5$, $F = 0$.

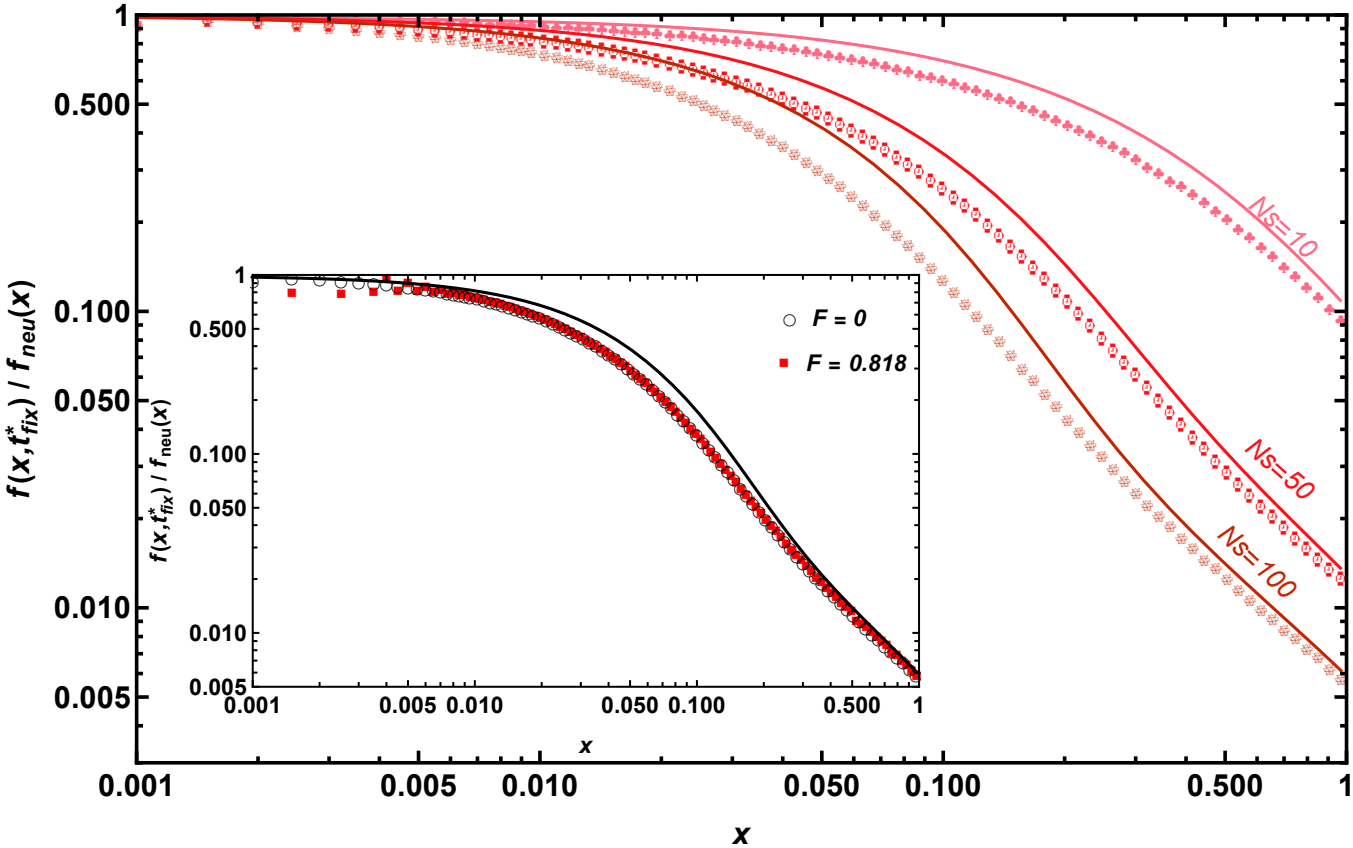

Figure 3: The SFS at neutral sites fully linked to a single new beneficial mutation immediately post-fixation relative to the neutral equilibrium SFS, $f_{neu}(x) = \theta/x$, for varying strengths of selection: $Ns = 10, 50, 100$ (top to bottom). Here, the points are obtained from the forward time simulations and the corresponding solid lines are obtained by numerically solving (1) for the SFS in diffusion theory for unlinked neutral loci. The inset shows the relative SFS for varying inbreeding coefficients. Points are obtained from forward time simulations whereas the black line is obtained using diffusion theory for unlinked neutral loci (1). The parameters used are $N = 10^3$, $\mu = 10^{-7}$, $L = 10^5$, $F = 0$ (main), $s = 0.1$ (inset), and $h = 0.5$.

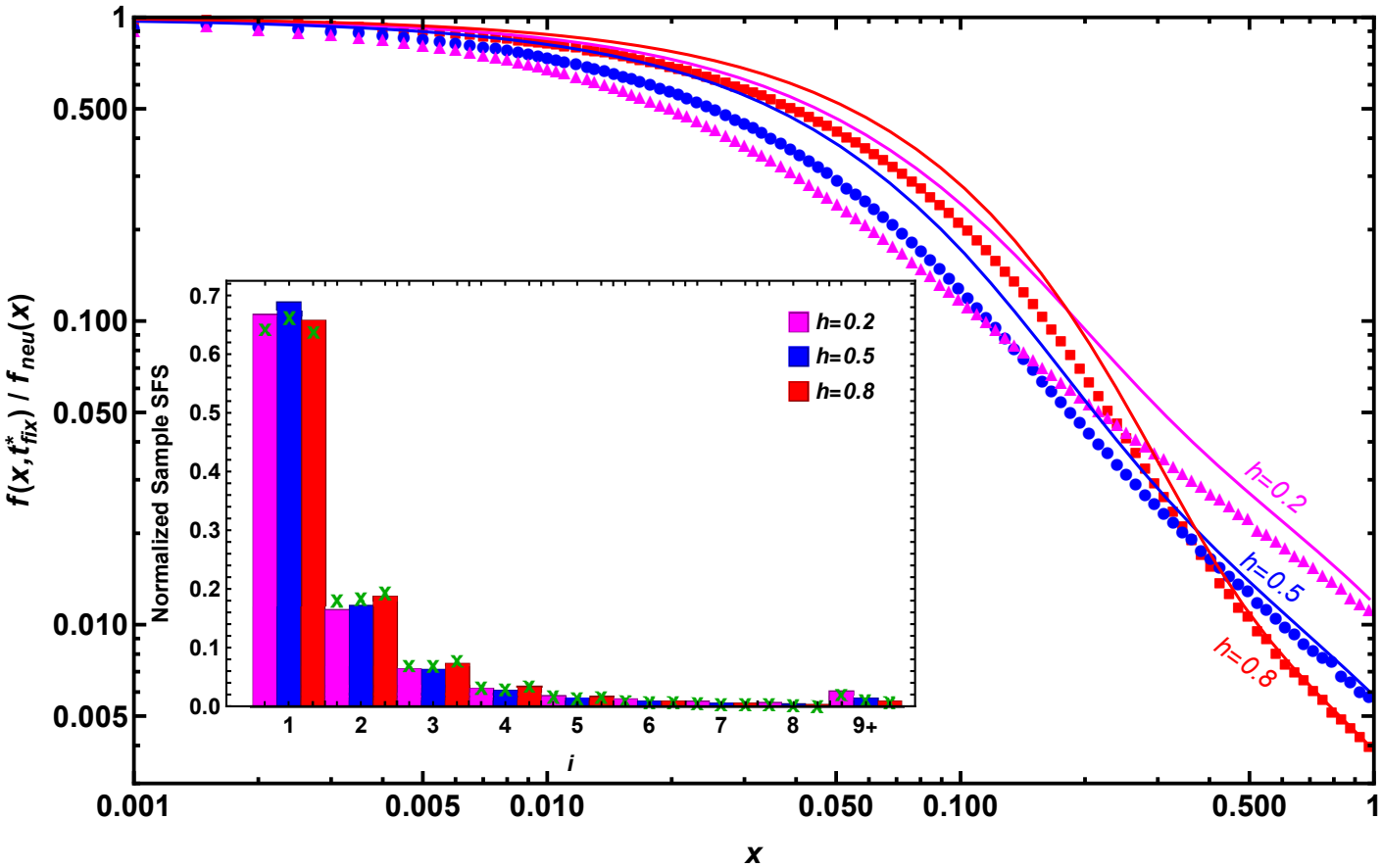

Figure 4: The SFS at neutral sites fully linked to a single new beneficial mutation immediately post-fixation relative to the neutral equilibrium SFS, $f_{neu}(x) = \theta/x$, for various dominance coefficients: $h = 0.2, 0.5, 0.8$. Here, the points are obtained from forward time simulations whereas the corresponding lines of the same color are obtained using diffusion theory for unlinked neutral loci by numerically solving (1). The inset shows the normalized sample SFS for various dominance coefficients where the bars and crosses show the data obtained from simulations and diffusion theory respectively. Note that in the normalized sample SFS shown here, the frequency of derived alleles for counts larger than 8 (i.e., $i \geq 9$) have been pooled together. The other parameters used are $N = 10^3$, $\mu = 10^{-7}$, $s = 0.1$, $F = 0$, and $L = 10^5$.

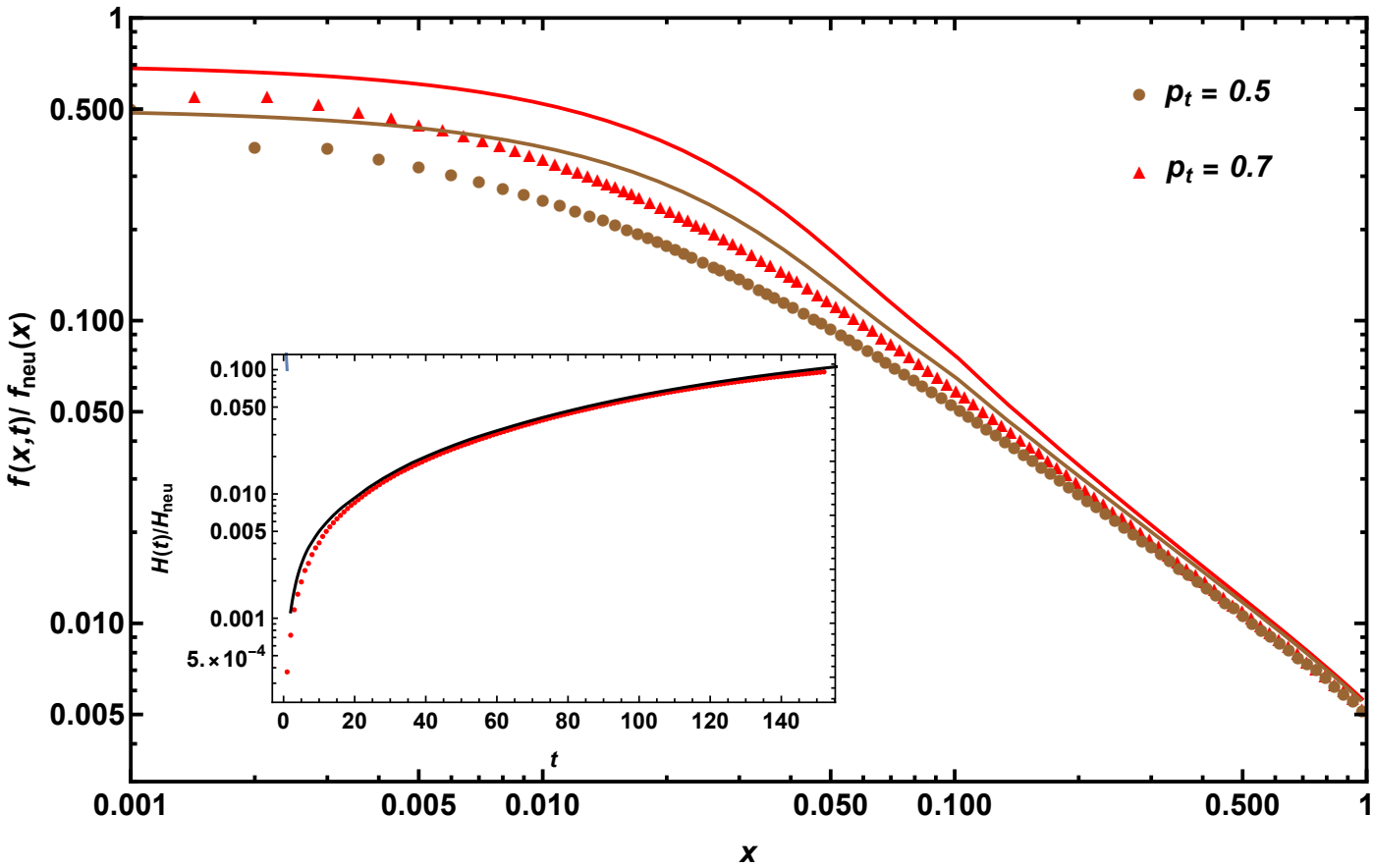

Figure 5: The time-dependent SFS at neutral sites fully linked to a single new beneficial mutation relative to the neutral equilibrium SFS, $f_{neu}(x) = \theta/x$ when the beneficial mutation's frequency p_t is smaller than one. The parameters used for the main figure are $N = 10^3$, $\mu = 10^{-7}$, $s = 0.1$, $F = 0, h = 0.5$, and $L = 10^7$. The inset shows the time evolution of neutral expected heterozygosity ($H(t)$) in the subpopulation carrying the beneficial mutation, relative to the neutral equilibrium heterozygosity ($H_{neu} = 4N\mu L$), during the progression of the beneficial mutation to fixation. The parameters used in the inset are $N = 500$, $\mu = 10^{-4}, h = 0.5$, $F = 0$, and $s = 0.05$. In both the figures, the points are obtained using forward time simulations, and the solid lines are obtained using the diffusion equations for unlinked neutral loci by numerically solving (1).

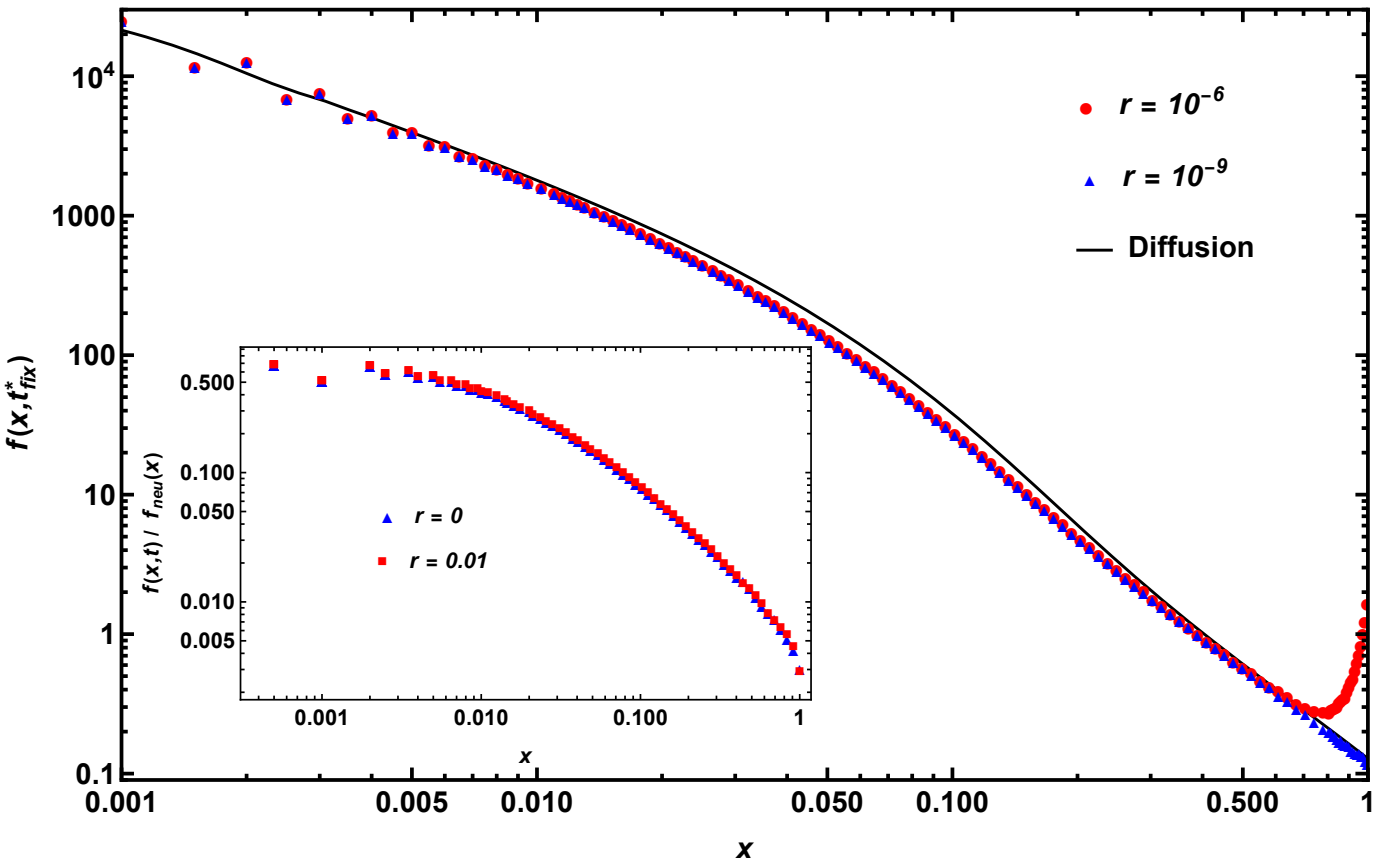

Figure 6: The SFS at neutral sites linked to a single beneficial mutation immediately post-fixation when recombination occurs at scaled rate $Nr = 10^{-3}$ (circles) and 10^{-6} (triangles). Points are obtained from forward time simulations while the solid line shows the numerical solution of (1) for the SFS from diffusion theory for unlinked neutral loci. Inset: The inset shows the time-dependent SFS relative to the neutral equilibrium SFS, $f_{neu}(x) = \theta/x$ in a strictly neutral expanding population with and without recombination; the simulation data are recorded when the deterministic allele frequency $p_t = 0.35$. In both figures, the parameters used are $N = 10^3$, $\mu = 10^{-6}$, $s = 0.1$, $L = 10^4$, $F = 0.818$ (main), $F = 0$ (inset), and $h = 0.5$.

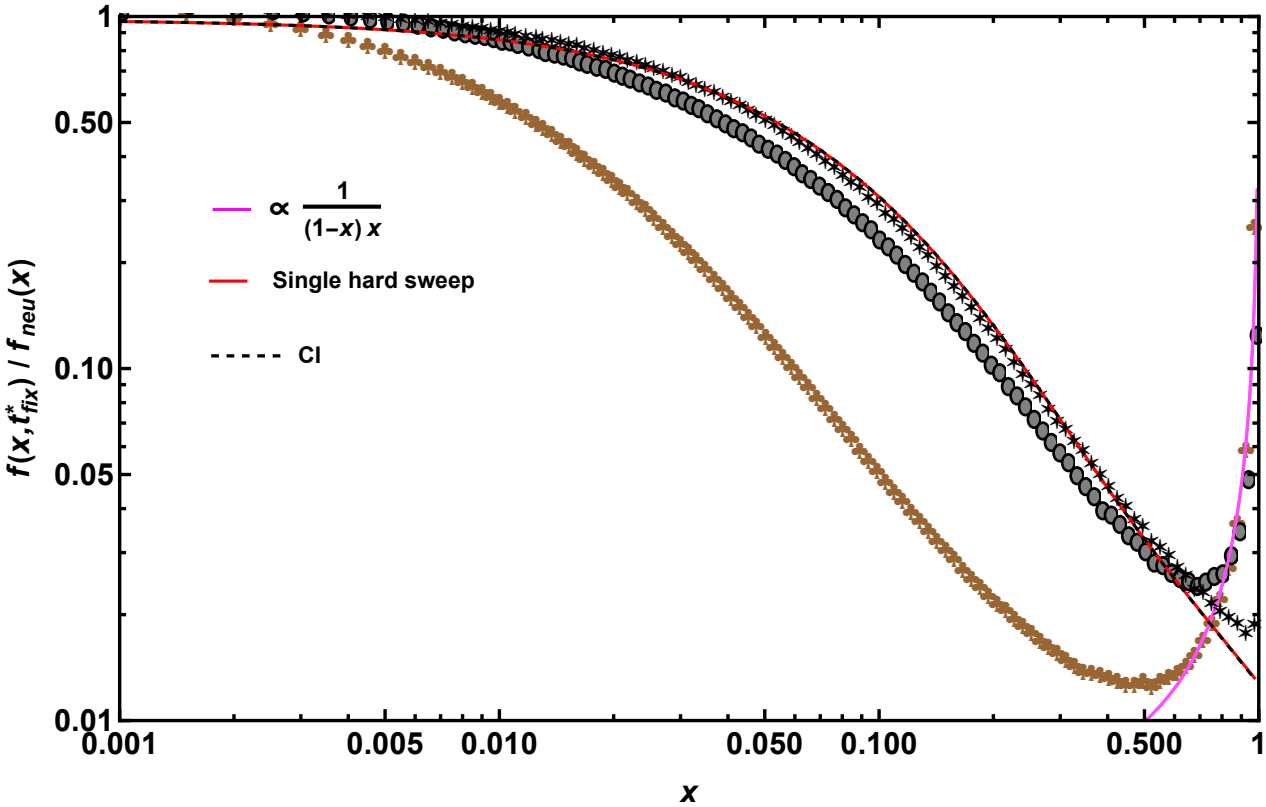

Figure 7: The SFS at linked neutral sites immediately post-fixation of a beneficial allele with $Ns_2 = 50$ experiencing clonal interference due to strongly beneficial mutations whose distribution is assumed to be exponential with mean $Ns_1 = 100$. The SFS is shown relative to the neutral equilibrium SFS when the expected number of interfering mutations obtained using (11) is low ($P_I = 0.00184, *$), moderate ($P_I = 0.184$, ●) and high ($P_I = 1.84$, ♣). The points are obtained from forward time simulations, while the red and black dashed lines, respectively, are obtained by numerically solving (1) assuming no clonal interference and with clonal interference (that is, using (10) instead of (6) in the diffusion theory). The magenta line is obtained by integrating (12) for the equilibrium SFS for selected sites over the exponential distribution of fitness effects of new beneficial mutations. The parameters are $N = 10^3$, $L = 11,000$, $\mu = 10^{-6}$, $F = 0$ and $h_1 = h_2 = 0.5$.

Appendix A Diffusion theory for changing population size

We first explain how the size $N(t)$ of the subpopulation carrying the selected allele A^* evolves (see Fig. 1). At the selected locus, we denote the three genotypes as A^*A^*, A^*a^* and a^*a^* with fitnesses $1+s, 1+hs$ and 1, respectively, where $0 < h < 1$ is the dominance coefficient and $0 < s < 1$ is the selection coefficient. In an infinitely large population, if the frequency of the selected mutant allele is p and that of wild-type allele is $q = 1-p$, under Hardy-Weinberg equilibrium (HWE), the frequencies of the three genotypes are given by $p^2, 2pq, q^2$, respectively. We also assume that inbreeding occurs with probability $0 \leq F < 1$ so that the contribution to either homozygote's frequency due to the union of heterozygotes is Fpq and therefore, with inbreeding, the genotypic frequency vector in an infinitely large population is given by (Charlesworth and Charlesworth, 2010)

$$\vec{y} = \{p^2 + Fpq, pq(1-F), pq(1-F), q^2 + Fpq\} \tag{A.1}$$

As expected, the frequencies are not in HWE due to non-random mating. The deterministic allele frequency $y(A^*, t) = y(A^*A^*, t) + y(A^*a^*, t)$ of A^* allele is then updated according to

$$y(A^*, t+1) = \frac{(1+s)(p^2 + Fpq) + (1+hs)pq(1-F)}{1 + sp[p + F(1-2h)(1-p) + 2h(1-p)]} \tag{A.2}$$

Then in the Wright-Fisher process, the number of mutants, $N(t+1)$ in generation $t+1$ in a population of (census) size N are obtained from a binomial distribution with mean $\hat{N}k(t)$,

$$N(t+1) \sim B\left(\hat{N}, k(t)\right) \tag{A.3}$$

where $\hat{N} = \frac{N}{1+F}$ (Pollak, 1987; Glémin, 2012) to take care of the reduction in effective population size due to inbreeding, and the probability $k(t)$ is given by the right-hand side (RHS) of (A.2) as the mean of the binomial distribution (A.3) is given by $\hat{N}k$. Note that while the population size $N(t)$ is a binomially distributed random variable, as we are interested in the SFS when the selected allele sweeps, we require the distribution F_c of the allele frequency *conditioned* on fixation of the selected allele, an approximate theory for which has been developed for strong positive selection (Martin and Lambert, 2015; Götsch and Bürger, 2024).

We now consider the neutral population with alleles a and A whose number evolves according to a Wright-Fisher process with population size $N(t)$ (see Fig. 1). If $n_a(t)$ and $n_A(t)$, respectively, denote the number of neutral wildtype and mutant individuals in generation t with $N(t) = n_a(t)+n_A(t)$, the number of mutant individuals in the next generation is chosen according to a binomial distribution,

$$n_A(t+1) \sim B\left(N(t+1), \frac{n_A(t)}{N(t)}\right) \tag{A.4}$$

That is, the neutral mutants in the previous generation are sampled $N(t+1)$ times with equal probability.

To develop a diffusion theory for the neutral SFS in the changing population size for the model described above, we need to find the change in mean and variance of the neutral allele frequency, $x(t) = \frac{n_A(t)}{N(t)}$ in one generation where $n_A(t)$ and $N(t)$ are distributed according to (A.4) and F_c, respectively. If $\overline{...}$ and $\langle ... \rangle_{F_c}$ denote the average with respect to these respective distributions, it immediately follows from (A.4) that the mean neutral allele frequency at time $t+1$ is given by

$$\left\langle \overline{x(t+1)} \right\rangle_{F_c} = \left\langle \frac{1}{N(t+1)} \sum_{n_A=0}^{N(t+1)} n_A(t+1) P(n_A) \right\rangle_{F_c} = \left\langle \frac{n_A(t)}{N(t)} \right\rangle_{F_c} \tag{A.5}$$

so that there is no change in the mean allele frequency. Similarly, the variance of the change in the allele frequency is given by

$$\left\langle \overline{[x(t+1) - x(t)]^2} \right\rangle_{F_c} = \left\langle \frac{1}{N(t+1)} \right\rangle_{F_c} x(t)[1-x(t)] \approx \left\langle \frac{1}{N(t)} \right\rangle_{F_c} x(t)[1-x(t)] \tag{A.6}$$

on using that $\langle \overline{x(t+1)} \rangle = x(t)$. Thus, we find that the distribution of the mutant allele at the neutral locus in a changing population obeys the following forward Fokker-Planck equation,

$$\frac{\partial P(x,t)}{\partial t} = \frac{1}{4N_e(t)} \frac{\partial^2 [x(1-x)P(x,t)]}{\partial x^2} \tag{A.7}$$

where, the variance effective population size

$$\frac{1}{N_e(t)} = \left\langle \frac{1}{N(t)} \right\rangle_{F_c} \tag{A.8}$$

is the harmonic mean of the size of the subpopulation carrying selected allele conditioned on fixation.

On using (A.8) in (1) for the SFS, our numerics (not shown) suggest that this approximation performs worse than if we assume that the size of the population varies deterministically. For this reason, in this article, we assume that the size of the subpopulation changes according to (B.1) as described in Appendix B.

Appendix B Deterministic dynamics of population size

If selection is strong ($Ns \gg 1$) and the time at which the SFS is measured is not too small, we expect that the stochasticity in the population size can be ignored. Furthermore, the stochastic events leading to the eventual fixation of one copy of selected allele can be taken into account if the initial frequency of the selected allele is $p_0 = (2\hat{N}p_{fix})^{-1}$ (Maynard Smith, 1976) where p_{fix} denotes the eventual fixation probability of a single mutant.

To this end, we first note that for small selection coefficient ($s \to 0$), (A.2) shows that $k \approx p$ for any F, h, and in continuous time, it yields

$$\frac{dp}{dt} = sp(1-p)(\phi_1 p + \phi_2) \tag{B.1}$$

which gives

$$st = \frac{(\phi_1+\phi_2)\ln(\frac{p}{p_0}) - \phi_1 \ln(\frac{\phi_1 p+\phi_2}{\phi_1 p_0+\phi_2}) - \phi_2 \ln(\frac{1-p}{1-p_0})}{\phi_2(\phi_1+\phi_2)} \tag{B.2}$$

where,

$$\phi_1 = (1-F)(1-2h) \tag{B.3}$$

$$\phi_2 = F + h(1-F) \tag{B.4}$$

Thus, ignoring the fluctuations in the subpopulation size, we write

$$N_e(t) \approx \hat{N}p_t \tag{B.5}$$

where the frequency p_t is determined from (B.2) with the initial condition, $p_0 = (2\hat{N}p_{fix})^{-1}$ and for large Ns, the backward equation counterpart of (A.7) gives $p_{fix} = 2s(F + h(1-F))$ (Ewens, 2004).

Appendix C Site frequency spectrum for changing population size

If the allele frequency at a neutral locus in a population of changing size evolves according to (A.7) and is under strong selection so that $N_e(t) \approx \hat{N}p_t$ with p being the deterministically evolving allele frequency, the neutral SFS $f(x,t)$ obeys (1) which can be solved exactly to give (Živković et al., 2015; Jain and Kaushik, 2022)

$$f(x,t) = \frac{\theta p_t}{x} + \theta j(x,t) \tag{C.1}$$

where,

$$j(x,t) = \sum_{m=1}^{\infty} a_m(t) \frac{(m+1)(2m+1)}{m} P_{m-1}^{(1,1)}(1-2x) \tag{C.2}$$

In the above expression, $P_n^{(a,b)}(x)$ is the Jacobi polynomial and the time-dependent coefficients,

$$a_m(t) = -\frac{1}{m+1} e^{-\frac{m(m+1)}{4\hat{N}} \int_0^t \frac{dt'}{p_{t'}}} \int_0^t dt' e^{\frac{m(m+1)}{4\hat{N}} \int_0^{t'} \frac{dz}{p_z}} \frac{dp_{t'}}{dt'} + e^{-\frac{m(m+1)}{4\hat{N}} \int_0^t \frac{dt'}{p_{t'}}} a_m(0) \tag{C.3}$$

with

$$a_m(0) = \int_0^1 dx \left(\frac{x(1-x)f(x,0) - \theta(1-x)p_0}{\theta} \right) P_{m-1}^{(1,1)}(1-2x) \tag{C.4}$$

For the initial condition $f(x,0) = 0$,

$$a_m(0) = -\frac{p_0}{m+1} \tag{C.5}$$

Appendix D SFS during and after single selective sweep

As described in Appendix C, the exact expression for the neutral SFS in the infinites-sites model with changing population size $\hat{N}p_t$ is given by (C.1). To our knowledge, the sum $j(x,t)$ has not been previously analyzed for the SFS, and below we find an approximate expression for it when $p_0 \to 0$ and $p_t \to 1$, that is, when the mutant is initially rare but has spread in the expanding population.

Due to (B.2), instead of time, it is convenient to express the SFS in terms of the allele frequency p_t. Then

$$a_m(p_t) = -\frac{1}{m+1}e^{-b_m D(p_t)} I_m(p_t) - e^{-b_m(D(p_t)-D(p_0))}\frac{p_0}{m+1} \tag{D.1}$$

where,

$$b_m = \frac{m(m+1)}{4\hat{N}s} \tag{D.2}$$

$$I_m(p_t) = \int_{p_0}^{p_t} dp e^{b_m D(p)} \tag{D.3}$$

and due to (B.1),

$$D(p) = s\int^t \frac{dt'}{p_{t'}} = \int^p \frac{dp'}{p'p'(1-p')(\phi_1 p' + \phi_2')} \tag{D.4}$$

$$= \frac{\frac{\phi_1^2 \ln(\phi_1 p + \phi_2)}{\phi_1+\phi_2} - (\phi_1 - \phi_2)\ln p - \frac{\phi_2}{p}}{\phi_2^2} - \frac{\ln(1-p)}{\phi_1+\phi_2} \tag{D.5}$$

Since the population is expanding, $dD/dp = (pdp/dt)^{-1} > 0$, and the function D increases from $D(p) \overset{p\to 0}{\approx} -\frac{1}{\phi_2 p}$ towards $D(p) \overset{p\to 1}{\approx} -\frac{\ln(1-p)}{\phi_1+\phi_2}$ with increasing p. On expanding $D(p)$ for small p and carrying out the integral in $I_m(p_t)$, we obtain

$$I_m(p_t) \approx \int_{p_0}^{p_t} dp e^{-a/p} = a\Gamma\left(0, \frac{a}{p_0}\right) - a\Gamma\left(0, \frac{a}{p_t}\right) + p_t e^{-\frac{a}{p_t}} - p_0 e^{-\frac{a}{p_0}} \tag{D.6}$$

$$\overset{p_0\to 0}{\approx} -a\Gamma\left(0, \frac{a}{p_t}\right) + p_t e^{-\frac{a}{p_t}} \tag{D.7}$$

where $a = \frac{m^2}{4\hat{N}s\phi_2}$. Also, for $p_t \to 1$, as $e^{-b_m D(p_t)} \approx e^{-\frac{m^2}{4\hat{N}s}\frac{|\ln(1-p_t)|}{\phi_1+\phi_2}}$ decays rapidly with m, the product $e^{-b_m D(p_t)} I_m(p_t)$ receives contribution from $I_m(p_t)$ for small a/p_t, and we can therefore write

$$I_m(p_t) \overset{a/p_t \ll 1}{\approx} p_t \left(1 + \frac{a}{p_t} \ln \frac{a}{p_t}\right) \tag{D.8}$$

Furthermore, for large m, on approximating the Pochammer symbol $(b)_n = b(b+1)...(b+n-1) \overset{b \gg 1}{\approx} b^n$, we can write

$$\frac{P_{m-1}^{(1,1)}(1-2x)}{m} = \sum_{n=0}^{m-1} \frac{(1-m)_n (2+m)_n}{(2)_n} \frac{x^n}{n!} \tag{D.9}$$

$$\overset{m \gg 1}{\approx} \sum_{n=0}^{\infty} \frac{(-1)^n (m^2 x)^n}{n! \Gamma(n+2)} \tag{D.10}$$

$$= \frac{J_1\left(2m\sqrt{x}\right)}{m\sqrt{x}} \tag{D.11}$$

where $J_n(x)$ is Bessel function of the first kind.

Using the approximate expressions described above and approximating the sum over m on the RHS of (C.2) by an integral, we obtain

$$j(x, p_t) \approx p_t \int_0^\infty dm e^{-\frac{m^2}{4\hat{N}s}\frac{|\ln(1-p_t)|}{\phi_1+\phi_2}} [1 + \frac{m^2}{4\hat{N}s\phi_2 p_t} \ln \frac{m^2}{4\hat{N}s\phi_2 p_t}] \times 2m \times \frac{J_1\left(2m\sqrt{x}\right)}{m\sqrt{x}} \tag{D.12}$$

$$= \frac{p_t}{x} \left[\frac{e^{-\frac{w}{4v}} \left(w \left(\mathrm{Ei}\left(\frac{w}{4v}\right) + \ln\left(\frac{4v^2}{w}\right)\right) + 4v \right)}{4v^2} + e^{-\frac{w}{4v}} - \frac{1}{v} - 1 \right] \tag{D.13}$$

where $v = \frac{\phi_2}{\phi_1+\phi_2}|\ln(1-p_t)|$, $w = 16\hat{N}sx\phi_2$, and $\mathrm{Ei}(z) = -P[\int_{-z}^{\infty} dt \frac{e^{-t}}{t}]$ is the exponential integral function and P denotes the principal value of the integral. The constants ϕ_1 and ϕ_2 are given in (B.3) and (B.4), respectively.

References

Balick, D. J., 2023. A field theoretic approach to non-equilibrium population genetics in the strong selection regime. bioRxiv. doi: https://doi.org/10.1101/2023.01.16.524324

Behillil, S., *et al.*, 2020. Oseltamivir resistance in severe influenza A(H1N1)pdm09 pneumonia and acute respiratory distress syndrome: A French multicenter observational cohort study. Clin Infect Dis. **71(4):** 1089–1091.

Braverman, J. M., R. R. Hudson, N. L. Kaplan, C. H. Langley, and W. Stephan, 1995. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics. **140:** 783–796.

Charlesworth, B. and D. Charlesworth, 2010. Elements of Evolutionary Genetics. Roberts and Company Publishers, Greenwood Village, Colorado, USA, 734 p.

Charlesworth, B., 2013. Background selection 20 years on: The Wilhelmine E. Key 2012 invitational lecture. J. Hered. **104(2):** 161–171.

Charlesworth, B. and K. Jain, 2014. Purifying selection, drift, and reversible mutation with arbitrarily high mutation rates. Genetics. **198(4):** 1587–1602.

Charlesworth, D., 2017. Evolution of recombination rates between sex chromosomes. Phil Trans R Soc B. **372:** 20160456.

Charlesworth, B. and J. D. Jensen, 2021. Effects of selection at linked sites on patterns of genetic variability. Annu Rev Ecol Evol Syst. **52:** 177-197.

Charlesworth, B. and J. D. Jensen, 2022. How can we resolve Lewontin's paradox? Gen Biol and Evol. **14:** evac096.

Cheek, D. and T. Antal, 2018. Mutation frequencies in a birth–death branching process. Ann Appl Probab. **28(6):** 3922–3947.

Chevin, L. M., S. Billiard, and F. Hospital, 2008. Hitchhiking both ways: effect of two interfering selective sweeps on linked neutral variation. Genetics. **180(1):** 301–316.

Chevin, L. M. and F. Hospital, 2008b. Selective sweep at a quantitative trait locus in the presence of background genetic variation. Genetics. **180(3):** 1645–1660.

Coop, G. and P. Ralph, 2012. Patterns of neutral diversity under general models of selective sweeps. Genetics. **192:** 205–224.

Cvijović, I., B. H. Good and M. M. Desai, 2018. The effect of strong purifying selection on genetic diversity. Genetics. **209:** 1235–1278.

Durrett, R., 2015. Branching process models of cancer. Branching process models of cancer. Springer International Publishing, Cham, pp. 1–63.

Evans, S. N., Y. Shvets and M. Slatkin, 2007. Non-equilibrium theory of the allele frequency spectrum. Theor Popul Biol. **71:** 109–119.

Ewens, W., 2004. Mathematical population genetics. Springer, Berlin.

Ewing, G., J. Hermisson, P. Pfaffelhuber, and J. Rudolf, 2011. Selective sweeps for recessive alleles and for other modes of dominance. J Math Biol. **63:** 399–431.

Fay, J. C., and C. I. Wu, 2000. Hitchhiking under positive Darwinian selection. Genetics. **155(3):** 1405–1413.

Fu, Y. X., 1997. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics. **147:** 915–925.

Gerrish, P. J., and R. E. Lenski, 1998. The fate of competing beneficial mutations in an asexual population. Genetica. **102/103:** 127–144.

Glémin, S., 2012. Extinction and fixation times with dominance and inbreeding. Theor Pop Biol. **81:** 310–316.

Götsch, H. and R. Bürger, 2024. Polygenic dynamics underlying the response of quantitative traits to directional selection. Theor Popul Biol. **158:** 21–59.

Graitcer, S. B., *et al.*, 2011. Characteristics of patients with oseltamivir-resistant pandemic (H1N1) 2009, United States. Emerg Infect Dis. **17(2):** 255–257.

Gunnarsson, E. B., K. Leder and J. Foo, 2021. Exact site frequency spectra of neutrally evolving tumors: A transition between power laws reveals a signature of cell viability. Theor Popul Biol **142:** 67–90.

Haller , B. C., and P. W. Messer, 2023. SLiM 4: Multispecies eco-evolutionary modeling. Am Nat. **201:** E127–E139.

Jain, K., 2019. Interference effects of deleterious and beneficial mutations in large asexual populations. Genetics. **211(4):** 1357–1369.

Jain, K. and S. Kaushik, 2022. Joint effect of changing selection and demography on the site frequency spectrum. Theor Popul Biol. **146:** 46–60.

Johri, P., B. Charlesworth, and J. D. Jensen, 2020. Toward an evolutionarily appropriate null model: jointly inferring demography and purifying selection. Genetics. **215(1):** 173–192.

Johri, P., B. Charlesworth, E. K. Howell, M. Lynch, and J. D. Jensen, 2021. Revisiting the notion of deleterious sweeps. Genetics. **219(3):** iyab094.

Kaplan, N. L., R. R. Hudson, and C. H. Langley, 1989. The "hitchhiking effect" revisited. Genetics. **123:** 887–899.

Kaushik, S. and K. Jain, 2021. Time to fixation in changing environments. Genetics. **219:** iyab148.

Kaushik, S., 2023. Effect of beneficial sweeps and background selection on genetic diversity in changing environments. J Theor Biol. **562:** 111431.

Kim, Y., and W. Stephan, 2000. Joint effects of genetic hitchhiking and background selection on neutral variation. Genetics. **155:** 1415–1427.

Kim, Y., and W. Stephan, 2002. Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics. **160:** 765–777.

Kim, Y., and W. Stephan, 2003. Selective sweeps in the presence of interference among partially linked loci. Genetics. **164:** 389–398.

Kim, Y., and R. Nielsen, 2004. Linkage disequilibrium as a signature of selective sweeps. Genetics. **167:** 1513–1524.

Kim Y, 2006. Allele frequency distribution under recurrent selective sweeps. Genetics. **172:** 1967–1978.

Kimura, M., 1968. Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. Genet Res. **11:** 247–269.

Kimura, M., 1969. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics. **61:** 893–903.

Kimura, M., 1971. Theoretical foundations of population genetics at the molecular level. Theor Pop Biol. **2:** 174–208.

Kosheleva, K. and M. M. Desai, 2013. The dynamics of genetic draft in rapidly adapting populations. Genetics. **195(3):** 1007–1025.

Martin, G., and A. Lambert, 2015. A simple, semi-deterministic approximation to the distribution of selective sweeps in large populations. Theor Pop Biol. **101:** 40–46.

Maruyama, T., and M. Kimura, 1974. A note on the speed of gene frequency changes in reverse directions in a finite population. Evolution. **28:** 161–163.

Min, J., M. Gupta, M. M. Desai and D. B. Weissman, 2022. Spatial structure alters the site frequency spectrum produced by hitchhiking. Genetics. **222(3):** iyac139.

Moghadasi, S. A., *et al.*, 2023. Transmissible SARS-CoV-2 variants with resistance to clinical protease inhibitors. Sci Adv. **9(13):** eade8778.

Pollak, E.,, 1987. On the theory of partially inbreeding finite populations. I. Partial selfing. Genetics. **117:** 353–360.

Przeworski, M., 2002. The signature of positive selection at randomly chosen loci. Genetics. **160(3):** 1179–1189.

Simonsen, K. L., G. A. Churchill, and C. F. Aquadro, 1995. Properties of statistical tests of neutrality for DNA polymorphism data. Genetics. **141:** 413–429.

Smith, J. M., and J. Haigh, 1974. Hitchhiking effect of a favourable gene. Genet Res. **23:** 23–35.

Maynard Smith, J., 1976. What determines the rate of evolution? Am Nat. **110(973):** 331–338.

Sniegowski, P. D., and P. J. Gerrish, 2010. Beneficial mutations and the dynamics of adaptation in asexual populations. Philos Trans R Soc Lond B Biol Sci. **365:** 1255–1263.

Stephan, W., T. H. E. Wiehe, and M. W. Lenz, 1992. The effect of strongly selected substitutions on neutral polymorphism: analytical results based on diffusion theory. Theor Pop Bio. **41:** 237–254.

Tung, H. R., and R. Durrett, 2021. Signatures of neutral evolution in exponentially growing tumors: A theoretical perspective. PLoS Comput Biol. **17(2):** e1008701.

Watterson, G. A., 1975. On the number of segregating sites in genetical models without recombination. Theor Popul Biol. **7(2):** 256–276.

Williams, M. J., B. Werner, C. P. Barnes, T. A. Graham, and A. Sottoriva, 2016. Identification of neutral tumor evolution across cancer types. Nat Genet. **48:** 238–244.

Wright, S., 1938. The distribution of gene frequencies under irreversible mutations. Proc Natl Acad Sci. USA **24:** 253–259.

Wright, S., 1969. Evolution and the genetics of populations, Volume 2: The theory of gene frequencies. University of Chicago Press, Chicago, IL, USA.

Živković, D., Y. S. Steinrücken, M. Song and W. Stephan, 2015. Transition densities and sample frequency spectra of diffusion processes with selection and variable population size. Genetics. **200:** 601–617.

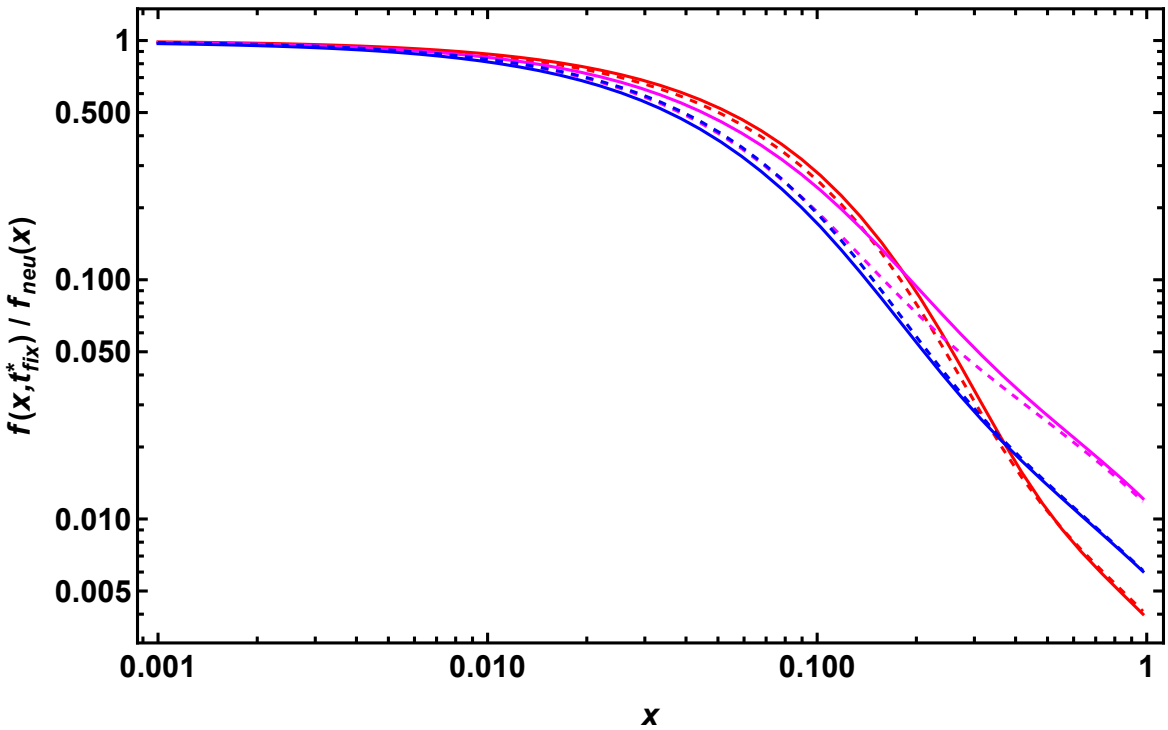

Figure S1: The SFS at neutral sites fully linked to a single new beneficial mutation immediately post-fixation relative to the neutral equilibrium SFS, $f_{neu}(x) = \theta/x$, for various dominance coefficients: a) $h = 0.2$ (magenta), b) $h = 0.5$ (blue), and c) $h = 0.8$ (red). The solid lines represent $f(x, t^*_{fix})$, calculated at the conditional mean fixation time of the beneficial mutation [see (7)], while the dashed lines depict the SFS averaged over the entire conditional fixation time distribution. Both sets of lines are derived using diffusion theory for unlinked neutral loci. The parameters used are $N = 10^3$, $\mu = 10^{-7}$, $s = 0.1$, $F = 0$, and $L = 10^5$.

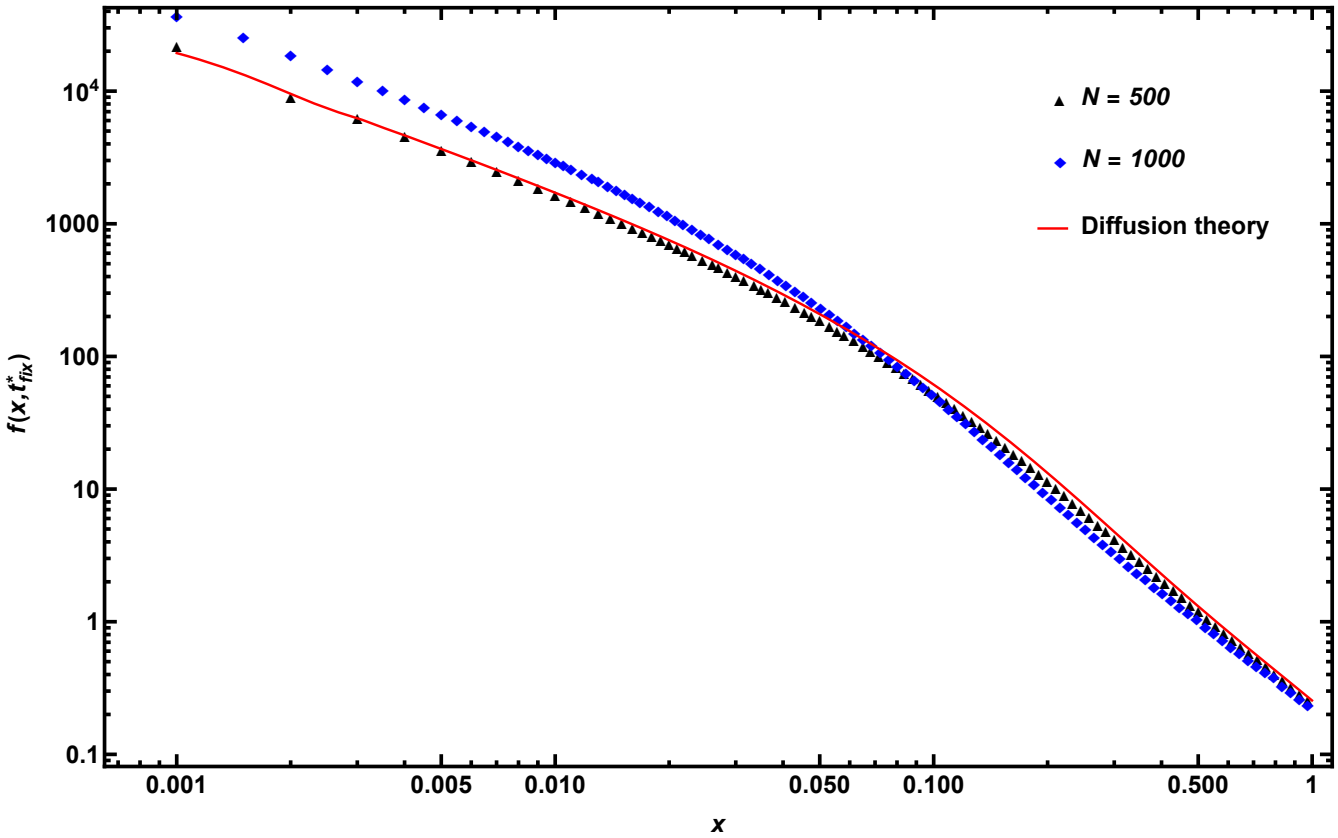

Figure S2: The SFS at neutral sites fully linked to a single new beneficial mutation immediately post-fixation is shown for varying population sizes: a) $N = 500$, and b)$N = 1000$, keeping the selection coefficient s fixed. The points are obtained using forward time simulations while the solid line is obtained using diffusion theory (1) for $N = 500$. The parameters used are $s = 0.1$, $\mu = 10^{-7}$, $L = 10^5$, $F = 0$, and $h = 0.5$.

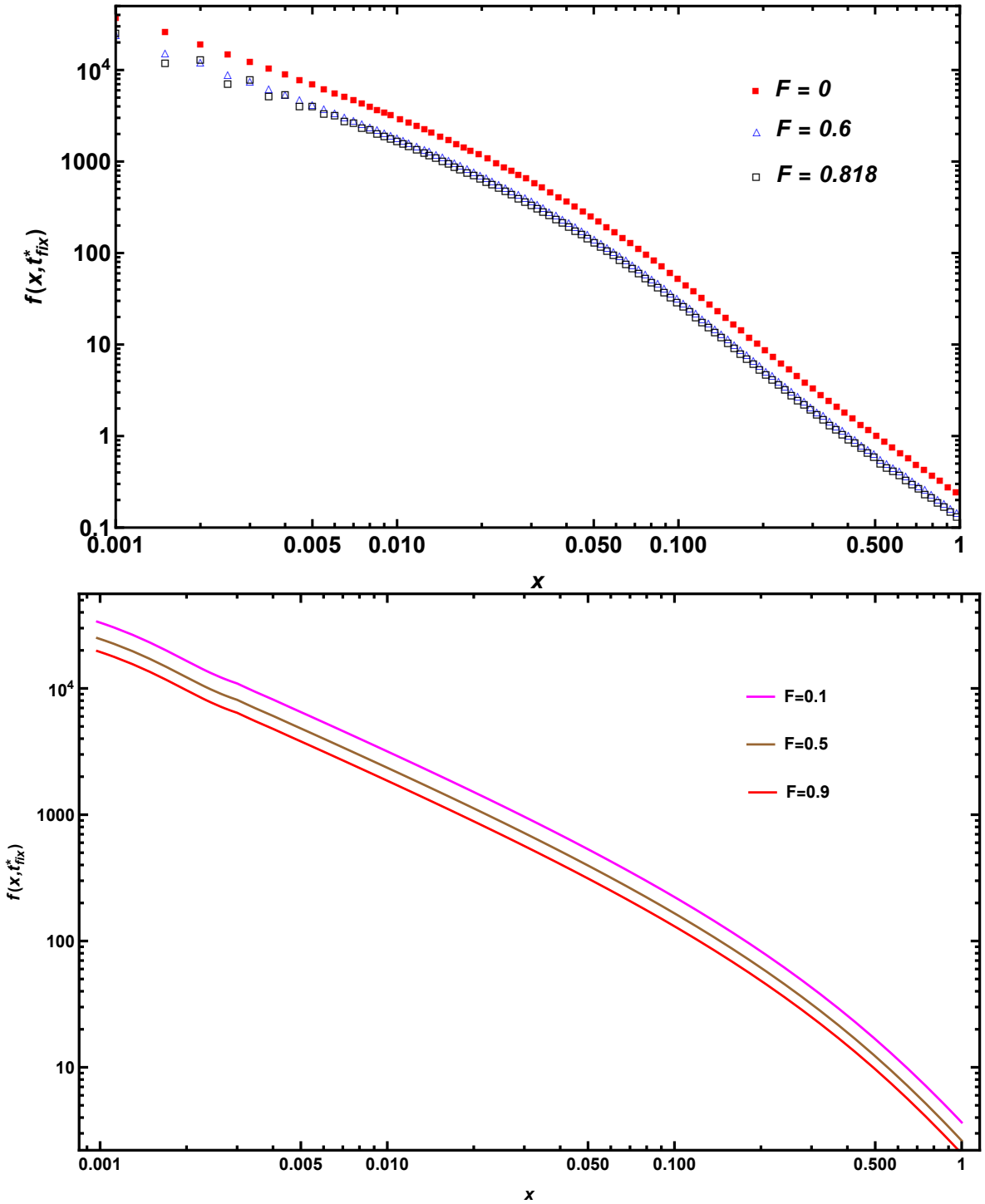

Figure S3: The SFS at neutral sites fully linked to a single new beneficial mutation immediately post-fixation is obtained from simulations (top) and diffusion theory (bottom) for various inbreeding coefficients. The other parameters used are $N = 10^3$, $\mu = 10^{-6}$ (top) and 10^{-7} (bottom), $s = 0.1$, $h = 0.5$ and $L = 10^4$ (top) and 10^5 (bottom).

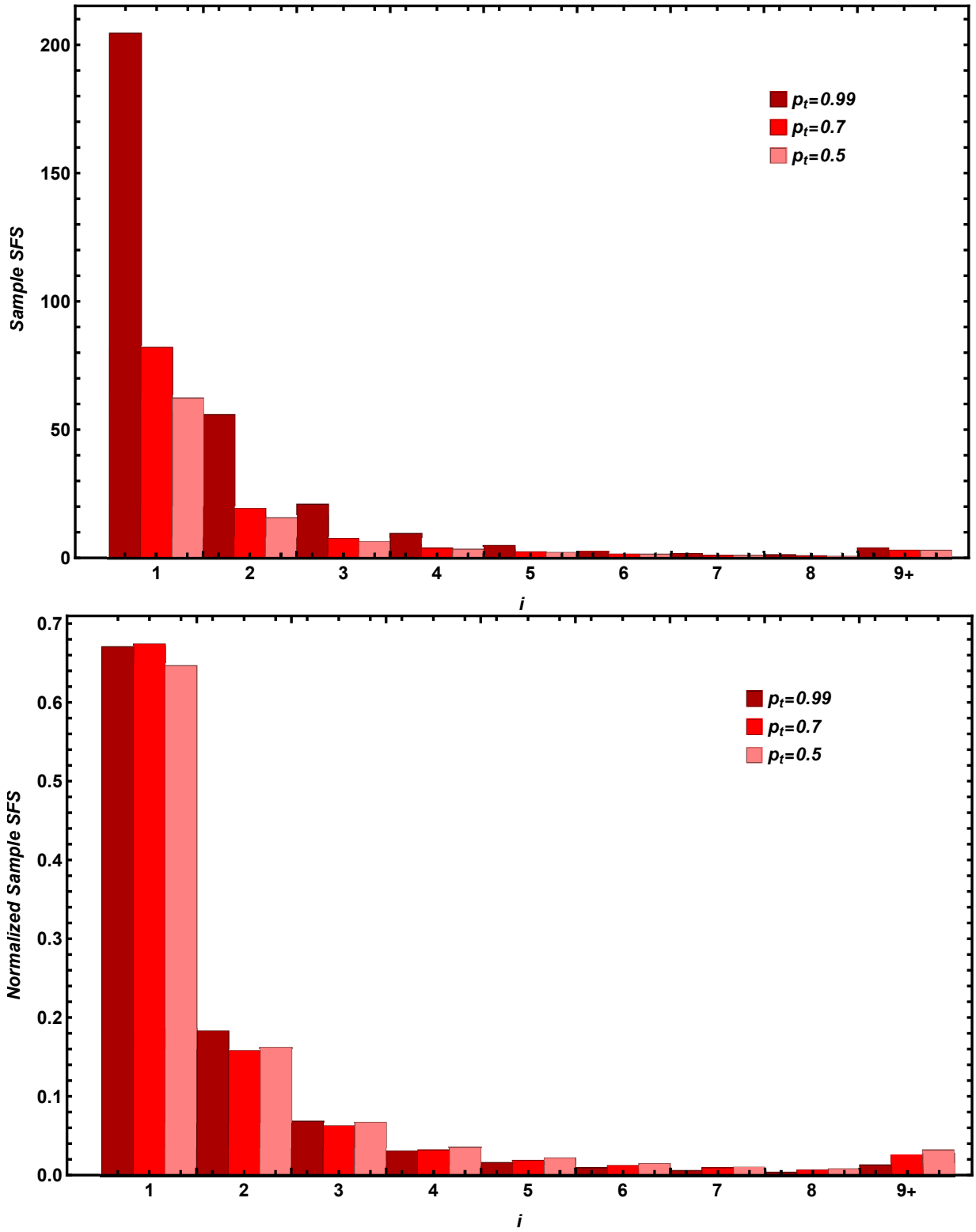

Figure S4: The sample SFS (above), and normalized sample SFS (below) at neutral sites when beneficial mutation frequency reaches the value mentioned in the legend. The data has been generated for 20 genomes from forward time simulations. The parameters used are $N = 10^3$, $\mu = 10^{-7}$, $s = 0.1$, $F = 0$, and $L = 10^6$.

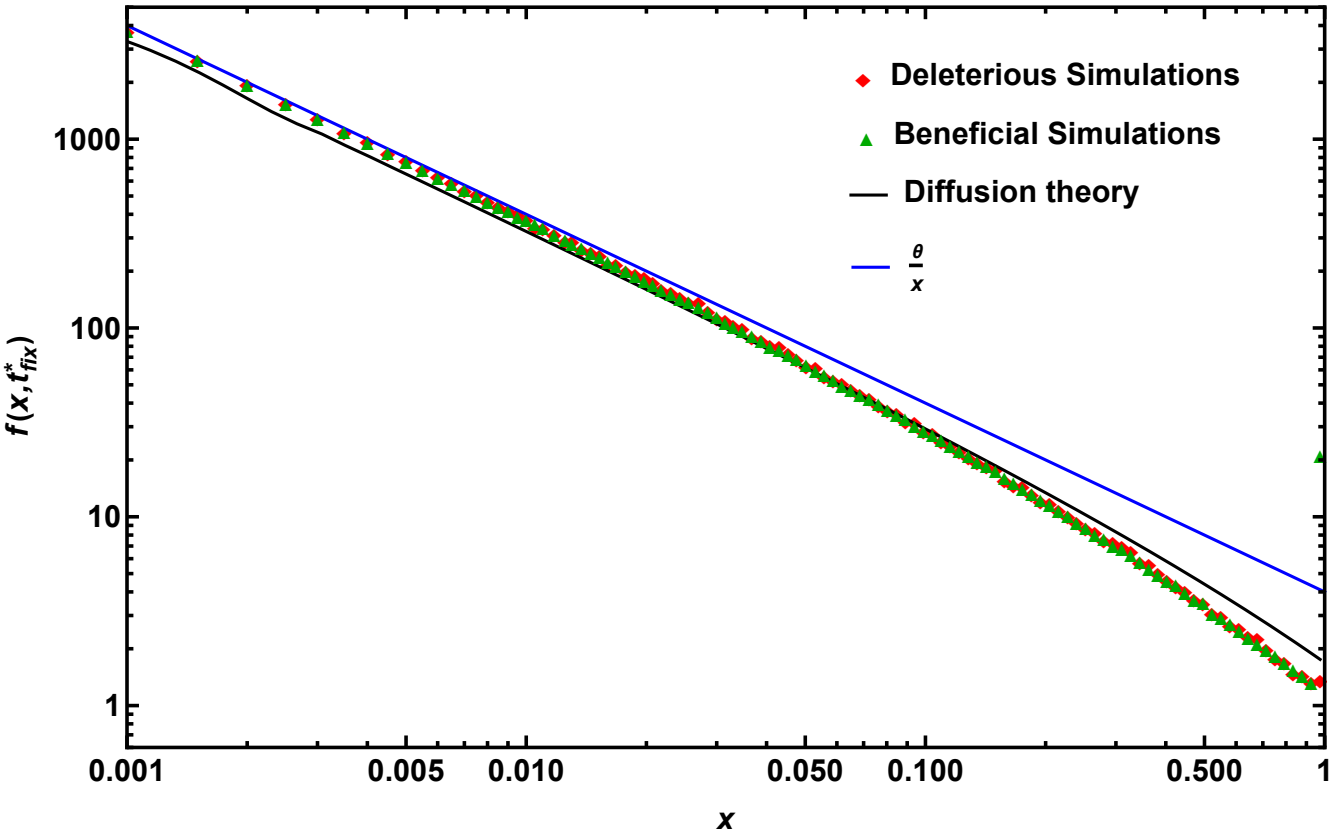

Figure S5: The simulation data for SFS at neutral sites when they are fully linked to a beneficial (green points) or a deleterious (red points) site immediately post-fixation are shown for $N = 1000$, and $|s| = 0.002$. The blue line shows the neutral equilibrium SFS, θ/x. The solid black line obtained by numerically solving (1) for the SFS in diffusion theory for beneficial hard sweep is shown for comparison and does not agree as well with the simulations results. The parameters used are $\mu = 10^{-7}$, $L = 10^4$, $F = 0$ and $h = 0.5$.

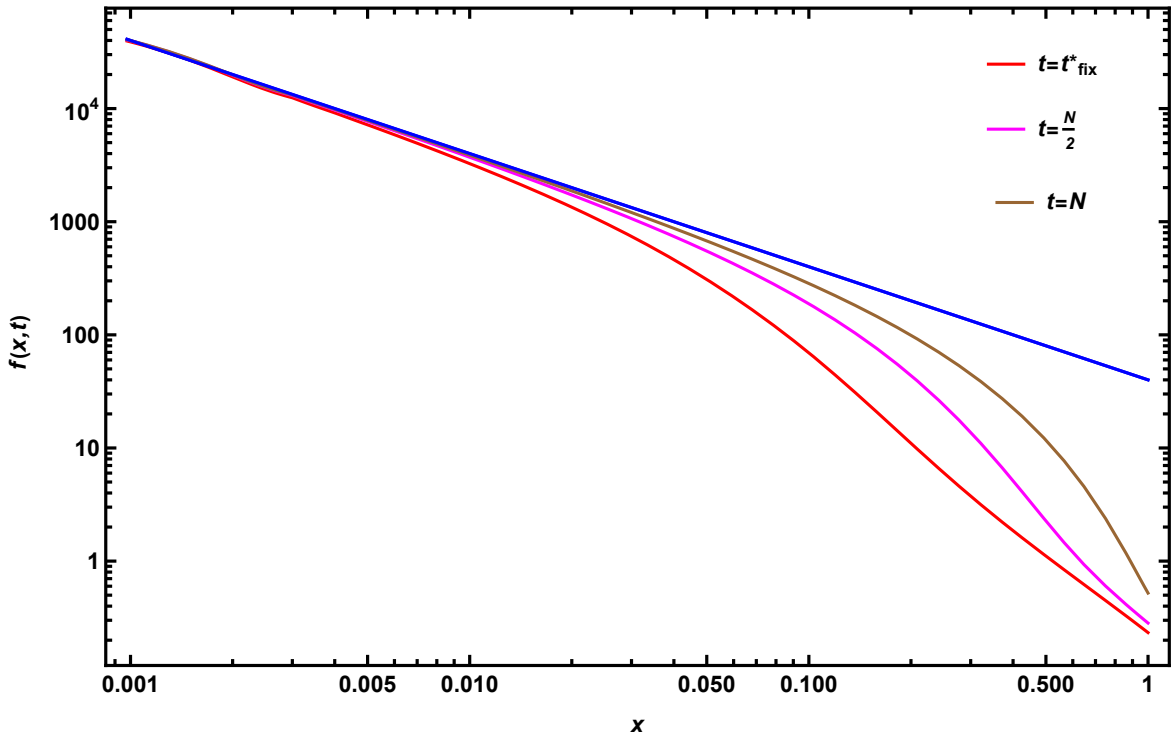

Figure S6: The recovery of the SFS at neutral sites fully linked to a single new beneficial mutation following a selective sweep is shown at different time points: $t = N/2$ (magenta) and $t = N$ (brown). The SFS immediately post-fixation is depicted in red. Over time, the SFS gradually returns to the neutral equilibrium form, θ/x (blue). The parameters used for this analysis are $N = 10^3$, $\mu = 10^{-7}$, $s = 0.1$, $h = 0.5$, $F = 0$, and $L = 10^5$.