# Strong neutral sweeps occurring during a population contraction

Antoine Moinet[1,2,3], Flavia Schlichta[2,3], Stephan Peischl*[1,2,4] and Laurent Excoffier*[2,3]

1. Interfaculty Bioinformatics Unit, University of Bern, Baltzerstrasse 6, 3012 Bern, Switzerland;
2. Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland;
3. Institute of Ecology and Evolution, University of Bern, Baltzerstrasse 6, 3012 Bern, Switzerland;
4. Corresponding author, *stephan.peischl@bioinformatics.unibe.ch*;

*. These authors contributed equally.

Keywords: bottleneck, selective sweep, genetic drift, range expansion, genome scan.

## Abstract

A strong reduction in diversity around a specific locus is often interpreted as a recent rapid fixation of a positively selected allele, a phenomenon called a selective sweep. Rapid fixation of neutral variants can however lead to similar reduction in local diversity, especially when the population experiences changes in population size, e.g., bottlenecks or range expansions. The fact that demographic processes can lead to signals of nucleotide diversity very similar to signals of selective sweeps is at the core of an ongoing discussion about the roles of demography and natural selection in shaping patterns of neutral variation. Here we quantitatively investigate the shape of such neutral valleys of diversity under a simple model of a single population size change, and we compare it to signals of a selective sweep. We analytically describe the expected shape of such “neutral sweeps” and show that selective sweep valleys of diversity are, for the same fixation time, wider than neutral valleys. On the other hand, it is always possible to parametrize our model to find a neutral valley that has the same width as a given selected valley. Our findings provide further insight in how simple demographic models can create valleys of genetic diversity similar to those attributed to positive selection.

## Introduction

Past demography and natural selection play a critical role in shaping extant genetic diversity. A central question in population genetics is to quantify their respective impact on observed genomic diversity. Because selection interferes with demographic estimates and vice versa, estimation of one of these two components is difficult without accounting for the other (Charlesworth *et al.* 1993, 1995; Kaiser and Charlesworth 2009; O'Fallon *et al.* 2010; Charlesworth 2013; Nicolaisen and Desai 2013; Johri *et al.* 2020, 2021b). Moreover, the relative importance of demography and selection as determinants of genome wide diversity is currently hotly debated, and may vary extensively among species (Corbett-Detig *et al.* 2015; Rousselle *et al.* 2018; Pouyet and Gilbert 2019; Galtier and Rousselle 2020). It has been shown that selection and demography can leave very similar footprints on the genetic diversity of a population (Andolfatto and Przeworski 2000; Teshima *et al.* 2006; Thornton and Jensen 2007; Johri *et al.* 2021a). Disentangling the effects of demography and selection is therefore crucial to avoid erroneous inference of evolutionary scenarios from genomic data (Jensen *et al.* 2005; Wares 2009; Mathew and Jensen 2015; Johri *et al.* 2020).

Hard selective sweeps lead to valleys of strongly reduced diversity around positively selected sites due to the hitchhiking of linked neutral loci (Maynard Smith and Haigh 1974), such observations of strong depletions of diversity in some genomic regions are often interpreted as due to past episode of positive selection, because the probability to observe a fast fixation of a neutral variant in a population of constant size is extremely low. However, during a range expansion for instance, some neutral or even mildly deleterious mutations can go quickly to fixation due to the low effective size of populations on the front of the range (Edmonds *et al.* 2004; Klopfstein *et al.* 2006; Hallatschek and Nelson 2008; Peischl *et al.* 2013), a phenomenon termed allele surfing (Klopfstein *et al.* 2006). Theoretical studies have shown that the average neutral diversity on the wave front decays exponentially as the range expands (Hallatschek and Nelson 2008), similarly to what happens when a population experiences a sudden decay of the population size, i.e. a population contraction, due to a drastic change in the environment for example. In both cases, a mutation appearing when the population size is shrinking might go quickly to fixation, inducing a strong decrease of diversity in the surrounding genomic region, whereas the average level of diversity might stay quite high depending on the strength and the duration of the contraction. As a result, the coalescent tree of alleles sampled in a population with strongly reduced effective population size will have short external branches, and long internal branches, depending on the parameters of the model (Excoffier *et al.* 2009). The average site frequency spectrum associated to such a tree resembles a neutral SFS, but with a lack of rare alleles and an excess of high frequency sites, i.e. it becomes “flatter” (Sousa *et al.* 2014; Peischl and Excoffier 2015). The footprint left by the rapid fixation of a neutral allele on the surrounding genomic diversity, might thus be like that of a positively selected allele sweeping through a constant size population.

The expected shape of nucleotide diversity in genomic regions surrounding a site undergoing a rapid neutral fixation has been investigated analytically and numerically. Tajima (1990) studied the reduction of diversity during a neutral fixation at a given recombination distance from the fixing site. His results rely on rigorous mathematical arguments based on diffusion theory, but no closed form solution is provided for the shape of a neutral sweep. Johri *et al.* (2021a) described the valley of diversity occurring around a neutral fixation using an approach introduced for selective sweeps, assuming that the evolution of the allele frequency is that of a selected allele except in the initial stochastic phase. Here, we extend this work by inferring the dynamics of fixation of neutral alleles after a population contraction and we examine their effects on neighboring regions of the genome. We provide an analytical result for the expected coalescence time as a function of the recombination distance from the locus undergoing a fast fixation. Importantly, our results apply regardless of the process driving the allele going to fixation (neutrality, positive selection, background selection), as it only relies on the typical trajectory of an allele going to fixation in a given time, even though this trajectory differs depending on the underlying driver of this fixation (i.e., neutrality or selection). We compare our results against simulations and find that they hold for a wide range of realistic parameter combinations. We compare our results about the signature of neutral sweeps to patterns expected under selective sweeps and discuss potential differences between the signatures that could potentially allow us to discriminate between neutral and selective processes for a given demographic scenario. Finally, we investigate the similarity between the genomic signature of an allele going to fixation either selectively or neutrally and observe that a selective sweep signal can in principle be replicated in a neutral model with an appropriate choice of demographic parameters. We conclude that strong diversity depletions in the genome of a population, often attributed to the effect of positive selection, can be obtained with demographic effects only, and we call for caution when trying to detect signals of adaptation from genomic data, adding support to previous studies reaching similar conclusions (Thornton and Jensen 2007; Crisci *et al.* 2013; Jensen *et al.* 2019).

## Model

We model here the effect of an instantaneous population contraction on genomic diversity. Throughout the whole manuscript, time is measured backwards. We assume that $t_c$ generations before the present, the population size instantaneously dropped from $N_0$ diploid individuals to $N_c$ individuals with $N_c < N_0$. We assume a standard coalescent model (Kingman 1982a; b) with discrete non-overlapping generations, random mating, monoecious individuals, and no selection. Two haplotypes sampled in the current population at time $t = 0$ have, as we go backwards in time, a constant probability $(2N_c)^{-1}$ of coalescing at each generation, for the first $t_c$ generations, and then this probability switches to $(2N_0)^{-1}$ as we enter the ancestral uncontracted population. We can approximate the distribution of coalescence time $T$ of these two haplotypes as a piecewise exponential distribution (see Appendix) with expected value:

$$\mathrm{E}[T] = 2(N_0 - N_c)\, e^{-t_c/2N_c} + 2N_c\,. \qquad (1)$$

We see that the expected coalescence time decreases exponentially with the age of the contraction $t_c$ and that it approaches $2N_c$ for a very old contraction. Coalescence times cannot be measured directly from empirical data, but they are closely related to nucleotide diversity $\pi$. Under the infinitely many sites model, the number of nucleotide differences between two homologous DNA segments is proportional to their coalescence time $T$ as $\pi = 2\mu T$, where $\mu$ is the total mutation rate for the whole segment. Multiplying eq. (1) by $2\mu$ shows that an instantaneous population contraction leads to an exponential decrease of the expected nucleotide diversity along the genome with the age of the contraction $t_c$. However, it does not inform us on the distribution of nucleotide diversity $\pi$ along the genome, or on spatially correlated patterns of diversity such as local depletion or excess of diversity relative to the expectation.

Fig. 1 shows the evolution of the distribution of $\pi$ as a function of the time $t_c$ elapsed since the contraction. For $t_c = 0$, there is no contraction, and the population size remains constant and equal to $N_0$. In this case we see (Fig. 1a,1b, $t_c = 0$) that the distribution of π is symmetric and centered at $E[\pi] = 4N_0\mu$. For an older contraction, we see that the distribution is not only shifted to lower values of diversity as expected from eq. (1), but that it also becomes strongly peaked around $\pi = 4N_c\mu$. This bimodality of the distribution can be understood intuitively in the following way. There are two possible types of coalescent trees for haplotypes sampled after the population contraction (note that the tree depends on the locus considered because of recombination). Indeed, the most recent common ancestor (MRCA) of the sample lived either before the contraction ($T_{\mathrm{MRCA}} > t_c$), or after the contraction ($T_{\mathrm{MRCA}} < t_c$). In the former case, the tree at this locus has long inner branches and short outer branches, whereas in the latter case, the tree is essentially a (short) neutral tree corresponding to a population of constant size $N_c$ (Excoffier *et al.* 2009). Both types of trees occur at different loci and correspond to the two observed modes in the distribution of the nucleotide diversity along the chromosome. The precise shape of the distribution of nucleotide diversity across sites depends on the relative frequency of both types of trees, which itself depends on the age of the contraction $t_c$. For a sample of size two, the probability that the MRCA lived after the contraction, that is, $T_{\mathrm{MRCA}} < t_c$

is $1 - e^{-t_c/2N_c}$. For a larger sample of haplotypes, there is no closed form solution for this probability, but the trees rooted after the contraction are rare for $t_c << 2N_c$ and very frequent when $t_c >> 2N_c$ (Tavaré 1984). Therefore, the evolution of the distribution of $\pi$ for increasing contraction age $t_c$ appears to be a transition from a unimodal distribution centered at $4N_0\mu$ to another unimodal distribution centered at $4N_c\mu$, with both modes coexisting for intermediate ages (Fig. 1). This bimodality has been pointed out previously in the context of population bottlenecks (Austerlitz *et al.* 1997); however, those studies mainly focused on long duration bottlenecks (the effect of a contraction or a bottleneck on nucleotide diversity is the same provided that the bottleneck is not yet finished, or that it finished very recently so that the effect of population recovery is negligible). In the present work, we investigate the effect of short contractions on the genetic diversity and make the claim that this short contraction regime is of particular interest as it can lead, such as in Fig. 1c, to genomic signatures similar to those generated by positive selection acting on a few sites in an otherwise neutral genome. More specifically, we want to quantitatively describe the reduction of diversity along the genome that is observed around a locus with a small $T_{\mathrm{MRCA}}$ (such as in Fig. 1c in the regions around 10-11 and 19-20 Mb), where we observe a valley or trough of diversity. Akin to what is done for selective sweeps, we consider the (neutral) fast fixation of an allele and analyze the impact of hitchhiking on the genetic diversity of neighboring loci, and we refer to this process as a neutral sweep.

To investigate neutral sweeps in our model, we consider the following scenario: $t_{\mathrm{m}}$ generations ago a mutation occurred at a single site on the chromosome, which we call the focal site. We further assume that this mutation has just fixed in the population, i.e., that it was segregating at a frequency strictly lower than one in the last generation (at $t = 1$) and has now (at $t = 0$) a frequency equal to one. We assume that the population contraction occurred $t_c$ generations ago, with $t_c \geq t_{\mathrm{m}}$. As the mutant enters the population as a single allelic copy at the focal locus, defined as a non-recombining region surrounding the focal site, this copy is a common ancestor for all the copies ($2N_c$) present at fixation. However, it is not necessarily the most recent common ancestor. Fig.2 shows a sketch of our model to help visualize how recombination can maintain diversity at linked loci around a locus where a new mutation quickly fixed in the population.

# Results

## Average coalescence time at a linked locus

We can calculate the expected coalescence time $T^{(\mathrm{l})}$ of two randomly sampled haplotypes at a linked locus as a function of the recombination rate *r* from the focal locus. The idea is to consider two haplotypes with a given coalescence time $T^{(\mathrm{f})}$ at the focal locus, and then follow the genealogy of the gene copies carried by these two haplotypes at the linked locus backward in time, while considering possible recombination events. The expected coalescent time at the linked locus is then

$$\mathrm{E}\left[T^{(\mathrm{l})}\right] = \left(1 - E\left[e^{-2r\sum_{\mathrm{t}=1}^{\mathrm{T}^{(\mathrm{f})}}(1-\overline{x}_t)}\right]\right)(t_m + T_m) + E\left[T^{(\mathrm{f})}\ e^{-2r\sum_{t=1}^{\mathrm{T}^{(\mathrm{f})}}(1-\overline{x}_t)}\right] \qquad (2)$$

where $\overline{x}_t$ is the average frequency of the mutant (derived) allele at the focal locus at time *t* counting backward from present. A detailed derivation of this equation is given in Appendix A4. The first term of the right-hand side of eq. (2) corresponds to cases where lineages escape the neutral sweep due to recombination, and still have not coalesced after $t_{\mathrm{m}}$ generations. In this case we need to wait on average $T_m = 2(N_0 - N_c)\, e^{-(t_c - t_m)/2N_c} + 2N_c$ extra generations before the lineages coalesce, due to the contraction that happened $t_{\mathrm{c}}$ - $t_{\mathrm{m}}$ generations before the focal mutation. The second term of the right-hand side of eq. (2) corresponds to cases where the lineages cannot escape the sweep and are forced to coalesce at a time $T^{(\mathrm{l})} \leq t_{\mathrm{m}}$.

## Distribution of coalescence times at the focal locus

To evaluate eq. (2), we need to determine the probability distribution of the pairwise coalescence times $T^{(\mathrm{f})}$ at the focal locus, as well as the expected frequency trajectory of the derived allele. Even though this allele fixes neutrally in a population of constant size (the contraction occurs prior to the mutation), the distribution of coalescent times at the focal locus

$T^{(f)}$ departs from the usual exponential distribution for a neutral coalescent process because the allele fixes in exactly $t_m$ generations, and hence the coalescence time for a randomly chosen pair of haplotypes is at most $t_m$. Slatkin (1996) investigated the coalescent process within a “mutant allelic class” that originated from a single mutation at a given time in the past. He derived exact analytical results for the average pairwise coalescence time, but the coalescence distribution itself can only be expressed with multidimensional integrals and obtaining a closed form expression does not appear feasible. We therefore use a different approach: given a particular fixation trajectory of the mutant allele, i.e. given the number of mutant copies $N_\mu$ at each generation between $t$ = 0 and $t$ = $t_m$, we can express the coalescence time distribution within the mutant allelic class, using the result of a coalescent in a population with a time-dependent (but deterministic) size $N_\mu(t)$ (Griffiths and Tavaré 1994). Averaging over all possible trajectories of the mutation, we obtain:

$$\mathrm{P}\left(\mathrm{T}^{(\mathrm{f})}\right) = \sum_{\{x_t\}} \left[\frac{1}{2N_c x_{\mathrm{T}^{(\mathrm{f})}}} \prod_{t=1}^{\mathrm{T}^{(\mathrm{f})}-1} \left(1 - \frac{1}{2N_c x_t}\right)\right] P(\{x_t\}) \quad (3\mathrm{a})$$

where $x_t = N_\mu(t)/(2N_c)$ is the frequency of the mutant $t$ generations from fixation, and $P(\{x_t\})$ is the probability of a given trajectory. $P(\{x_t\})$ can be evaluated (see Appendix A2) and the sum in eq. (3a) can in principle be computed numerically; however, the number of trajectories to consider is prohibitive. As a first approximation, we can replace $x_t$ by its expectation $\overline{x}_t$, i.e., we neglect the fluctuations of the trajectory around the mean to obtain

$$\mathrm{P}\left(\mathrm{T}^{(\mathrm{f})}\right) \simeq \frac{1}{2N_c\, \overline{x}_{\mathrm{T}^{(\mathrm{f})}}} \prod_{t=1}^{\mathrm{T}^{(\mathrm{f})}-1} \left(1 - \frac{1}{2N_c\, \overline{x}_t}\right). \qquad (3b)$$

The last step is to determine the average trajectory of an allele fixing in exactly $t_m$ generations. Zhao *et al.* (2013) as well as Maruyama and Kimura (Maruyama and Kimura 1975) have investigated the characteristic trajectory of an allele fixing in a given time but they do not provide a closed form solution. Here, we use a different approach (also based on diffusion theory to obtain an approximation for the average trajectory of an allele fixing in exactly $t_m$ generations, starting from a frequency $p_0$. As detailed in the Appendix A2, we obtain

$$\overline{x}_t = 1/2\left(1 - (1 - 2p_0)e^{-(\mathrm{t_m}-t)/N_c} + e^{-t/N_c}\right), \qquad (4a)$$

which is valid for $t_{\mathrm{m}} \gg 2N_c$. For very fast fixations, i.e., when $t_{\mathrm{m}} \ll 2N_c$, the frequency of the allele increases approximately linearly as

$$\overline{x}_t = 1 - (1 - p_0)\frac{t}{\mathrm{t_m}}. \qquad (4b)$$

We remind the reader that *t* is counted backwards from fixation. Fig. 3 compares equations (4a) and (4b) to trajectories obtained from simulations of a Wright-Fisher diploid population. We find good agreement between the simulations and the analytical results. Importantly, the typical neutral trajectory for large values of the fixation time has an “inverse-sigmoid shape” (Fig. 3c), contrary to the typical sigmoid trajectory of a positively selected allele going to fixation in a constant size population (see Fig. 5a). This neutral trajectory occurs because, conditional on non-loss, neutral alleles need to quickly escape loss at the beginning and remain at intermediate frequencies to stay away from both fixation and loss until they eventually fix in the population at *t* = 0 (*i.e.* in exactly $t_{\mathrm{m}}$ generations). Fig. 3e-3h also shows the coalescence time distribution for several values of the fixation time $t_{\mathrm{m}}$. The comparison of the distribution of pairwise coalescence time with numerical simulations of a Wright-Fisher model shows that our approximation eq. (3b) is quite accurate but overestimates the probability of coalescence for large coalescence times when $t_{\mathrm{m}}$ is small (Fig. 3d). Notably, coalescence (simulated or theoretical) is more probable at large times *(i.e.* when the mutant appeared) for short fixation times (Fig. 3d), whereas it is more probable at small times (i.e. close to fixation) for large fixation times (Fig. 3e). The coalescence rate within the mutant allelic class is given by the inverse of the number of mutant copies and is for all values of the fixation time slightly more than $1/2N_{\mathrm{c}}$ at the first generation. However, when the fixation time is short (Fig. 3e), there is a fast increase of the coalescence rate backwards in time, and many lineages are forced to coalesce at $t = t_{\mathrm{m}}$. When the fixation time is large (Fig. 3h), the coalescence rate also increases backwards in time, but the increase is much slower. In that case, most coalescence events happen in much less than $t_{\mathrm{m}}$ generations, so that the early increase in frequency of the mutant has almost no influence on the coalescence distribution.

## Effect of a neutral sweep on linked diversity

Combining equations (3b), (4a) with eq. (2) allows us to get an approximation for the average coalescence time at linked loci. Since the derivation of eq. (2) assumes that there is at most one recombination event in the genealogy of a randomly chosen pair of gene copies, we expect it to be only accurate for small values of the recombination rate *r*. For large values of *r* we use a heuristic approach combining the result of eq. (2), which is accurate for small *r*, and the expected diversity at unlinked loci, which is equal to $T_0 = 2(N_0 - N_c)\, e^{-t_c/2N_c} + 2N_c$ as stated in eq. (1). We fit the trough of diversity with an exponential function of the form:

$$\mathrm{E}\left[\mathrm{T}^{(\mathrm{l})}\right](r) = T_0(1 - ce^{-ar}), \qquad (5)$$

where the coefficients $c = 1 - E\left[\mathrm{T}^{(\mathrm{f})}\right]/T_0$ and $a = 2E[(t_m + T_m - \mathrm{T}^{(\mathrm{f})})\sum_{t=1}^{\mathrm{T}^{(\mathrm{f})}}(1 - \overline{x}_t)]/(T_0 - E\left[\mathrm{T}^{(\mathrm{f})}\right])$ are obtained by imposing that eqs. (2) and (5) coincide for small values of *r* (using a linear expansion in *r*). On Fig. 4 we compare the result of eq. (5) to Wright-Fisher simulations with two recombining loci. We see in Fig. 4a that the exponential function fits the data accurately at large values of the recombination distance, but that the fit is biased for intermediate values of *r*. In Fig. 4b we see that the approximation is very good for low values of the recombination distance, although there still is a slight bias. This discrepancy at small *r* can be corrected (solid lines in Fig. 4) if we use numerical estimations of $\overline{x}_t$ and $\mathrm{P}\left(\mathrm{T}^{(\mathrm{f})}\right)$, instead of eqs. (4) and (3b), to evaluate eq. (5).

We observe, as expected, on Fig. 4 that the troughs of diversity induced by neutral sweeps are wider and deeper for short fixation times. Similarly to what happens after a selective sweep, there is less opportunity for linked loci to escape the sweep by recombination and maintain diversity when the fixation is fast. In addition, the diversity level at the center of the valley is given by the average coalescence time at the focal locus, which quickly decreases for small fixation times $t_m$.

## Comparison of neutral sweeps and selective sweeps

Since we did not make any assumption regarding the process driving the mutant allele to fixation when deriving the average coalescence time at linked loci (eq. (2)) and the coalescence time distribution at the focal locus (eq. (3b)), our framework allows us to directly compare the signatures of different processes that can drive mutations to fixation in a given number of generations. We illustrate this by comparing the effect of neutral and hard selective sweeps on linked diversity. Later we will discuss how neutral sweeps compare to a larger variety of scenarios (e.g. background selection, small selection coefficients, or dominant alleles). Here we assume that the neutral and selected fixations occurred over the same time interval, that is in both cases in exactly $t_{\mathrm{m}}$ generations. The selected fixation is assumed to be codominant (*h*=0.5) and occurs on an autosomal locus in a randomly mating diploid population of constant size $N_1$, and we consider a strong selection strength ($2N_1s >> 1$) so that the allele frequency follows the deterministic trajectory

$$\overline{\mathrm{x}}_{\mathrm{t}} = \frac{1}{1 + (2N_1 - 1)\, e^{-2(1-t/t_{\mathrm{m}})\log(2\mathrm{N}_1)}}, \qquad (6)$$

where the fixation time is given by $t_{\mathrm{m}}(s) = 2\log(4N_1 s)/s$ (Barton 1995). Then combining eqs. (5), (3b) and (6), we can compute the average coalescence time at linked loci as a function of the recombination distance *r* to the focal locus, after replacing $T_{\mathrm{m}}$, the average coalescence time at $t = t_{\mathrm{m}}$, by $2N_1$ in eq. (5) and $N_{\mathrm{c}}$ by $N_1$ in eq. (3b). This approach yields results similar to Charlesworth (2020), where the author investigated signals of selective sweeps correcting for coalescent events that happen during the sweep, thus going beyond the common assumption of a star tree structure at the focal locus. For sake of simplicity in the neutral case, we consider that the mutant appeared at the time of the contraction, i.e. $t_{\mathrm{m}} = t_{\mathrm{c}}$. Furthermore, we will assume that the average coalescence times (and consequently the genetic diversity) are equal in both scenarios, i.e. that $T_0 = 2N_1$ which implies that

$$N_0(\mathrm{t_m}) = (N_1 - N_c)\, e^{t_m/2N_c} + N_c\ . \qquad (7)$$

In the neutral case we want the diversity to remain as high as $4N_1\,\mu$ after the contraction, which is possible only if the ancestral diversity was even higher, i.e. we have in general $N_0 > N_1 > N_c$.

In Fig. 5a, we compare the mutant average frequency as a function of time for a selected and a neutral fixation. The dynamics of the neutral fixation is the opposite of that of the selected allele in the sense that when one is increasing, the other is "resting" and vice versa. These different trajectories translate into different coalescence distributions at the focal locus (Fig. 5b). If selection drives the fixation of the mutation, the distribution of coalescence time is peaked at large coalescence times. In contrast, in the neutral case the distribution is skewed towards small coalescence times. Correspondingly, the coalescence tree for the selected case has a star-like structure (Hermisson and Pennings 2017), whereas the tree for the neutral case has shorter outer branches. Therefore, for a given recombination distance, there will be fewer recombinations on the neutral tree because it has a much smaller total length. As recombination helps maintain diversity at linked loci, we would expect neutral troughs of diversity to be wider than in the selected case. However, this is at odds with the valleys of diversity observed in Fig. 5c, where the selective trough is wider than the neutral trough. Even though recombinations occur less frequently on the neutral tree as compared to a selected tree, a recombination on the neutral tree is more likely to lead to a change of genomic background from derived to ancestral allele due to the inverse sigmoid neutral trajectory of the derived allele. Recombination on the neutral tree will thus more often lead to a lineage escaping the sweep, resulting in more efficient recovery of diversity in the neutral case for a given genomic distance from the focal locus. Furthermore, we see that the trough is deeper in the neutral case (Fig. 5c), since the average coalescence time is smaller at the focal site due to the smaller total length of the coalescence tree.

To determine if these differences between selective and neutral troughs hold for other fixation times and population sizes, we define two quantities that characterize the shape of a trough, as well as its propensity to be detected in real data: i) the trough relative depth and ii) the width of the trough. The relative depth is defined as the difference between the background level of diversity and the diversity at the focal locus, divided by the background diversity, and the width is measured at half depth, *i.e*. halfway between the background diversity and the diversity at the focal locus. On Fig. 6 we plot the relative depth of neutral and selective troughs as a function of their width for different fixation times $t_{\mathrm{m}}$, calculated with our analytical expressions. We see that the neutral troughs are not only always narrower than the selective troughs for the same value of $t_{\mathrm{m}}$, but also deeper. This is due to differences in the focal tree structure between the selective case and the neutral case as well as difference in the ancestral background level in both cases, as explained above. For very short fixation times (corresponding to selection coefficients larger than 0.1), there is almost no difference between troughs generated by selective and neutral sweeps. Indeed, for such values of $t_{\mathrm{m}}$, in both cases the focal coalescence tree is essentially a star tree because the increase in frequency is very fast, and the ancestral backgrounds of diversity, $2N_0$ and $2N_1$, are also practically equal. Note however that at small $t_{\mathrm{m}}$ the corresponding value of the selection coefficient *s* (see legend of Fig. 6) may be unrealistically high. For realistic values of the selection coefficient/fixation time, the neutral troughs tend to be quite deep but narrow, whereas selective troughs are wider and their depth decreases quickly for low selection coefficients. From Fig. 6, we see that the shape of a neutral trough is generally different from a selective sweep signal, but in practice those differences might be hidden due to the noise inherent present in real genomic data, and it might be difficult to decide whether a genomic signal is a due to a neutral sweep or a selective sweep.

## Discussion

It has repeatedly been suggested that strong depletions of diversity in the genome are not necessarily due to the presence of positive selection (Johri *et al.* 2020), and can also be the result of demographic effects only, such as the allele surfing phenomenon occurring at the front of a range expansion (Klopfstein *et al.* 2006). In this work, we considered a model of population contraction to analyze quantitatively the genomic signature of the rapid fixation of a mutation during a population contraction, but it should also apply in case of range expansions or recurrent founder events by considering the harmonic mean of population sizes. Taking a step further from previous work that focused on the impact of range expansion on mere allele frequencies, we have studied here the impact of a neutral allele fixation on neighboring genomic diversity. We show that the diversity profile around a recently fixed locus crucially depends on the frequency trajectories of the allele going to fixation, and we outline the fact that neutrally fixing alleles have an inverse-sigmoid trajectory (Fig. 3d), as compared to the standard sigmoid frequencies observed for positively selected alleles. For the same fixation time, this difference translates into different genomic signatures (see figs. 5c and 6). Our results demonstrate that there is a short period after a demographic contraction (or during a range expansion) where observed profiles of genomic diversity would look like those usually attributed to selection (Fig. 1c), and that selective sweep signals can be mimicked by neutrally fixing mutations without the need to invoke complex histories of population size changes.

Our results allow for a systematic comparison of selective and neutral troughs of diversity, and we used our results to investigate trough shapes for range of neutral and selected scenarios (see Fig. 6), which in principle can be used to decide whether a given empirical trough is due to selection or demography, and to infer the corresponding parameters. However, we did not consider the whole spectrum of possible selection scenarios. It would be indeed interesting to use our results to study cases of background selection, small selection coefficients, and a variety of dominance coefficients. All these cases should have their own characteristic trajectories of fixation, and hence potentially different genomic signatures. In addition, in our model we do not consider mutations that fixed in the past (we always assume that the allele has just reached fixation), nor do we consider mutations appearing before the population contraction, i.e., with $t_m > t_c$. The average coalescence time in the former case can be expressed as a function of the coalescence time at fixation using conditional probabilities, and we can show that a sweep signal vanishes exponentially with the time elapsed since fixation (see Appendix A4). In the latter case, we can solve the problem by considering the number of gene copies at $t_c$ that descend from the original copy that appeared at $t_m$. One could extend our results by considering an allele starting from an arbitrary number of copies at $t_c$, akin to soft selective sweeps; however, the analytic calculations are complex, and we leave this study for future research. In any case, those additional scenarios must be considered when trying to infer models from the study of troughs found in empirical data. Another phenomenon that renders the inference of parameters cumbersome is a possible interference between troughs. Indeed, when two loci fix neutrally in the population, the genetic diversity in the region between those loci will be influenced by both fixations and will differ from the diversity expected in the vicinity of a single fixing locus. As in the case of interference between the fixation of selected alleles (Weissman and Barton 2012), this should limit the number of independent neutral fixations. The effect of trough interference is stronger for neighboring troughs, and the probability to observe close troughs depends on the relative frequency of troughs along the genome, which itself depends on the distribution of the $T_{MRCA}$. In Fig. 1d for example, the distribution of $T_{MRCA}$ has a mode centered around $4N_c$ (not shown) and correspondingly the nucleotide diversity is peaked around $4N_c\,\mu$. As a result, we see many regions of the chromosome with a low diversity. It is likely that those troughs interfere with each other and that they do not correspond to the profile of an isolated trough. On the other hand, in Fig. 1c, the first mode of the $T_{MRCA}$ distribution is truncated because $t_c$ is much smaller than $4N_c$, and only $T_{MRCA}$s equal or close to $t_c$ are observed (plus all the $T_{MRCA}$s corresponding to the second mode centered at $4N_0$). In this case there is no interference and the (rare) troughs, such as the one in Fig. 7, are correctly fitted by their theoretical expectation. Those considerations imply that, even though we know the forward in time probability that an allele will fix in $t_m$ generations, it is difficult to infer the parameters of a fixation scenario from a single observed neutral valley of diversity. It appears therefore difficult to perform model selection from a single trough signal, i.e., to decide whether a particular trough is due to selection or demographic effects, because alternative demographic scenarios that we did not consider here could also lead to similar signals.

We performed simulations to investigate the signature of a neutral rapid fixation on the Site Frequency Spectrum (SFS) (see Appendix . We chose demographic parameters such that troughs are not numerous along the genome, and leave a strong footprint on genomic diversity.

Out of 10,000 simulations of 20Mb chromosomes, only 432 exhibit a (single) region of highly reduced diversity (here arbitrarily set to less than 7% of the background diversity). By averaging over all these valleys of diversity, we calculated the average SFS observed in a 15Kb window at the center of the valley, and obtained a U-shape SFS, which is also expected around a selective sweep (Huber *et al.* 2016). However, contrary to a fixation driven by selection (Suppl. Figure S1), the SFS around a neutral fixation shows a slight excess of variants at intermediate frequencies. This is probably due to the fact that some neutral haplotypes have spent more time at intermediate frequencies before going to fixation than selected haplotypes that rapidly "jump" from very low to very high frequencies (see Fig. 5a). Note also that the background (genome wide) SFS away from neutral sweeps has a global excess of intermediate and high frequency variants compared to a constant size population. This excess of high frequency variants is typical of populations having gone through a recent population size reduction or a bottleneck (Marth *et al.* 2004) due to the higher coalescence rate during the population contraction. These differences in expected SFS around neutral and selected sweeps could help decide whether regions of low diversity observed in empirical data are due to selection or to demographic processes. However, since very few variants are usually observed in the vicinity of single troughs, the empirical SFS in such a region might be too noisy to confidently identify the cause of the diversity reduction. In principle, if several troughs of diversity were observed in a genome, one could use the distribution of trough shapes and pooled SFS expected under a given simple demographic model and a distribution of fitness effect to compare neutral and selection models under a likelihood framework, but such an exploration is beyond the scope of the present paper.

In conclusion, our results suggest that any empirical valley of diversity found in empirical data can be reproduced neutrally with a population contraction using appropriate parameters. One could argue that this identifiability problem disappears once the true evolutionary history is correctly inferred. However, inferring the true demographic history requires precise knowledge about how selection has shaped genomic diversity (Johri *et al.* 2020). In humans, for instance, it has been estimated that roughly 95 % of genomic diversity is affected by some form of non-neutral forces such as background selection or biased gene conversion (Pouyet *et al.* 2018) potentially biasing demographic inference (Ewing and Jensen 2016). These considerations indicate than genome scans in search for signals of adaptation might be more affected by past demography than previously thought. We thus believe that despite current advances using supervised machine learning or similar approaches (Schrider and Kern 2018), it remains important to further study the effect of neutral fixations in various demographic scenarios using localized genomic approaches such as the present analytical work (Johri *et al.* 2021b), as well as with controlled experiments on real living organisms where both the selected locus and the population history are known (Orozco-terWengel *et al.* 2012). Such work will be critical in order to develop more appropriate evolutionary null models for statistical inference (Hahn 2008; Johri *et al.* 2020).

## Data availability

The authors affirm that all data necessary for confirming the conclusions of the article are present within the article, figures, and tables.

## Acknowledgment

We are grateful to Montgomery Slatkin, Brian Charlesworth, Jeff Jensen, Rimberly Gilbert and three anonymous reviewers for their helpful comments.

## Funding

This work was partially supported by a Swiss NSF grant No 310030_188883 to LE.

## Competing interest

None to declare

# Bibliography

Andolfatto P., and M. Przeworski, 2000 A genome-wide departure from the standard neutral model in natural populations of Drosophila. Genetics 156: 257–268.

Austerlitz F., B. Jung-Muller, B. Godelle, and P.-H. Gouyon, 1997 Evolution of coalescence times, genetic diversity and structure during colonization. Theor. Popul. Biol. 51: 148–164.

Barton N. H., 1995 Linkage and the limits to natural selection. Genetics 140: 821–841.

Charlesworth B., M. T. Morgan, and D. Charlesworth, 1993 The effect of deleterious mutations on neutral molecular variation. Genetics 134: 1289–1303.

Charlesworth D., B. Charlesworth, and M. T. Morgan, 1995 The pattern of neutral molecular variation under the background selection model. Genetics 141: 1619–1632.

Charlesworth B., 2013 Background selection 20 years on: the Wilhelmine E. Key 2012 invitational lecture. J. Hered. 104: 161–171.

Charlesworth B., 2020 How Good Are Predictions of the Effects of Selective Sweeps on Levels of Neutral Diversity? Genetics 216: 1217–1238.

Corbett-Detig R. B., D. L. Hartl, and T. B. Sackton, 2015 Natural selection constrains neutral diversity across a wide range of species. PLoS Biol. 13: e1002112.

Crisci J. L., Y.-P. Poh, S. Mahajan, and J. D. Jensen, 2013 The impact of equilibrium assumptions on tests of selection. Front. Genet. 4: 235.

Edmonds C. A., A. S. Lillie, and L. Luca Cavalli-Sforza, 2004 Mutations arising in the wave front of an expanding population. Proc. Natl. Acad. Sci. U. S. A. 101: 975–979.

Ewens W. J., 2004 *Mathematical Population Genetics: I. Theoretical Introduction*. Springer, New York, NY.

Ewing G. B., and J. D. Jensen, 2016 The consequences of not accounting for background selection in demographic inference. Mol. Ecol. 25: 135–141.

Excofffier L., N. Marchi, D. A. Marques, R. Matthey-Doret, A. Gouy, *et al.*, 2021 fastsimcoal2: demographic inference under complex evolutionary scenarios. Bioinformatics. https://doi.org/10.1093/bioinformatics/btab468

Excoffier L., M. Foll, and R. J. Petit, 2009 Genetic Consequences of Range Expansions. Annu. Rev. Ecol. Evol. Syst. 40: 481–501.

Galtier N., and M. Rousselle, 2020 How Much Does Ne Vary Among Species? Genetics 216: 559–572.

Griffiths R. C., and S. Tavaré, 1994 Ancestral inference in population genetics. Stat. Sci. 9: 307–319.

Hahn M. W., 2008 Toward a selection theory of molecular evolution. Evolution 62: 255–265.

Hallatschek O., and D. R. Nelson, 2008 Gene surfing in expanding populations. Theor. Popul. Biol. 73: 158–170.

Haller B. C., and P. W. Messer, 2019 SLiM 3: Forward genetic simulations beyond the Wright-Fisher model. Mol. Biol. Evol. 36: 632–637.

Hermisson J., and P. S. Pennings, 2017 Soft sweeps and beyond: understanding the patterns and probabilities of selection footprints under rapid adaptation. Methods Ecol. Evol. 8: 700–716.

Huber C. D., M. DeGiorgio, I. Hellmann, and R. Nielsen, 2016 Detecting recent selective sweeps while controlling for mutation rate and background selection. Mol. Ecol. 25: 142–156.

Jensen J. D., Y. Kim, V. B. DuMont, C. F. Aquadro, and C. D. Bustamante, 2005 Distinguishing between selective sweeps and demography using DNA polymorphism data. Genetics 170: 1401–1410.

Jensen J. D., B. A. Payseur, W. Stephan, C. F. Aquadro, M. Lynch, *et al.*, 2019 The importance of the Neutral Theory in 1968 and 50 years on: A response to Kern and Hahn 2018. Evolution 73: 111–114.

Johri P., B. Charlesworth, and J. D. Jensen, 2020 Toward an Evolutionarily Appropriate Null Model: Jointly Inferring Demography and Purifying Selection. Genetics 215: 173–192.

Johri P., B. Charlesworth, E. K. Howell, M. Lynch, and J. D. Jensen, 2021a Revisiting the Notion of Deleterious Sweeps. Genetics. https://doi.org/10.1093/genetics/iyab094

Johri P., K. Riall, H. Becher, L. Excoffier, B. Charlesworth, *et al.*, 2021b The Impact of Purifying and Background Selection on the Inference of Population History: Problems and Prospects. Mol. Biol. Evol. 38: 2986–3003.

Kaiser V. B., and B. Charlesworth, 2009 The effects of deleterious mutations on evolution in non-recombining genomes. Trends Genet. 25: 9–12.

Kingman J. F. C., 1982a The coalescent. Stochastic Process. Appl. 13: 235–248.

Kingman J. F. C., 1982b On the genealogy of large populations. J. Appl. Probab. 19A: 27–43.

Klopfstein S., M. Currat, and L. Excoffier, 2006 The fate of mutations surfing on the wave of a range expansion. Mol. Biol. Evol. 23: 482–490.

Marth G. T., E. Czabarka, J. Murvai, and S. T. Sherry, 2004 The Allele Frequency Spectrum in Genome-Wide Human Variation Data Reveals Signals of Differential Demographic History in Three Large World Populations. Genetics 166: 351–372.

Maruyama T., and M. Kimura, 1975 Moments for sum of an arbitrary function of gene frequency along a stochastic path of gene frequency change. Proc. Natl. Acad. Sci. U. S. A. 72: 1602–1604.

Mathew L. A., and J. D. Jensen, 2015 Evaluating the ability of the pairwise joint site frequency spectrum to co-estimate selection and demography. Front. Genet. 6: 268.

Maynard Smith J., and J. Haigh, 1974 The hitch-hiking effect of a favorable gene. Genet. Res. 23: 23–35.

Nicolaisen L. E., and M. M. Desai, 2013 Distortions in genealogies due to purifying selection and recombination. Genetics 195: 221–230.

O'Fallon B. D., J. Seger, and F. R. Adler, 2010 A continuous-state coalescent and the impact of weak selection on the structure of gene genealogies. Mol. Biol. Evol. 27: 1162–1172.

Orozco-terWengel P., M. Kapun, V. Nolte, R. Kofler, T. Flatt, *et al.*, 2012 Adaptation of Drosophila to a novel laboratory environment reveals temporally heterogeneous trajectories of selected alleles. Mol. Ecol. 21: 4931–4941.

Peischl S., I. Dupanloup, M. Kirkpatrick, and L. Excoffier, 2013 On the accumulation of deleterious mutations during range expansions. Mol. Ecol. 22: 5972–5982.

Peischl S., and L. Excoffier, 2015 Expansion load: recessive mutations and the role of standing genetic variation. Mol. Ecol. 24: 2084–2094.

Pouyet F., S. Aeschbacher, A. Thiéry, and L. Excoffier, 2018 Background selection and biased gene conversion affect more than 95% of the human genome and bias demographic inferences. Elife 7: e36317.

Pouyet F., and K. J. Gilbert, 2019 Towards an improved understanding of molecular evolution: the relative roles of selection, drift, and everything in between. arXiv [q-bio.PE].

Rogers R. L., T. Bedford, A. M. Lyons, and D. L. Hartl, 2010 Adaptive impact of the chimeric gene Quetzalcoatl in Drosophila melanogaster. Proc. Natl. Acad. Sci. U. S. A. 107: 10943–10948.

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

Schrider D. R., and A. D. Kern, 2018 Supervised Machine Learning for Population Genetics: A New Paradigm. Trends Genet. 34: 301–312.

Slatkin M., 1996 Gene genealogies within mutant allelic classes. Genetics 143: 579–587.

Sousa V., S. Peischl, and L. Excoffier, 2014 Impact of range expansions on current human genomic diversity. Curr. Opin. Genet. Dev. 29: 22–30.

Tajima F., 1990 Relationship between DNA polymorphism and fixation time. Genetics 125: 447–454.

Tavaré S., 1984 Line-of-descent and genealogical processes, and their applications in population genetics models. Theor. Popul. Biol. 26: 119–164.

Teshima K. M., G. Coop, and M. Przeworski, 2006 How reliable are empirical genomic scans for selective sweeps? Genome Res. 16: 702–712.

Thornton K. R., and J. D. Jensen, 2007 Controlling the false-positive rate in multilocus genome scans for selection. Genetics 175: 737–750.

Wares J. P., 2009 Evolutionary dynamics of transferrin in Notropis. J. Fish Biol. 74: 1056–1069.

Weissman D. B., and N. H. Barton, 2012 Limits to the rate of adaptive substitution in sexual populations. PLoS Genet. 8: e1002740.

Zhao L., M. Lascoux, A. D. J. Overall, and D. Waxman, 2013 The characteristic trajectory of a fixing allele: a consequence of fictitious selection that arises from conditioning. Genetics 195: 993–1006.

***Figure 1**. Nucleotide diversity of a population experiencing a contraction, as a function of the time $t_c$ elapsed since the contraction, measured in units of $2N_c$. (a) distribution of nucleotide* *diversity as a function of time, nucleotide diversity along the chromosome at $t_c = 0$ (panel b), at $t_c = 0.25$ (panel c) and at $t_c = 0.75$ (panel d). Population size before contraction $N_0 = 2.37\times10^6$ and after contraction $N_c = 4{,}400$. Mutation rate $\mu = 5.42\times10^{-10}$ per site per generation. Recombination rate $r = 3.5\times10^{-8}$ per site per generation. Chromosome size L = 20 Mb. Window size 10 Kb sliding at 1 Kb intervals. Sample size: 30 haplotypes. These parameters are taken from* Rogers *et al.* (2010). *Simulations were performed with fastsimcoal2* (Excofffier *et al.* 2021).

***Figure 2**. Instantaneous population contraction with a subsequent neutral fixation. A mutant (green star) appeared $t_m$ generations ago and has just fixed neutrally in a diploid population that experienced a contraction $t_c$ generations ago. We represent the population as a set of $2N_c$ two-locus haplotypes that are painted so that the gene copies present at t = 0 can be traced back to $t = t_m$. Due to recombination, haplotype i carries a red gene copy at the linked locus at t = 0. Correspondingly, the coalescence time $T^{(l)}$ of the haplotypes i and j at the linked locus (black tree) is larger than $t_m$. On the other hand, the coalescence time $T^{(f)}$ at the focal locus (green tree) is smaller than $t_m$ because at this locus all gene copies descend from the same haplotype (due to the fixation of the focal mutation).*

***Figure 3**. Average frequency (a-d) and coalescence time distribution (e-h) of an allele fixing in a diploid population of constant size $N_c = 20$ in exactly $t_m$ generations, starting as a single copy (i.e. $p_0 = (2N_c)^{-1}$). The red dots are the results of Wright-Fisher simulations, and the black and white dashed lines are calculated with eqs. (4b) (first and second columns) (4a) (third and fourth columns) and (3b). In panes (a-d) we show the variability of the fixation process by overlapping 1780 fixing trajectories. The (numerically estimated) probability, for a mutant that appears at the onset of the contraction, to fix in less than $t_m$ generations is 0.006, 0.16, 0.64 and 0.86 for $t_m$ = 20, 40, 80 and 120 respectively (for this particular value of $N_c$).*

***Figure 4**. Average coalescence time at a linked locus, as a function of the recombination distance from the focal locus where a mutant fixed in exactly $t_m$ generations, starting from a single copy $t_m$ generations ago. $t_m = 15$ in black, $t_m = 20$ in red and $t_m = 40$ in blue. The dots are calculated with two-locus WF simulations, and compared to eq. (5) with either a numerical estimation (solid lines) or a theoretical estimation (dashed lines) of $\overline{x}_t$ and $P\left(T^{(\mathrm{f})}\right)$. $N_c = 20$. $N_0 = 1500$. The population experienced a contraction $t_c = t_m$ generations ago.*

***Figure 5**. Comparison between troughs of diversity resulting from a selective sweep (black) and a neutral sweep (red), for the same fixation time $t_m = 120$ (corresponding to s ≈ 0.1 in the selective case). Frequency of the fixing allele as a function of time (a), coalescence time distribution (b) and diversity around the fixing site along the genome using eq. (5) (c). $N_1 = 1500$, $N_c = 20$ and $N_0 = 2.97\times10^4$.*

***Figure 6**. Relative depth as a function of the width of the diversity troughs, for different values of $t_m$ and $N_c$ in the neutral case and for selective scenarios with identical fixation times. $t_m$ goes from 1 to 333 by increments of 1, the corresponding values of the selection coefficient s are indicated on the left of the legend bar (for all of them we have $N_1 s >> 1$). $N_1 = 1500$. $N_0$ is given by eq. (7) and depends on $N_c$ and $t_m$. The jumps in the neutral curves for $N_c = 20, 40, 60, 80$ and 100 are due to the use of two different approximations for the frequency of the mutant, eqs. (4a) and (4b) and are located at $t_m = 2N_c$.*

# Appendix

## A1. Coalescence distribution after a contraction

We want to determine the coalescence time of two lineages in a population that experienced a contraction $t_m$ generations ago, from a diploid size $N_0$ to $N_c$. As we go backward in time, the coalescence rate switches from $(2N_c)^{-1}$ to $(2N_0)^{-1}$ at $T = t_c$. The probability distribution might still be approximated by a piecewise exponential density:

$$\begin{aligned} f_0(T) &= \frac{1}{2N_c}\exp\left(-\frac{T}{2N_c}\right) \quad \text{for} \quad 0 < T < t_c \\ &= \frac{1}{2N_0}\exp\left(-\frac{t_c}{2N_c}\right)\exp\left(-\frac{T-t_c}{2N_0}\right) \quad \text{for} \quad T \geq t_c \end{aligned}$$

The corresponding expectation for this distribution is

$$\begin{aligned} E[T] = T_0 &= \int_0^\infty T\ f_0(T)\,\mathrm{d}T \\ &= 2N_0\, e^{-t_c/2N_c} + 2N_c\left(1 - e^{-t_c/2N_c}\right) \end{aligned}$$

## A2. Average frequency of an allele fixing in exactly $t_m$ generations

In this section time is counted forward from the mutation, which appears after the contraction, so that during the fixation the diploid population size is constant and equal to $N_c$. We condition on the fixation time $t_m$ of the mutant. We define the trajectory of a mutant as the list of frequencies at all generations: $\{x_t\} = \left(x_0, x_1, \ldots, x_{t_m-1}, x_{t_m}\right)$. We assume that the mutant fixes in exactly $t_m$ generations, starting from a frequency $p_0$, i.e. $x_0 = p_0$, $0 < x_{t_m-1} < 1$ and $x_{t_m} = 1$. The probability that the mutant follows a given trajectory might be expressed as the product of the transition probabilities

$$\mathrm{P}(\{x_t\}) = \prod_{t=0}^{\mathrm{t_m}-1} P(i, t \to j, t+1 \mid \text{fix in } t_m, p_0)$$

For an unconditional Wright Fisher model, $\mathrm{P(i, t \to j, t+1)}$ is the probability to have j copies of the new allele at $\mathrm{t+1}$ given that there were i copies at t. We note $P_t(i \to j)$ for brevity. If we only consider trajectories fixing in exactly $t_m$ generations and starting from a number $2N_c\, p_0$ of copies at $\mathrm{t} = 0$, then the transition probabilities are not equal to the transitions of the unconditional Wright-Fisher model. However, thanks to Bayes theorem, we can write

$$\begin{aligned} P_t(i \to j \mid \text{fix in } t_m, p_0) &= \frac{P_t(\text{fix in } t_m \mid i \to j, p_0) P_t(i \to j \mid p_0)}{P(\text{fix in } t_m \mid p_0)} \\ &= \frac{P(\text{fix in } t_m \mid j_{t+1}) P_t(i \to j)}{P(\text{fix in } t_m \mid p_0)} \end{aligned} \qquad (S1)$$

From the first to the second line, we use the Markov property. The three terms involved in the right-hand side of this equation can be approximated thanks to diffusion theory. In this framework, the probability for an allele to fix in $t_m$ generations, given that there were i copies at time t is approximately (Ewens 2004)

$$\mathrm{P}(\text{fix in } t_m \mid i_\mathrm{t}) = \frac{3}{2N_c}\left(1 - \frac{i}{2N_c}\right)\frac{i}{2N_c}\; e^{-(t_m - t)/2N_c} \qquad (S2)$$

The term $P_t(i \to j)$ is the unconditional binomial transition probability of the Wright Fisher model (which does not depend on t). In principle, eq. (S1) can be used to compute the exact distribution of coalescence times at the focal locus, using eq. (3a). However, the huge number of possible trajectories fixing in $t_m$ generations ($(2N_c - 1)^{\mathrm{t_m}-1}$) makes the average over trajectories impossible to evaluate numerically. For this reason, we use the approximation in eq. (3b).

We consider here the probability that the allele has frequency x at time t, given that it started at frequency $p_0$ at $\mathrm{t} = 0$. Again if we only consider trajectories that fix in exactly $t_m$ generations, this probability is not equal to the neutral diffusive result. However, similarly to the previous section, we can use Bayes theorem:

$$\mathrm{P}(x_t\,|\mathrm{fix\,in}\,t_m,\,p_0) = \frac{P(\mathrm{fix\,in}\,t_m\,|\,x_t)P(x_t\,|\,p_0)}{P(\mathrm{fix\,in}\,t_m\,|\,p_0)}$$

From diffusion theory (Ewens 2004), we also have

$$\mathrm{P}(x_t\,|\,p_0) = 6p_0(1-p_0)\,e^{-t/2N_c}\big(1+5(1-2p_0)(1-2x)e^{-t/N_c}\big)$$

which is a second order expansion of an infinite series involving vanishing exponential terms ($e^{-k(k+1)t/4N_c}$ for all $\mathrm{k} \geq 1$). This expansion is thus valid in the limit of large times $\mathrm{t} \gg 2N_c$. We deduce that the probability that an allele fixing in $t_m$ generations has frequency x at time t is

$$\mathrm{P}(x_t\,|\mathrm{fix\,in}\,t_m,\,p_0) = 6\mathrm{x}(1-x)\big(1+5(1-2p_0)(1-2x)e^{-t/N_c}\big)$$

654 which yields $\mathrm{E}[x_t\,|\mathrm{fix\,in}\,t_m,\,p_0] = 1/2\big(1-(1-2p_0)e^{-t/N_e}\big)$

This expression is valid for $\mathrm{t_m} \gg t \gg 2N_c$, and does not allow one to estimate the frequency close to fixation. If we evaluate this expression for a given value of *t*, we must assume that *t*<sub></sub>$_{\mathrm{m}}$ is much larger than *t* (otherwise (S2) is not accurate). It implies that we cannot evaluate the frequency close to fixation, because wherever we "look", the fixation is always much later in time. Consequently, we see that $\mathrm{E}[x_t]$ tends to $1/2$ when *t* is very large, which is the only possible value for an average frequency infinitely far away from both fixation (at *t* = $t_{\mathrm{m}}$) and loss (at *t* = 0) . However, we know that the frequency should be symmetric, *i.e*. the allele should on average approach fixation in the same way it escapes loss, because the neutral fixation of a derived allele is the same as the loss of the ancestral allele. We thus write

$$\mathrm{E}[x_t\,|\,\mathrm{fix\,in}\,t_m,\,p_0] = 1/2\big(1-(1-2p_0)e^{-t/N_c}+e^{-(\mathrm{t_m}-t)/N_c}\big)$$

When $\mathrm{t_m} \ll 2N_c$, we can use a linear approximation for the trajectory (based on the numerical observations)

$$\mathrm{E}[x_t\,|\mathrm{fix\,in}\,t_m,\,p_0] = p_0 + (1-p_0)\frac{t}{\mathrm{t_m}}$$

### A3. Coalescence distribution at linked loci around a neutral fixation

We now return to the scenario of Fig. 2, with a backward in time approach. Using Bayes theorem, we express the coalescence time of two haplotypes at the linked locus $T^{(l)}$, conditioning on the coalescence time at the focal locus $T^{(f)}$

$$\mathrm{P}\big(T^{(l)}\big) = \int_0^{\mathrm{t_m}} \mathrm{P}\big(T^{(l)}\,\big|\,T^{(f)}\big)\; P\big(T^{(f)}\big)\mathrm{d}T^{(f)} = \mathrm{E}\big[\mathrm{P}\big(T^{(l)}\,\big|\,T^{(f)}\big)\big]$$

We assume that the linked locus is close to the focal locus on the chromosome, more precisely that the recombination rate $\mathrm{r}$ is very small $\mathrm{r} \ll 1$, so that we consider at most one recombination, occurring on one of the two focal lineages. We distinguish cases where there is no recombination between $t$ = 0 and $t = T^{(f)}$, cases where the allele at the linked locus recombines (somewhere between $t$ = 0 and $t = T^{(f)}$) onto a haplotype carrying the ancestral allele at the focal locus, and cases where the allele at the linked locus recombines onto a haplotype carrying the derived allele at the focal locus. We call the second and third case heterozygous and homozygous recombination, respectively, referring to the zygosity at the focal locus of the recombining pair of haplotypes (note that are three haplotypes, the two first ones have a coalescence time $T^{(f)}$, and the third one recombines with one of these two). If there is no recombination, then the coalescence time is the same for both loci, $T^{(l)} = T^{(f)}$. To treat the case with a homozygous recombination, it is convenient to name the haplotypes: *i* and *j* coalesce at $T_{ij}^{(f)} = T^{(f)}$ at the focal locus, and *k* is a third haplotype, onto which the linked allele recombines (coming from *i*). The linked allele carried by *j* stays on the same haplotype (no more than one recombination), and after recombining onto *k*, the linked allele initially carried by *i* also stays on *k* (again, at most one recombination). This implies that those two linked alleles coalesce at $T_{ij}^{(l)} = T_{jk}^{(f)}$. This time is in general different than $T_{ij}^{(f)}$, however on average $T_{jk}^{(f)}$ tand $T_{ij}^{(f)}$ are equal (averaging over all possible coalescence trees at the focal locus). This implies that we can treat the case with homozygous recombination as if there was no recombination. If there is a heterozygous recombination between *i* and *k*, at some generation between $t$ = 0 and $t$

= $T^{(f)}$, then the linked alleles still have not coalesced at $t = t_m$ because after the recombination one of them is linked to a derived focal allele and the other one to an ancestral focal allele (and they stay linked because there is at most one recombination). In that case, $T_{ij}^{(l)}$ is equal to $t_m$ plus a random time given by (on average) $T_m$, and is independent of $T_{ij}^{(f)}$. Using again Bayes theorem and the previous results to write

$$\begin{aligned}\mathrm{P}\left(T^{(l)} \,\middle|\, T^{(f)}\right) = {} & \mathrm{P}\left(T^{(l)} \,\middle|\, T^{(f)}, \quad \textit{one het}.\mathrm{rec.\ in}\left[0,\, T^{(f)}\right]\right)P\left(\textit{one het}.\mathrm{rec.\ in}\left[0,\, T^{(f)}\right]\right) \\ & + \mathrm{P}\left(T^{(l)} \,\middle|\, T^{(f)}, \quad \textit{no het}.\mathrm{rec.\ in}\left[0,\, T^{(f)}\right]\right)P\left(\textit{no het}.\mathrm{rec.\ in}\left[0,\, T^{(f)}\right]\right) \\ & = f_m\left(T^{(l)} - t_m\right)\left[1 - P\left(\textit{no het}.\mathrm{rec.\ in}\left[0,\, T^{(f)}\right]\right)\right] \\ & + \delta\left(T^{(l)} - T^{(f)}\right)P\left(\textit{no het}.\mathrm{rec.\ in}\left[0,\, T^{(f)}\right]\right)\end{aligned}$$

Where $\delta(\cdot)$ is the Dirac delta function, and $f_m$ is the unconditional coalescence distribution of a pair of lineages sampled at $t = t_m$, i.e. it is equal to the function $f_0$ introduced above but replacing $t_c$ by $t_c - t_m$ (note also that $f_m(t) = 0$ if $t < 0$). We then have to evaluate the probability that there is no heterozygous recombination. At generation $t$ (counted backward) the probability that a linked allele recombines onto a haplotype carrying the ancestral allele at the focal locus is $r(1 - x_t)$, where $x_t$ is the frequency of the derived allele at the focal locus, we deduce that the probability that there is no heterozygous recombination on either lineage is

$$\begin{aligned}P\left(\textit{no het}.\mathrm{rec.\ in}\left[0,\, T^{(f)}\right]\right) & = \prod_{t\,=\,1}^{T^{(f)}} \left(1 - r[1 - x_t]\right)^2 \\ & \simeq \exp\left(-2r \sum_{\mathrm{t=1}}^{T^{(f)}} (1 - x_t)\right)\end{aligned}$$

This probability depends explicitly on the allele trajectory, which means that rigorously, all the calculations should be conditioned on a given trajectory, and then averaged over all trajectories. To allow for mathematical tractability, and to avoid heavy expressions, we consider that as a good approximation $x_t = \overline{x}_t$. Finally we obtain

$$\mathrm{P}\big(T^{(l)}\big) = E\left[\delta\big(T^{(l)} - T^{(f)}\big)\exp\left(-2r\sum_{\mathrm{t}=1}^{T^{(f)}}(1-x_t)\right)\right] + f_m\big(T^{(l)} - t_m\big)\, E\left[1 - \exp\left(-2r\sum_{\mathrm{t}=1}^{T^{(f)}}(1-x_t)\right)\right]$$

The expectation corresponding to this distribution yields eq. (2).

## A4. Average coalescence time at a linked locus around a mutation that completed fixation *t*~fix~ generations ago

Thanks to Bayes theorem we can write

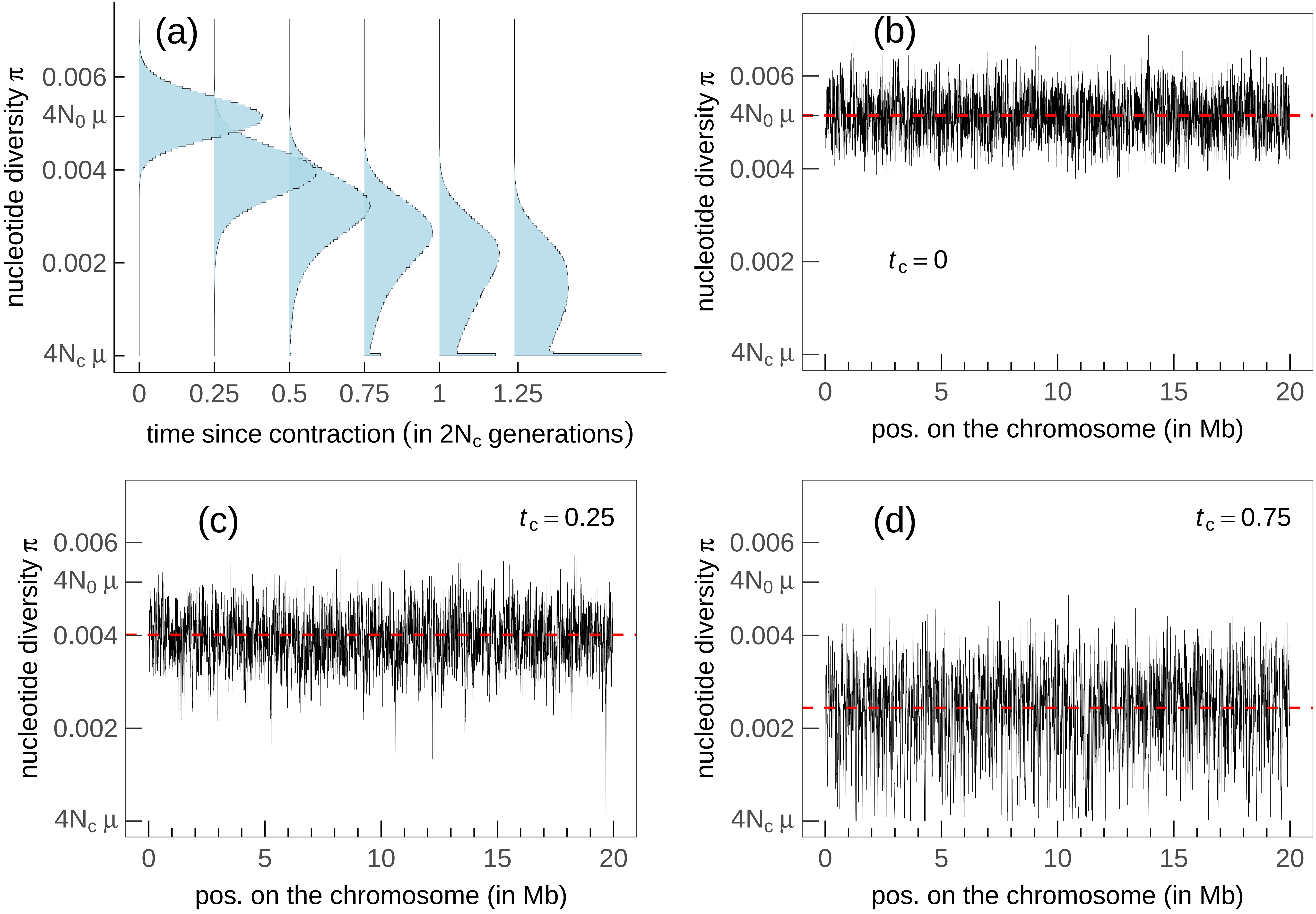
$$\mathrm{E}\big[T^{(l)}\big] = E\big[T^{(l)}\big|T^{(l)} < t_{\mathrm{fix}}\big]P\big(T^{(l)} < t_{\mathrm{fix}}\big) + E\big[T^{(l)}\big|T^{(l)} > t_{\mathrm{fix}}\big]P\big(T^{(l)} > t_{\mathrm{fix}}\big)$$

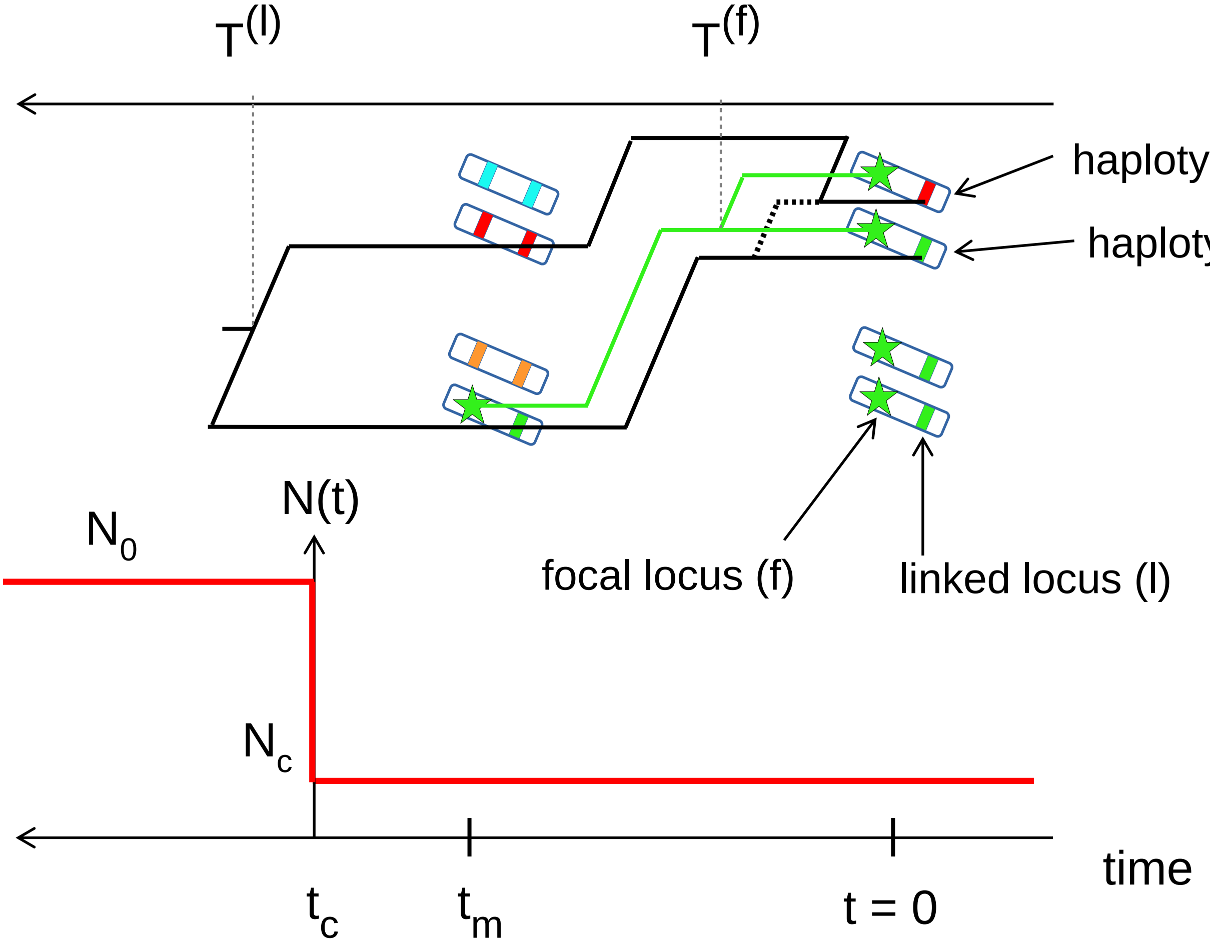
i.e. we distinguish coalescence events happening in less than $t_{\mathrm{fix}}$ generations or more than $t_{\mathrm{fix}}$ generations. In the former case, the coalescence is neutral, unconditional (the fixation is completed) and happens in a population of constant size $N_c$ which means that $E\big[T^{(l)}\big|T^{(l)} < t_{\mathrm{fix}}\big]$ and $P\big(T^{(l)} < t_{\mathrm{fix}}\big)$ can be worked out from the neutral exponential distribution. On the other hand, $E\big[T^{(l)}\big|T^{(l)} > t_{\mathrm{fix}}\big]$ is equal to $t_{\mathrm{fix}}$ plus the expectation from eq. (5) which we note here $\mathrm{E}\big[T^{(l)}\big](t = t_{\mathrm{fix}})$. We obtain

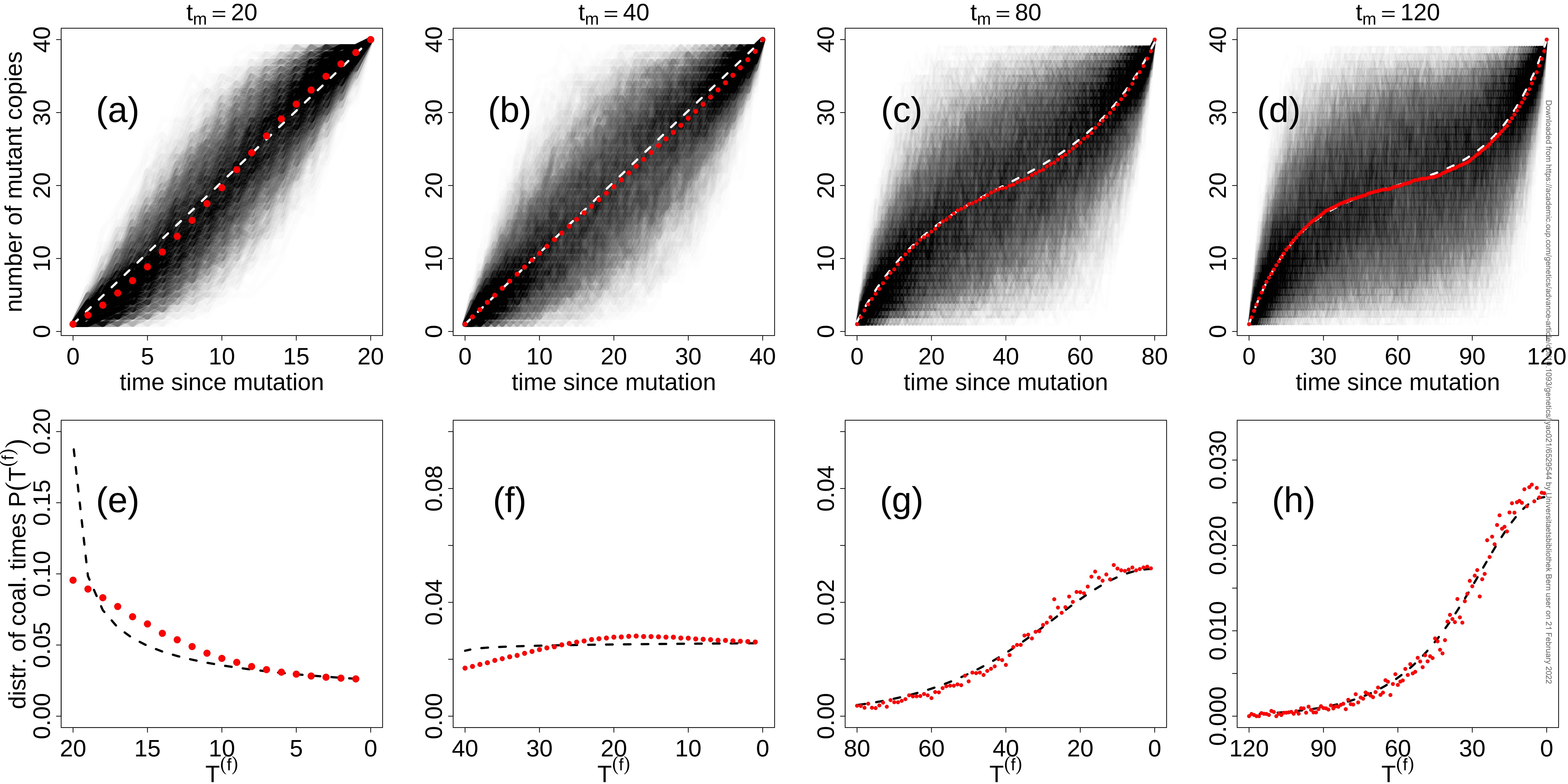
$$E\big[T^{(l)}\big] = 2N_c\big(1 - e^{-t_{\mathrm{fix}}/2N_c}\big) + \mathrm{E}\big[T^{(l)}\big](t = t_{\mathrm{fix}})\, e^{-t_{\mathrm{fix}}/2N_c}$$

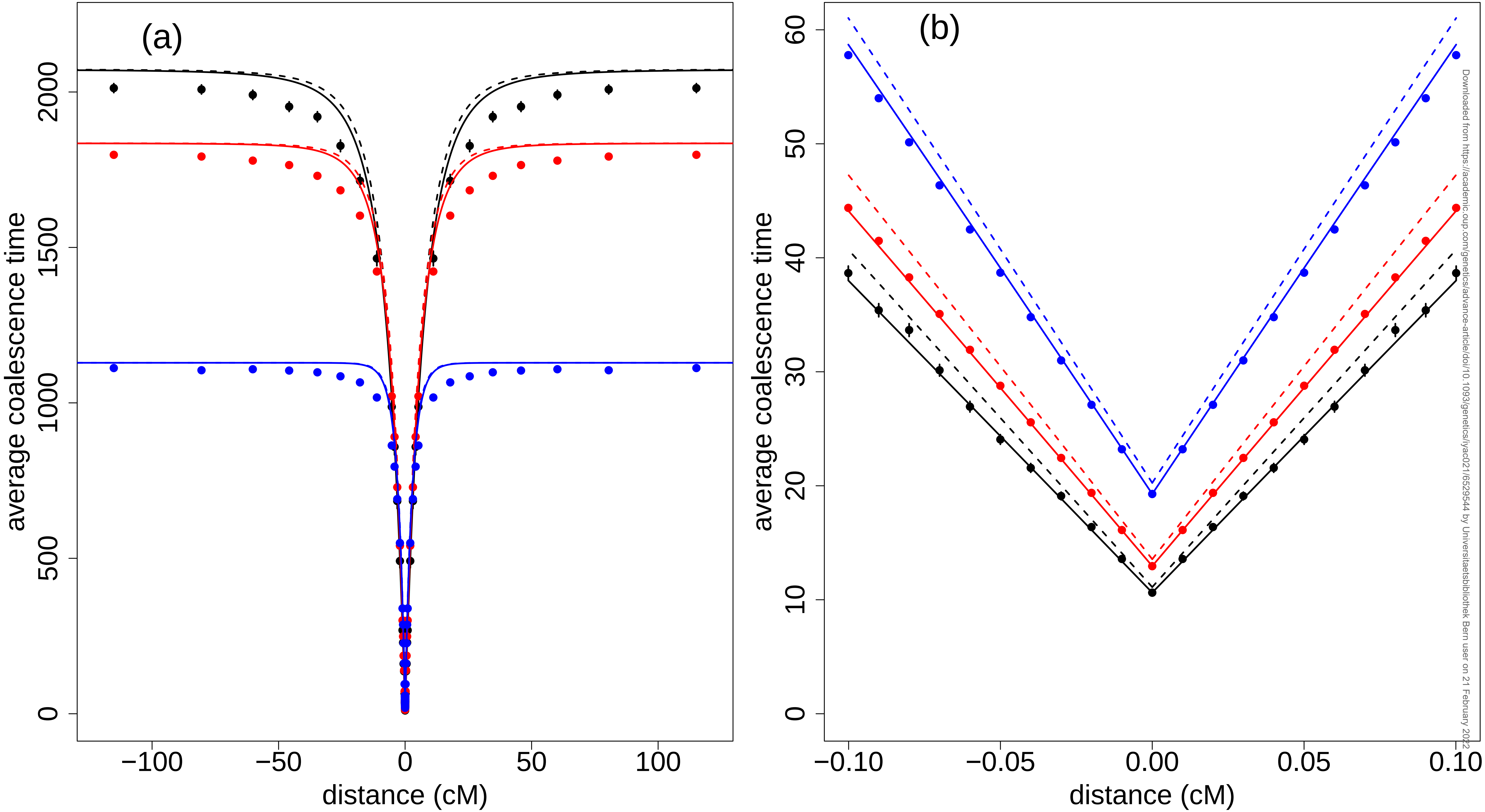
We see that the sweep signal vanishes exponentially with the time elapsed since fixation.

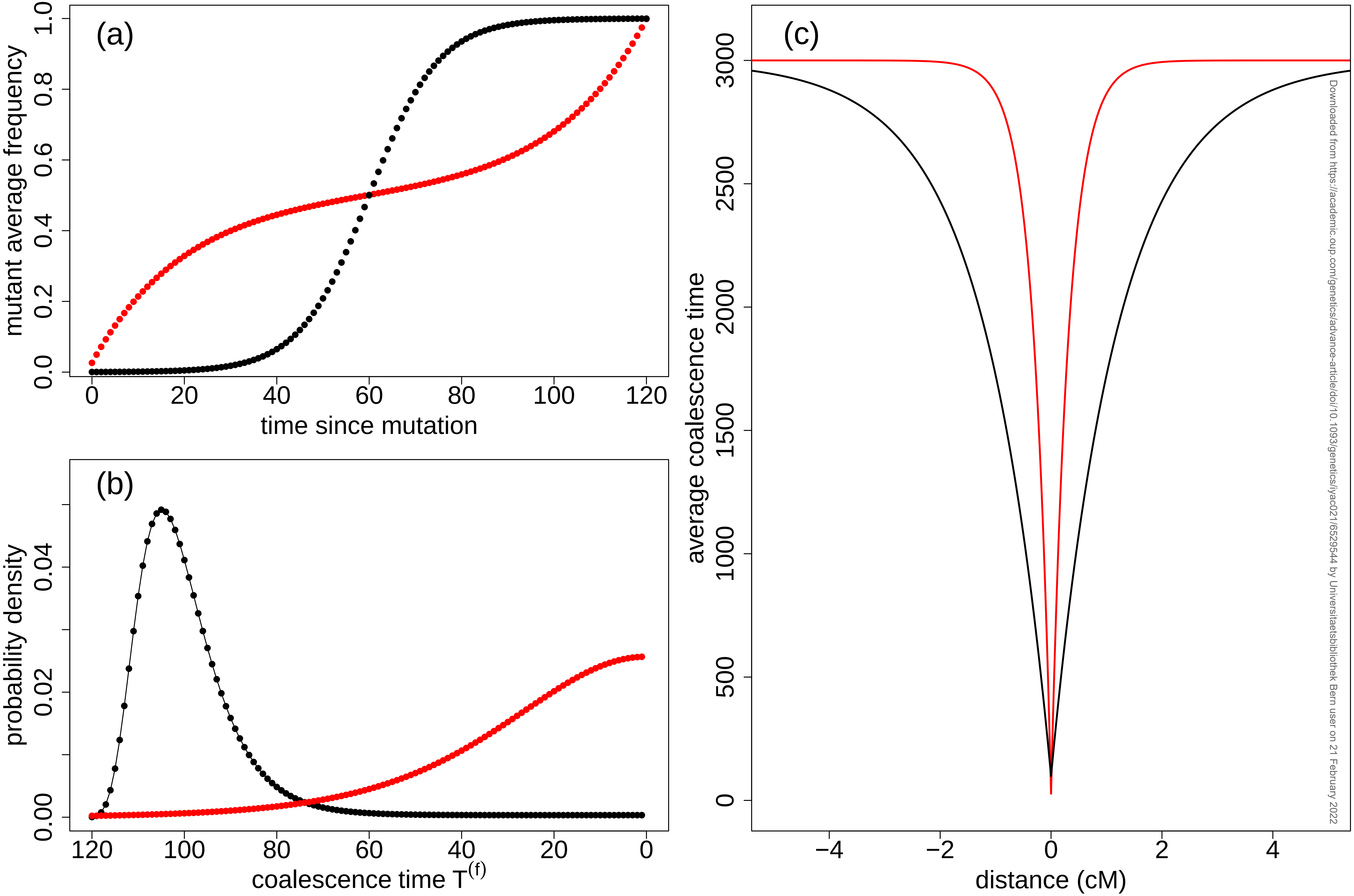
## A5. Site frequency spectrum around a neutral trough compared to a selective trough

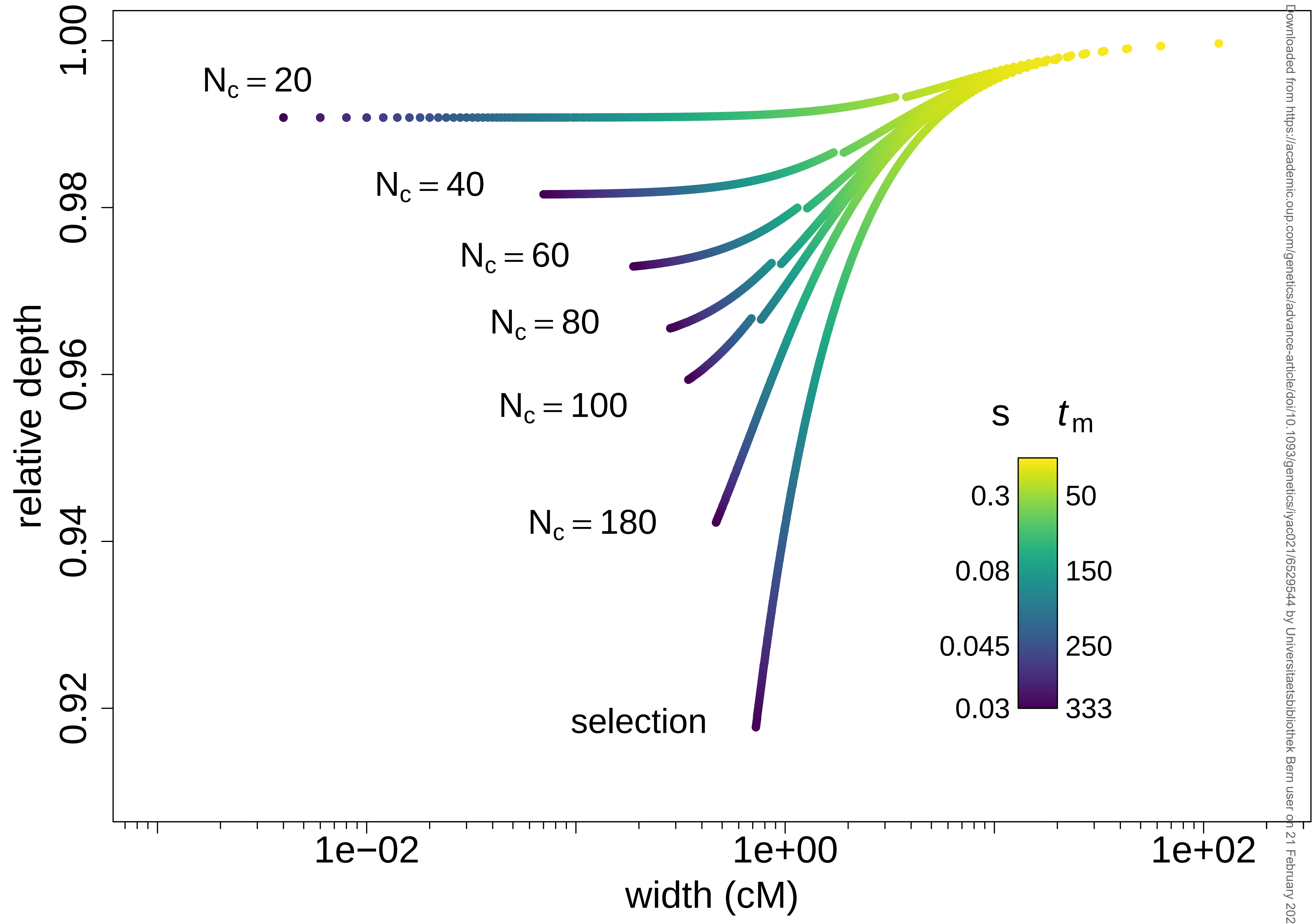

(a)
nucleotide diversity π
0.006
4N0 μ
0.004
0.002
4Nc μ
0
0.25
0.5
0.75
1
1.25
time since contraction (in 2Nc generations)
(b)
nucleotide diversity π
0.006
4N0 μ
0.004
0.002
4Nc μ
tc = 0
0
5
10
15
20
pos. on the chromosome (in Mb)
(c)
tc = 0.25
nucleotide diversity π
0.006
4N0 μ
0.004
0.002
4Nc μ
0
5
10
15
20
pos. on the chromosome (in Mb)
(d)
tc = 0.75
nucleotide diversity π
0.006
4N0 μ
0.004
0.002
4Nc μ
0
5
10
15
20
pos. on the chromosome (in Mb)


$T^{(l)}$
$T^{(f)}$
haplotype i
haplotype j
N(t)
$N_0$
$N_c$
focal locus (f)
linked locus (l)
time
$t_c$
$t_m$
t = 0


$t_m = 20$
$t_m = 40$
$t_m = 80$
$t_m = 120$
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
number of mutant copies
0
10
20
30
40
time since mutation
0
5
10
15
20
0
10
20
30
40
0
20
40
60
80
0
30
60
90
120
distr. of coal. times $P(T^{(f)})$
0.00
0.05
0.10
0.15
0.20
0.00
0.04
0.08
0.00
0.02
0.04
0.000
0.010
0.020
0.030
$T^{(f)}$
20
15
10
5
0
40
30
20
10
0
80
60
40
20
0
120
90
60
30
0


(a)
average coalescence time
0
500
1000
1500
2000
−100
−50
0
50
100
distance (cM)
(b)
average coalescence time
0
10
20
30
40
50
60
−0.10
−0.05
0.00
0.05
0.10
distance (cM)


(a)
mutant average frequency
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
120
time since mutation
(b)
probability density
0.00
0.02
0.04
120
100
80
60
40
20
0
coalescence time $T^{(f)}$
(c)
average coalescence time
0
500
1000
1500
2000
2500
3000
−4
−2
0
2
4
distance (cM)


1.00
0.98
0.96
0.94
0.92
relative depth
$N_c = 20$
$N_c = 40$
$N_c = 60$
$N_c = 80$
$N_c = 100$
$N_c = 180$
selection
s
$t_m$
0.3
0.08
0.045
0.03
50
150
250
333
1e−02
1e+00
1e+02
width (cM)