Weak genetic draft and Lewontin's paradox

G Achaz[1,2,*] and E Schertzer[3,*]

1: SMILE group, Center for Interdisciplinary Research in Biology, Collège de France, Paris, France

2: Université Paris-Cité, Paris, France

3: Dynamical Systems in Biomathematics, University of Vienna, Austria

*: Both authors equally contributed to this work

July 2025

1 Abstract

Neutral theory assumes that in a population of size N, diversity results from an equilibrium between new mutations arising at rate μ and genetic drift that purge them at rate $1/N$, predicting an equilibrium value proportional to $N\mu$. The difference between this expectation and the much lower observed molecular diversity is known as Lewontin's paradox of variation. Here, we investigate the effect of genetic draft, a regime of evolution where recurrent sparse selective sweeps entirely drive the diversity of surrounding loci. More specifically, we focus on the neglected distant effect of selective sweeps on remote neutral loci, where the effect of a single sweep is almost negligible. We derived novel mathematical approximations of this underexplored regime and show that under *weak genetic draft*, diversity at neutral loci is a power law of the population size: $N_e^{(\pi)} \propto N^{2A}$, for $A < 0.5$, where A is the ratio between recombination rate and coefficient of selection ($A = c/s$). Interestingly the Site Frequency Spectrum at neutral loci is identical to the one produced by genetic drift, as the underlying coalescent tree is an n-Kingman coalescent. In brief, weak genetic draft produces patterns of diversity that look entirely neutral, while being drastically reduced in magnitude. Ultimately, our study points to the need to explore evolutionary models for which diversity looks neutral but does not scale linearly with population size.

2 Introduction

The *neutral theory of molecular evolution* (Kimura, 1983) has become the reference model in molecular evolution since its first proposal at the end of the sixties (Kimura, 1968; King and Jukes, 1969). Under this model, the vast majority of the segregating polymorphisms results from a balance between incoming neutral mutations, at rate μ per individual per generation, and genetic drift purging them at the scale of N generations (Wright, 1931). To picture the effect of genetic drift, it is convenient to consider the ancestry of individuals backward in time (Malécot, 1948). This reversal in time led to what has since been named *coalescent theory* describing the genealogy of individuals sampled in the population (Kingman, 1982; Hein *et al.*, 2005; Wakeley, 2009). At a given locus, the most recent common ancestor of the whole present-day population lived on the order of N generations ago. Although the mean is exactly $2N$ for a haploid Wright-Fisher model where generations are non-overlapping, it reduces to N "generations" (N births and N deaths) for a haploid Moran model, where the replacement of individuals is an overlapping continuous process. Extending the Wright-Fisher model to N panmictic diploid individuals of two sexes, the time scale of genetic drift is unchanged ($4N$ generations as there are $2N$ haploid genomes), provided there is a equal number of males and females (Wright, 1931).

Consequently, in a standard neutral model (a constant-size panmictic neutral population, SNM hereafter), the expected molecular diversity is typically proportional to $N\mu$, regardless of the precise measure of molecular diversity (number of alleles, number of polymorphic sites, pairwise difference, etc.). Thus, under the SNM,

molecular diversity increases linearly with population size. When μ is known, one can infer N from the observed molecular diversity. This inferred value is necessarily linear with the amount of observed molecular diversity and is one of the many estimates of what is known as the *effective population size* N_e.

The concept of effective population size was originally proposed by Wright (1931), and it has become a central concept in population genetics. It was designed so that its inverse ($1/N_e$) measures the rate of genetic drift. Indeed, for more realistic populations, the population census size N is not the relevant parameter tuning genetic drift. As a simple illustration, in an aged structure of a population, only adult individuals contribute to the next generation, so that the rate of genetic drift only depends on the size of this age class and not on the total number of individuals. Several scenarios will result in important differences between N and N_e (Charlesworth, 2009; Waples, 2022): large demographic variations, recent foundation time of the species following recurrent species fragmentation, strong structuration in local demes, presence of linked selection, sex ratios, large variance of offspring's number, etc.

Finer patterns of genetic diversity can help unraveling the evolutionary processes at work in living species. First, between species, the apparent constancy of the rate of accumulation of molecular divergence between species, known as the molecular clock (Zuckerkandl, 1962), was used a pivotal argument in favor of the neutral theory (Kimura, 1968; Kimura and Ohta, 1971) and then later in favor the nearly neutral theory (Ohta and Kimura, 1971; OHTA, 1973). Interestingly, positive selection will not distort the molecular clock at the linked neutral polymorphisms as beneficial mutations that will fix have a uniform probability $1/N$ to occur in one genome of the population.

Second, within species, the distribution of frequencies of alleles at polymorphic sites, the *Site Frequency Spectrum* (SFS), provide a richer information than simple scalar metrics of diversity. As the frequency of a given allele depends on the branch of the complete coalescent tree where the mutation occurred, the SFS is a direct output of the probabilistic distribution of the coalescent trees. Under the SNM, the coalescent trees follows a Kingman distribution (Kingman, 1982) and the expected SFS is a hyperbolic function (Fu, 1995), for which alleles that are rare in the population (low frequency) are more abundant in the genome alignment than frequent alleles. The SFS is routinely used to assess the goodness-of-fit of the SNM using a large variety of neutrality tests (Achaz, 2009).

Soon after the advent of the neutral theory, Lewontin (1974) pointed out that the range of molecular diversity observed across species is much narrower than their range of population size, as if population size was a poor predictor of diversity. This observation has been since coined *Lewontin's paradox of variation* (Leffler *et al.*, 2012; Ellegren and Galtier, 2016; Galtier and Rousselle, 2020; Charlesworth and Jensen, 2022). In fact, most estimates of effective population size based on diversity are typically much smaller than what is expected under the SNM. It was nonetheless reported that few life history traits are good predictors of diversity. For example, species with small propagule size, that is likely inversely proportional to population size, harbor higher diversities (Romiguier *et al.*, 2014). More recently, using a data set of

172 species, a strong statistical support was reported in favor of a positive log-log linear relationship (*i.e.* a power law) between census size and diversity (Buffalo, 2021).

As neutral diversity scales with the coalescent times, also known as $N_e^{(\pi)}$, Lewontin's paradox can be recast as an enigma on the time scale of genetic ancestry. Hence, the main question raised in the present study is whether one can explain the difference in the order of magnitude between $N_e^{(\pi)}$ and N. Non-constant demography, population structure or linked selection can result in $N_e^{(\pi)}$ values that can be very different from N. However, most deviations of the SNM that were proposed to explain the reduction in diversity also results in strong distortions of other patterns of diversity such as the SFS. In addition, the previous explanations often predict that the effective population size would be either on the order of N (e.g. age structure, sex ratio, see Wright (1931)), or almost independent of N (e.g. strong linked selection, very large demographic fluctuations, see Coop and Ralph (2012)).

In this article, we show that weak linkage to a constant flow of beneficial mutations, a regime that we named *weak genetic draft*, solves the problem of coalescent time scale by predicting a power law between $N_e^{(\pi)}$ and N, without altering the neutral pattern of diversity at the SFS.

Linked selection is the influence of a nearby selected allele on a focal neutral locus. Both positive linked selection, named *hitchhiking* (Maynard Smith and Haigh, 1974), and negative linked selection, named *background selection* (Charlesworth *et al.*, 1993) reduces diversity in surrounding loci. The strongest empirical support for the reduction of diversity through linked selection stems from the observation that regions with low recombination show less diversity than regions with high recombination (Aguade *et al.*, 1989; Stephan and Langley, 1989; Begun and Aquadro, 1992). Indeed, both positive and negative regimes of linked selection reduce drastically surrounding diversity for tightly linked loci. Interestingly, this effect is stronger for abundant species (Corbett-Detig *et al.*, 2015), suggesting that for large N, linked selection is more intense, as is selection more generally.

In the hitchhiking model, a beneficial mutant of selective coefficient s appears at a selected locus located at a (genetic) map distance c of a neutral locus. Map distance is evaluated by the recombination rate c. The conclusion of Maynard Smith and Haigh (1974) is that diversity at the neutral locus is strongly lowered by the selective sweep when the coefficient $A = c/s$ is small: selection is strong (s large), recombination is low ($c \to 0$) or both.

When $A = 0$, there is no recombination during the sweep and diversity at the neutral locus equals diversity at the selected locus. For (very) strong sweep, the entire region is purged from diversity. At the opposite, when $A \gg 1$, the two loci are completely independent and the neutral locus is unaffected by the selective sweep. Although the original derivation of the model assumed an infinite population size and was mostly focused on the reduction of heterozygosity, subsequent works refined the results and eventually rephrase it in a diffusion setting (Wiehe and Stephan, 1993) or in a coalescent framework (Kaplan *et al.*, 1989; Fay and Wu, 2000). The effect hitchhiking has further been connected to the effect of inflation of the additive variance in fitness

(Barton, 2000), originally reported by Robertson (Robertson, 1961). Broader scenarios of hitchhiking effects were studied in fluctuating environment, spatial structure and soft sweeps (Barton, 2000; Barton and Etheridge, 2004; Hermisson and Pennings, 2005; Pennings and Hermisson, 2006a,b).

Under the regime of genetic draft (Gillespie, 2000a,b), it is the repeated occurrence of beneficial mutations at the selected locus that dictates the coalescence process at the neutral locus. Although the original derivation was done for two lineages and infinite population size (Gillespie, 2000a,b), several approximations were derived for finite population and n lineages (Durrett and Schweinsberg, 2004; Etheridge *et al.*, 2006; Coop and Ralph, 2012). As this effect of hitchhiking decreases with A, most of the previous work have focused on the strong selection, low-recombination regimes ($A \to 0$). At the limit, the genealogy of a sample of n individuals is well described by a Multiple Merger Coalescent (MMC) process (Pitman, 1999; Sagitov, 1999; Tellier and Lemaire, 2014; Berestycki, 2009). In MMC, several lineages (the ones that are captured by the selective sweep) coalesce simultaneously once the time has been properly rescaled, assuming the time scale of a sweep is much smaller than the time elapsed between two consecutive sweeps. Although the probability of coalescence in most approximations (Wiehe and Stephan, 1993; Barton, 1998, 2000; Durrett and Schweinsberg, 2004; Etheridge *et al.*, 2006) is a power law of N ($P_{\text{coal}} \propto N^{-2A}$), it becomes independent of N at the limit of $A \to 0$. This has led to a view where the coalescent process is entirely driven by the frequency of the selective sweeps and does not dependent anymore on the population size (Coop and Ralph, 2012).

Here, we show that contrarily to the regime $A \ll 1$ alluded in the previous paragraph, larger values of A generates a genealogy consistent with SNM (Kingman tree) but on drastically reduced time scale $N_e \propto N^{2A}$. Thus, this regime – that we call *weak genetic draft* – leaves the qualitative predictions of neutral theory unchanged (SFS) while predicting a relation between effective population size and census size consistent with previous studies (Buffalo, 2021).

To investigate the weak selection draft analytically, we first rederive a semi-deterministic model (Martin and Lambert, 2015) with Random Initial and Final frequencies, that we named *RIF model*, which approximates very well selective sweeps in finite population. Using a backward version of the RIF model, we derive a novel accurate approximation for the coalescent probability at the linked neutral locus for 2 lineages. We then show how the RIF model can be used to compute analytically or simulate very efficiently the effect of genetic draft. We further derived a second approximation using a diffusion approximation that is simpler to evaluate numerically and has also an excellent agreement with simulations. We finally numerically show how weak genetic draft generates a neutral Kingman tree, with a reduced time scale. We end by discussing the parameter space on which the model holds and how it fits observed patterns of molecular diversity.

We shall start by recapitulating notations, defining the strength of a selective sweep and provides a basic overview of the model. Most of the previous works have mainly investigated the regime where recombination is relatively small compared to selection. This regime is characterized by a small recombination to selection

ratio.

$$\textbf{Strong hitchhiking:} \quad 1/N \ll c, 1/N \ll s \ll 1 \quad \text{(moderate selection-recombination)}$$
$$\text{and} \quad A = \frac{c}{s} = O\left(\frac{1}{\log(s)}\right) \ll 1.$$

Under this scenario, the ratio A is small (selection is much stronger than recombination). As a consequence, the hitchhiking is *strong* in the sense that fixation of the advantageous selected allele has a sizeable effect on the genetic diversity at the strongly linked neutral locus. After completion of the sweep, a positive fraction of the population inherits the genetic material from the mutant type at the neutral locus. In the absence of new mutation at the neutral locus, the heterozygosity is entirely lost after a few sweeps.

In contrast with the aformentioned works, this article will focus on a different regime. Namely,

$$\textbf{Weak hitchhiking:} \quad 1/N \ll c, 1/N \ll s \ll 1 \quad \text{(moderate selection-recombination)}$$
$$\text{but} \quad A = \frac{c}{s} = O(1).$$

In this regime, recombination is high enough so that, *at first approximation*, the evolution of the genetic diversity at the neutral and the selected sites are asymptotically independent along the course of a single sweep. The genetic diversity at the neutral locus is then almost insensitive to the fixation of a newly introduced mutant type.

Although the effect of a single sweep is negligible, our aim is to show that when the sweeps are repeated at a high enough frequency (more frequent than the drift scale), the infinitesimal effect of weak sweeps accumulates and have a sizeable aggregated effect, more important than genetic drift, for large populations $N \to \infty$. In order to illustrate this idea, we assume the environment changes at constant rate λ and that each triggers the sweeps of new beneficial mutations at the locus under selection. Mutated loci have a $(1+s)$ relative fitness benefit over the resident background. Mutations are triggered by environmental changes.

Under the scenario of frequent sweeps, we analyze properties of the coalescent tree at the neutral locus, using the structured coalescent framework proposed by Hudson and Kaplan (1988). We show that for large enough N and A values (large population-weak hitchhiking), the genealogical structure at the neutral locus becomes indistinguishable from a neutral genealogy. More formally, the coalescent tree converges to a Kingman tree (neutral genealogy), except for its reduced time scale.

3 Model derivation and Results

3.1 The fixation trajectory and its approximation

3.1.1 The forward fixation of a beneficial allele

We build our approach on a diffusion approximation of a haploid population of constant size N, in which a derived allele at frequency $\bar{Z}_{t^\star}$ benefits from a relative fitness $(1+s)$ compared to the ancestral alleles of

relative fitness 1. The trajectory of this beneficial derived allele is modeled by a Wright-Fisher diffusion with selection, provided that usual diffusion conditions hold ($N \gg 1$ and $s \ll 1$), for which of one time unit of t^* corresponds to 1 discrete Wright-Fisher generation

$$d\bar{Z}_{t^*} \;=\; s\bar{Z}_{t^*}(1-\bar{Z}_{t^*})dt \;+\; \sqrt{\frac{1}{N}\bar{Z}_{t^*}(1-\bar{Z}_{t^*})}dW_{t^*}. \tag{3.1}$$

To model a fixation trajectory, we condition the latter diffusion on fixation and denote by Z the resulting process. To construct the conditioned process, we proceed in two steps (1) we start from $\bar{Z}_0 \in (0,1)$ and we follow the frequency of a fixing allele by first conditioning (3.1) on survival, and (2) we let the initial frequency go to 0 ($\bar{Z}_0 \downarrow 0$). Note that although the initial frequency of a Wright-Fisher model is $1/N$, the resulting diffusion approximation starts from 0 for a large population. The resulting conditioned process Z can be shown to be itself a diffusion obtained from the unconditioned diffusion through a Doob h-transform (Etheridge *et al.*, 2006):

$$dZ_{t^*} = sZ_{t^*}(1-Z_{t^*})\coth(NsZ_{t^*})dt \;+\; \sqrt{\frac{1}{N}Z_{t^*}(1-Z_{t^*})}dW_{t^*}, \quad Z_0 \;=\; 0. \tag{3.2}$$

(Note that we remove the bar to indicate the conditioning process.) The effect of the conditioning is to multiply the first term by $\coth(NsZ_{t^*})$. Since

$$\lim_{x\downarrow 0} x(1-x)\coth(Nsx) = 1/Ns$$

Z is repulsed away from 0 at small frequency. The process thus starts from 0 and is repulsed away from the origin so that it never returns to 0 on later times. *

From now on, we will assume a moderate selection regime where

$$\frac{1}{N} \ll s \ll 1. \tag{3.3}$$

Under this regime, it is convenient to accelerate time by a factor $1/s$ and thus introduce a rescaled process $z_t := Z_{t^*/s}$. In this new time scale t represents $1/s$ units of t^* (that were Wright-Fisher generations). A direct computation (dividing both sides of eq. (3.2) by s) yields to:

$$dz_t = \qquad z_t(1-z_t)\coth(Nsz_t)dt \;+\; \sqrt{\frac{1}{Ns}z_t(1-z_t)}dw_t, \quad z_0 \;=\; 0. \tag{3.4}$$

where $(w_t; t \geq 0) = (\sqrt{s}\; W_{t^*/s}; t^* \geq 0)$ is distributed as a standard Brownian motion.

3.1.2 The RIF approximation: a logistic equation with random initial and final frequencies

The aim of the present section is to show that the diffusive selective sweep (3.4) is well approximated by a deterministic logistic equation where the initial frequency and the final frequency are characterized by

*On a more formal level, the previous observation can be made rigorous by proving that 0 is an entrance boundary (Etheridge, 2011) for the process Z_t, i.e., there exists a solution starting at 0 and remaining strictly positive afterwards.

independent exponential random variables with mean $1/(2Ns)$. We name this approximation of the fixation trajectory the RIF-model, for Random Initial and Final conditions. This model has been previously described in the literature (Martin and Lambert, 2015) under the name of semi-deterministic. In their work, the authors derive this approximation starting from a Moran-like model. In this manuscript, we provide rapid insights on its justification in the diffusion setting.

The idea of the RIF approximation relies on the classical observation that the fixation trajectory is almost deterministic at intermediary frequency, and that this deterministic regime is interpolated by two stochastic phases – respectively at low and high frequency of the mutant allele – which are well approximated by branching processes.

Let us now spell out the argument in more details. When z_t is at intermediate frequency, since $Ns \gg 1$ (3.4) is well approximated by the deterministic logistic equation (time t is still in $1/s$ generations):

$$\frac{dx_t}{dt} = x_t(1-x_t). \tag{3.5}$$

To fully determine the deterministic dynamics, it now remains to determine the initial condition of the system. The idea is to observe the process at small frequency (i.e., when $t = o(1)$) and to observe that under (3.5)

$$x_t \approx x_0 e^t \tag{3.6}$$

so that it remains to identify x_0 such that the latter approximation holds at small times for the diffusion equation.

Set $y_t := Nsz_t$. At low frequency (*i.e.* at small time t such that $z_t \approx 0$), (3.4) writes

$$\begin{aligned} dy_t &= y_t(1-z_t)\coth(y_t)dt + \sqrt{y_t(1-z_t)}dw_t \\ &\approx y_t \coth(y_t)dt + \sqrt{y_t}dw_t. \end{aligned} \tag{3.7}$$

The latter approximation is identical in law to a supercritical continuous-state branching process (Feller diffusion, see e.g. Lambert (2008)) conditioned on eternal survival †. In the fixed size Wright-Fisher populations with directional selection, this Feller diffusion approximation translates into approximating the advantageous alleles at low frequency by a supercritical branching process (a classical approach used by Fisher and Haldane, see review by Patwa and Wahl (2008)) conditioned on survival.

A key property (known as the Yaglom's law) of the process y (described by the approximation (3.8)) is that it converges to a deterministic exponential growth with a random initial condition

$$\lim_{t\to\infty} y_t e^{-t} = \frac{\mathcal{E}}{2} \text{ almost surely,} \tag{3.8}$$

where $\mathcal{E}$ is a standard exponential random variable (of mean 1). Since $z_t = \frac{y_t}{2Ns}$, we can informally write:

$$z_t \approx \frac{\mathcal{E}}{2Ns}e^t, \quad \text{when } z_t \ll 1. \tag{3.9}$$

†Analogously to our discussion on the conditioned Wright-Fisher diffusion in Section 3.1.1 and the distinction between Z and $\bar{Z}$, the unconditioned Feller diffusion of the $\bar{y}_t$ is given by: $d\bar{y}_t = \bar{y}_t dt + \sqrt{\bar{y}_t}dw_t$. As for the Wright Fisher diffusion, the conditioning has the effect of changing the first term $\bar{y}_t$ into $y_t \coth(y_t)$.

Comparing this estimate with (3.6), we conclude that the process z observed at intermediary frequency is well approximated by the deterministic solution x started from the initial condition

$$x_0 = \frac{\mathcal{E}}{2Ns}$$

Finally, it has been noted by Coop and Griffiths (2004) that the Wright-Fisher fixation trajectory z is reversible in the sense that if τ denotes the fixation time then

$$(z_t; t \in [0, \tau]) \text{ and the reversed process } (\hat{z}_t := 1 - z_{\tau - t}; t \in [0, \tau]) \text{ are identical in law.} \tag{3.10}$$

By time reversal, this shows that the fixation time is well approximated by stopping the deterministic trajectory at

$$x_\tau = 1 - \frac{\mathcal{E}'}{2Ns}$$

where $\mathcal{E}'$ is an independent standard exponential random variable.

To summarize, the previous heuristics show that one can approximate the initial stochastic phase by an instantaneous jump from frequency 0 to x_0 and the final stochastic phase by an instantaneous jump from frequency x_τ to 1. The two phases are interpolated by a deterministic logistic phase. In the following, this approximation will be referred to as the RIF model.

3.1.3 Numerical evaluation of the RIF model

To assess the numerical precision of the RIF model, we compared individual based Wright-Fisher simulations of trajectories conditioned on fixation (using a naive rejection algorithm) with forward RIF-model realisations with random z_0 and z_τ (sampled by Monte-Carlo). We compute two distributions. First, we report the distribution of allele frequency at the rescaled time $t = \ln(Ns)$, where the trajectory has become deterministic (Figure 1, left panel). Second, we report the distribution of the fixation time (Figure 1, right panel).

In the RIF approximation, the rescaled fixation time of the trajectory is given by:

$$\tau = 2\ln(Ns) - \ln(\frac{1}{4}\mathcal{E}\mathcal{E}') \tag{3.11}$$

implying that the mean fixation time in generations is well approximated by

$$E[t^*_{\text{fix}}] = 2\ln(2Ns + \gamma_\varepsilon)/s \tag{3.12}$$

where γ_ε is the Euler constant ($\gamma_\varepsilon \approx 0.5772$).

Figure 1 shows that both distributions (derived allele frequency at time $\ln(Ns)/s$ and fixation time) display an excellent numerical agreement between both the Wright-Fisher simulations and the RIF-model realisations. They illustrate how the RIF-model constitutes an excellent approximation for selective sweeps in finite populations. Diffusion realisations based on eq. 3.2 also display excellent numerical agreement (data not shown).

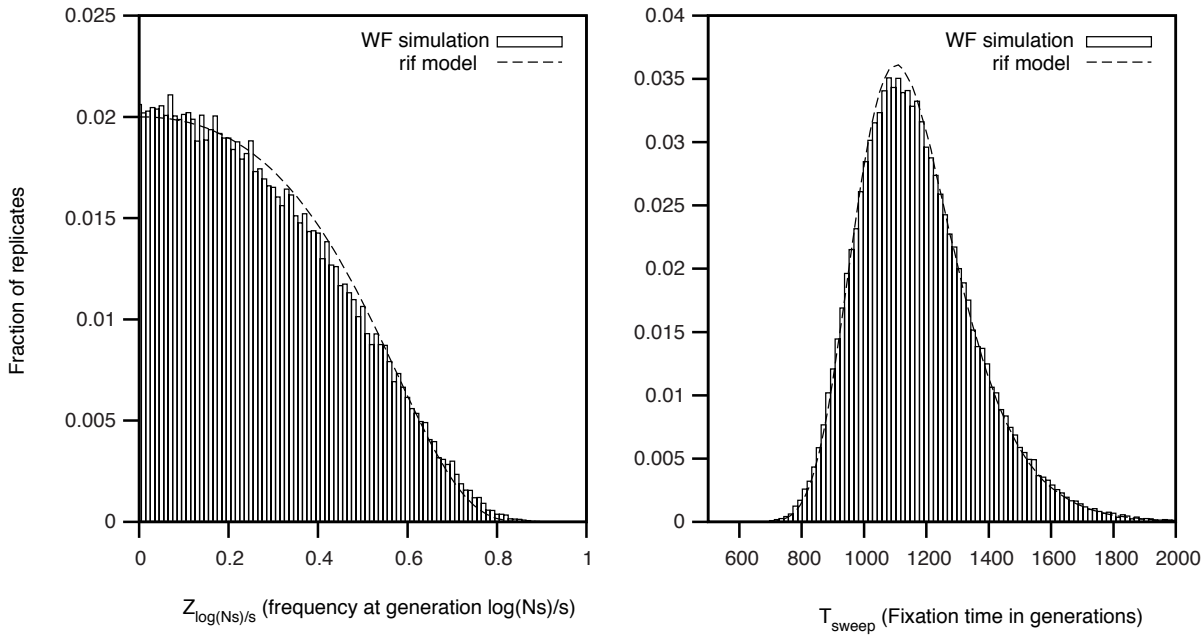

Figure 1: We compare the outputs of simulated selective sweeps in a Wright-Fisher model (discarding replicates where the beneficial allele is lost) to the prediction of the RIF model. Population size is $N = 10^5$ and the coefficient of selection is $s = 0.01$. On the left panel, we display the distribution (bins of 0.01) of frequency Z_{t^*} of the fixing beneficial allele at the generation $t^* = \ln(Ns)/s$. The RIF-model corresponds to eq. (3.9), with $\mathcal{E}$ generated by Monte-Carlo. On the right panel, we display the distribution (bins of 15 generations) of fixation time. The RIF-model corresponds to equation 3.11, with $\mathcal{E}$ and $\mathcal{E}'$ generated by Monte-Carlo.

We would like to mention that a single numerical realisation of the RIF model is instantaneous as it only relies on randomly drawing two exponential random variables (z_0 and z_t), whereas a forward realisation of a sweeping allele in a Wright-Fisher framework (only following frequencies) requires as many binomial samplings as the fixation time (plus few samplings for the runs where the beneficial allele is lost). To simulate using the diffusion approximation, conditioned on fixation (see eq. (3.2)), one draw as many normal random variables as time increments dt. Both Wright-Fisher and diffusion increase CPU time with population size since selective sweeps take more time to complete. As a crude comparison, the realisation of 10^5 replicates of selective sweeps ($N = 10^5$ and $s = 0.01$ as in Figure 1) took an a laptot: 0.3" for the RIF model, 2'45" for diffusion with $dt = 1$ and 4'01" for Wright-Fisher with rejections of losses. This makes the RIF approximation 10^3 times faster than a naive Wright-Fisher version. In this regard, the RIF model can be preferred in *most* numerical simulations of strong selective sweep ($Ns \gg 1$), both forward or backward (see below).

3.2 Coalescence probability at a linked neutral locus during a sweep

3.2.1 Model assumptions

We now turn on describing the backward coalescent process that is associated with a forward selective sweep in finite population. We recall that we considered two disctinct description of the sweep:

1. The diffusion approximation through the conditioned process z_t.
2. The RIF approximation x_t which is obtained as approximation of the rescaled process z under assumption (3.3).

Hereafter,we only describe the backward coalescent process for the RIF model, but an analog description for the diffusion model consists of replacing the RIF dynamics x_t by the conditioned diffusion process z_t in the upcoming discussion.

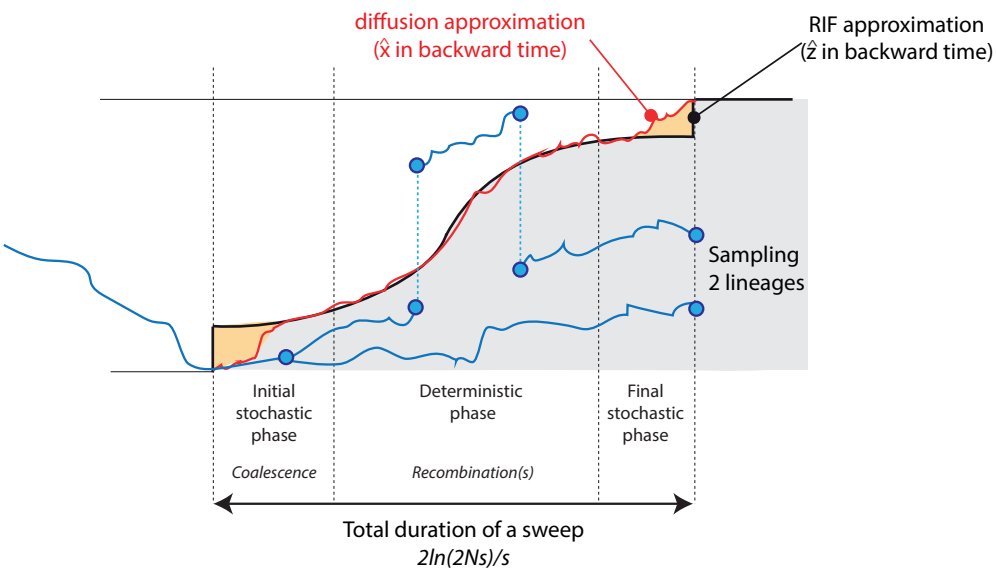

Figure 2: **A model of structured coalescent**. In this sketch of the model, we represent a single event of selective sweep that lasts on average $\approx 2\ln(2Ns)/s$. Two lineages at a linked neutral locus (represented in blue) are sampled after the completion of the selective sweep. On their journey backward in time, the lineages start linked to the beneficial allele at the selected locus (they are in the gray area). They can nonetheless recombine back and forth during the selective event changing their association to the ancestral allele or back to the beneficial allele. Although coalescence could in principle occur anytime during the whole sweep, it occurs in practice only at the onset of the selective sweep (as justified in the appendix). We also illustrate the two approximations used to derived the probability of coalescence: the RIF approximation (in black) and a diffusion approximation (in red).

As time is now counted backward, we define a reverse trajectory into the past noted $\hat{x}_t = 1 - x_{\tau-t}$. Since time is reversed, $\hat{x}$ starts from the initial condition $\hat{x}_0 = 1 - x_\tau$, where $x_\tau = \mathcal{E}'/2Ns$ in the RIF model ($\hat{z}_0 = 0$ in the diffusion framework). $\hat{x}$ terminates at $\hat{x}_\tau = 1 - x_0$, where $x_0 = \frac{\mathcal{E}}{2Ns}$ ($\hat{z}$ terminates at 1). Recall that time is rescaled by $1/s$ (see Section 3.1.1 and eq. 3.4), so that the recombination rate becomes $A = c/s$. Given the reversed fixation trajectory $\hat{x}_t$, we follow the construction of Hudson and Kaplan (1988) who model the genealogy at the neutral locus as a structured coalescent process. For an schematic view of this structured coalescent is provided in Figure 2.

1. Lineages can be of type 0 (linked to the beneficial allele – grey zone on Fig. 2) or type 1 (linked to the ancestral allele). The two lineages start as type 0.

2. Lineages transition from type 0 to type 1 at effective recombination rate $A\hat{x}_t$ and from type 1 to type 0 at effective rate $A(1 - \hat{x}_t)$.

3. Coalescence can only occur within lineages of the same type: pairs of type 0 coalesce at rate $1/Ns(1 - \hat{x}_t)$ and pairs of type 1 coalesce at rate $1/Ns\hat{x}_t$. Coalescence rates are multiplied by $1/s$ since time has been rescaled by the same amount.

4. In addition, we also assume that if two lineages are of type 0 at time τ, the two lineages trace back to the ancestral mutant and are forced to coalesce (this is especially relevant for the RIF approximation).

3.2.2 Two new approximations

We have derived two new approximations for the probability of coalescence of two neutral lineages that are linked to the selected locus. We first derived an approximation under the RIF model (that requires to evaluate numerically integrals, see eq. 3.13) and then derived a second approximation that is numerically simpler to compute (see eq. 3.14). Details of the derivation are provided in the Appendix.

$$\mathbb{P}^{\text{RIF}}_{\text{coal2}} \approx \frac{1}{(2Ns)^{2A}} \left(\frac{\pi A}{\sin(\pi A)}\right)^2 \left[2^{2A} \int_0^\infty \gamma(1-2A, \frac{2}{x}) e^{-x} dx \; + \; \int_0^\infty x^{2A} e^{-x-\frac{2}{x}} dx\right] \tag{3.13}$$

$$\mathbb{P}^{\text{diff}}_{\text{coal2}} \approx \frac{1}{(2Ns)^{2A}} \left(\frac{\pi A}{\sin(\pi A)}\right)^2 \frac{\Gamma(2(A+1))}{1-2A} \tag{3.14}$$

These two approximations both contain 3 elements which reflect the decomposition of the sweep as indicated on Fig. 2. The first element is $((2Ns)^{-2A})$ and corresponds to the average selective duration ($\approx 2 \ln 2Ns/s$) during which lineages should not recombine ($e^{-\tau/s}$) and provides the global time scale.

The other elements come from the fine description of the lineages during a sweep. During the "deterministic phase" of the sweep (see Fig. 2), lineages can recombine and the second element – given by $(\pi A/\sin(\pi A))^2$ – corresponds to the probability that both lineages are of type 0 at the end of the deterministic phase (first vertical dashed line from the left in Fig. 2.) The difference between the two approximations account for

different behavior of the two processes in the "initial stochastic" phase, that is the phase where the original fluctuations of the process z matter. Despite the apparent difference, both approximations exhibit very similar numerical values (see Figure 3).

We then compare the numerical values of both approximations to forward Wright-Fisher and backward RIF Monte-Carlo simulations as well as to approximations previously derived in the literature in the next section.

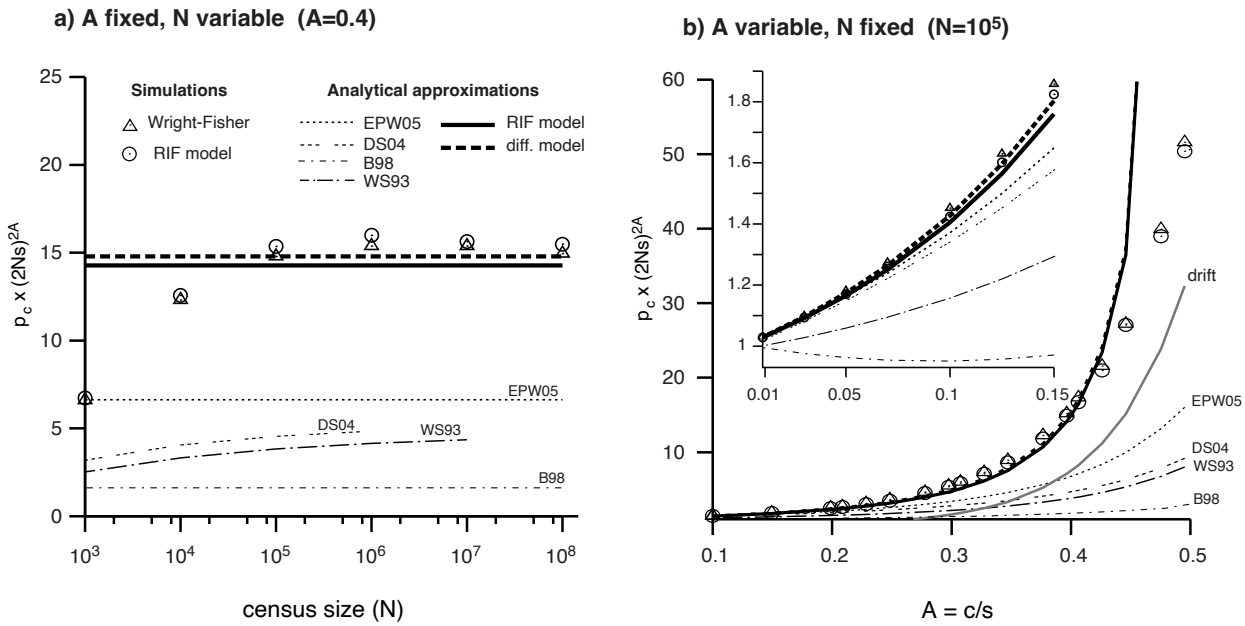

Figure 3: **Numerical evaluation of coalescence probability**. We compare the outputs of two different simulation procedures with six analytical approximations of $p_c \times (2Ns)^{2A}$ the probability of coalescence deprived of the first term. For all cases, the selection coefficient of the sweeping allele is $s = 0.01$. The 'Wright-Fisher' simulations are discrete non-overlapping generations of a Wright-Fisher model with selection ignoring cases where the beneficial allele is lost. The 'RIF model' is a deterministic sweep with random initial and final frequencies. WS93, B98, DS04, EPW05 are the analytical approximations provided respectively in Wiehe and Stephan (1993); Barton (1998); Etheridge *et al.* (2006); Durrett and Schweinsberg (2004). W93 and DS04 could not be computed for large N. B98 corresponds to the eq7 of the cited article. The two other expressions are derived in the present work ('RIF model' is eq. (3.13) and 'diff model' is eq. (3.14). a) On this panel, the ratio $A = c/s$ is fixed to 0.4 and the population size varies from 10^3 to 10^8. b) On this panel, the population size N is fixed to 10^5 and the A ratio varies from 0.01 to 0.5. The insert is a zoom for small A values. The "drift" curve is computed using the approximate coalescence probability by drift during the sweep: $p_c = 1 - (1 - \frac{1}{N})^{\mathbb{E}[\tau/s]}$.

3.2.3 Comparison between the approximations and the simulations.

We will now compare different analytical approximations for the coalescence probability under a single selective sweep. We consider the two expressions derived earlier (under the RIF and the diffusion model) together with previous work of Wiehe and Stephan (1993), Barton (1998) (we used eq.7, as the "exact solution" of eq.13 was numerically too difficult to evaluate), Durrett and Schweinsberg (2004) and Etheridge *et al.* (2006).

$$\mathbb{P}_{\text{coal2}}^{\text{WS93}} \approx 1 - \frac{2A}{(2Ns)^{2A}} \int_{1/2Ns}^{\infty} t^{1-2A} e^{-t} dt$$

$$\mathbb{P}_{\text{coal2}}^{\text{B98}} \approx \frac{1}{(2Ns)^{2A}} \Gamma(1+A)^2 \Gamma(1-A)^2 \Gamma(1+2A)$$

$$\mathbb{P}_{\text{coal2}}^{\text{DS04}} \approx \prod_{l=2}^{\lfloor 2Ns \rfloor} (1 - \frac{2A}{l-1}) + \sum_{i=2}^{\lfloor 2Ns \rfloor} \frac{2A}{i(i+1)} \prod_{l=i+1}^{\lfloor 2Ns \rfloor} (1 - \frac{2A}{l+1})$$

$$\mathbb{P}_{\text{coal2}}^{\text{EPW05}} \approx \frac{1}{(2Ns)^{2A}} \sum_{f=2}^{\infty} \frac{2}{f(f+1)} \exp\left(2A \times (\sum_{i=1}^{f-1} \frac{1}{i} - \gamma_\varepsilon)\right)$$

All those analytical expressions are compared with two types of Monte-Carlo simulations. The first one is a Wright-Fisher forward-then-backward (like in e.g. msms (Ewing and Hermisson, 2010)), where the fixation trajectory is simulated forward and a structured coalescent is then simulated backward conditioned by this trajectory as in Simonsen *et al.* (1995). The second is a simple structured coalescent assuming a backward RIF-model realization (randomly generating x_0 and x_τ), that is the RIF model but without analytical approximation.

In Figure 3, we report results for either A fixed (panel A) or N fixed (panel B). We first see that the two types of Monte-Carlo simulations are numerically almost indistinguishable, again pointing to the usefulness of using a CPU efficient RIF-model, here for coalescent numerical simulations.

We further compare all formulas with the simulations. Although, our two approximations provide very close numerical values, the approximations in Wiehe and Stephan (1993); Durrett and Schweinsberg (2004); Barton (1998); Etheridge *et al.* (2006) are only satisfactory for low A values: typically $A < 0.1$. In particular, we see from panel B Figure 3 that all previous approximations predict a coalescence probability lower than the neutral case when A is large enough. In addition, our two approximations become better when N gets larger (Panel A of Figure 3) and when A is not too close to 0.5 (Panel B of Figure 3). Although A could take in principle any values in $[0, \infty]$, our approximations only hold when $A < 0.5$, and even more when A is not too close to 0.5, where the approximations tend to infinity.

In Figure 4, we report the probability of coalescence of two lineages as a function of $A = c/s$ from $A = 0.01$ that corresponds to a strong draft regime to $A = 5$ that is a drift regime.

Numerical results shows that the probability of coalescence is increased compared to pure genetic drift when the ratio A is below 1 (and not 0.5). Results also shows that our approximation of the coalescence

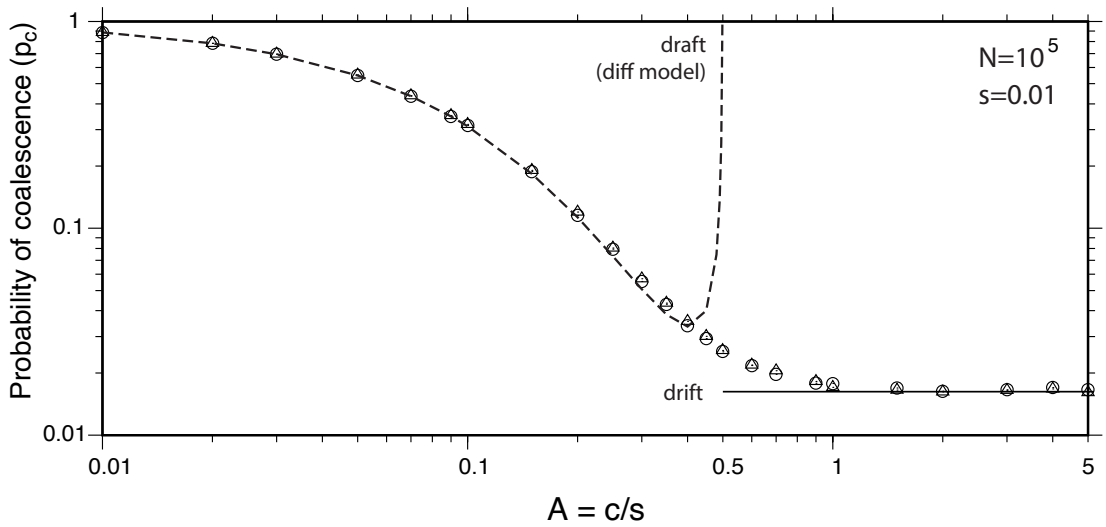

Figure 4: **From draft to drift**. The probability of coalescence is plotted as a function of the $A = c/s$ ratio, for a census size of $N = 10^5$ and a selective coefficient of $s = 0.01$. the dashed line represents the approximation derived from diffusion (eq. 3.14), whereas the drift line is $1 - (1 - \frac{1}{N})^{\mathbb{E}[\tau/s]}$. Like in Figure 3, circles are averages of forward Wright-Fisher simulations and triangles from backward RIF simulations (both with 10^6 replicates).

probability (eq. 3.14) is very good when $A < 0.4$ for a population size of 10^5 but becomes unreliable when $A \to 0.5$, where it is infinite. Extra work would thus be needed to analytically characterize the transition between genetic draft and genetic drift in the critical region surrounding $A = 0.5$.

3.3 Weak genetic draft

3.3.1 Effective population size

We now consider a population subject to recurrent selective sweeps. We assume, for the sake of simplicity, that each new beneficial allele appears at the same selected locus and benefits from the same coefficient of selection s over the resident allele. We shall further assume that successful selective sweeps occur at rate $\mathcal{S}$. Following (Kaplan *et al.*, 1989), one could consider two different cases, depending whether the adaptation process is mutation limited or not.

In the first case (mutation limited), the waiting time to the next successful selective sweep depends on the product of the number of new beneficial alleles per generation ($N\mu$, assuming a rate of beneficial μ) times its probability of fixation (well approximated by $2s$, when $s \ll 1$). This results in a rate of selective sweep of $\mathcal{S}_1 \approx 2N\mu s$ that is proportional to N. For any fixed value for s and μ, the rate of selective sweep will increase to ∞ with N and the expected time between two selective sweeps will converge to 0. We do not consider this case here, where selective sweeps would, at the limit, occurred all nested one into the other and

interfere. This would also result in a regime where N_e decreases as N increases as described in Weissman and Barton (2012).

Alternatively, we consider here another population model where the rate of selective sweep does not depend on the population size and where an external factor allows for a new beneficial mutation to arise. An example of such regime is evolving in a fitness landscape with a moving optimum. Here we model environment changes at rate λ and the fitness optimum is shifted by s. After each change of environment, the population adapts to the environmental shift by a selective sweep.

If the time of adaptation is small compared to the time scale of the environmental changes, $\lambda \ll 2N\mu s$ changes in the environment are almost instantaneously followed by a rapidly occurring selective sweep. In this setting, the sweep rate is then given by λ and is independent of N, a case that will always occur for sufficiently large values of N. In this second scenario, we have

$$\mathcal{S}_2 \approx \lambda$$

In the following, we will only focus on the case $\mathcal{S} \approx \lambda$. We monitor the fate of a neutral locus located at a map distance c of the selected locus. If we sample 2 individuals in the extant population under changing environment, we expect their coalescence to be driven by recurrent selective sweeps. We have shown in the previous section that the probability of coalescence between two lineages in a single sweep is well approximated by $p_c = \bar{C}_{A,s} N^{-2A}$, where

$$\bar{C}_{A,s} = (2s)^{-2A} C_A = (2s)^{-2A} \left(\frac{\pi A}{\sin(\pi A)} \right)^2 \frac{\Gamma(2(A+1))}{1-2A}$$

in the diffusion approximation (eq. A.10). It follows that the number of selective sweeps required for the two lineages to coalesce follows a geometric distribution with parameter p_c.

Consequently, when $A < 0.5$, the expected time of coalescence between the two lineages at the neutral locus will be given by:

$$\mathbb{E}[T_2] = \frac{1}{p_c \mathcal{S}} = (N^{2A})/(\bar{C}_{A,s}\lambda) \tag{3.15}$$

neglecting the effect of genetic drift (that occurs at a rate $1/N$) that can be ignored for $A < 0.5$ (when $N^{2A} \ll N$).

As a commonly used estimator of the effective population size is proportional to the average T_2 ($\hat{N_e}^{(\pi)} = \mathbb{E}[T_2]$, see Introduction), we come to the conclusion that for a locus subject to genetic draft with $A < 0.5$, the effective population size will no longer be linear in census size but a power function of it:

$$\hat{N}_e \propto N^{2A}, \text{ with } 2A < 1.$$

This implies that the time scale of the coalescent process is N^{2A}, which potentially will be much smaller that the census size N and is consistent with the power law relation described in Buffalo (2021). The smaller the A, the stronger the reduction in coalescent time. Coalescent time remains nonetheless geometrically distributed, provided that $N^{2A} \ll N$, which will happen when A is not very close to 0.5.

To check the validity of this result, we have compared results from Monte-Carlo simulations (forward and backward) to the predictions of equation (3.15), for the case of changing environments.

Forward simulations corresponds to a haploid Wright-Fisher model with a genome of two loci (one neutral one selected) for which the environment changes at an exponential rate λ. Beneficial mutations can occur only once per individual at the selected locus after an environmental change. Population evolves for a large number of generations (about $10N$ generations) where all parent-offspring relationship are recorded. At the end, two lineages are sampled at the last environmental change. Their ancestries are tracked backward until they coalesce.

Backward simulations corresponds to a continuous time backward RIF coalescent model with no overlapping selective sweeps. We follow two lineages sampled just before an environmental change. Environment change occurs at an exponential rate λ and the waiting time to the first selective sweep (after the change) is exponential with rate $2N\mu s$. If the later happens by chance to be longer than the former, an extra environmental change is considered. Lineages can coalesce during a selective sweeps when they have the same type (as we did for simulations in Figure 3). Time is discretized in units of $0.001s$ during the sweep. Alternatively, lineages can coalescence between selective sweeps at a rate $1/N$.

Results (figure 5) again show that forward and backward simulations matches closely for all values of A. For small A (in the figure $A = 0.01$), the effective population size is almost independent of A, almost $1/\lambda$. For $A < 0.5$, the prediction of the simple approximation is good for large enough values of N. For smaller N values, the standard genetic drift dominates and simulations matches $N_e = N$. For values larger than $A > 0.5$, although we have no analytical approximation, the coalescent rate is on the order of $N_e = O(N)$.

We do note that the averages of T_2 in the simulations are often slightly larger than the ones expected from equation (3.15) for $A = \{0.01, 0.1, 0.25\}$. This is mainly due to the existence in the simulations of few rapid environment changes where selective sweeps have no time to occur. In the theoretical predictions, any change of environment leads to a new selective sweep. Hence the small difference.

3.3.2 Convergence to the Kingman tree

When A is large enough (for $A > 0.1$ in most of our simulations), the coalescence probability of two lineages for each selective event is small. In fact, it gets smaller as A gets closer to 0.5. As most of the selective events will not result in coalescing the two lineages, the fate of the neutral locus becomes mostly insensitive to the sweeps occurring at the selected locus.

Under weak genetic draft, we hypothesize that the probability of observing k lineages coalescing in a single selective sweep is much smaller than the pairwise lineages coalescence rate. We can thus define an *effective* population size

$$N_e = \frac{N^{2A}}{\bar{C}_{A,s}\lambda} \tag{3.16}$$

as the inverse of the expected pairwise coalescent time (equation 3.15). From there, the negligible probability of the multiple mergers coalescence events leads to a genealogy of k-sample that converges to a k-Kingman

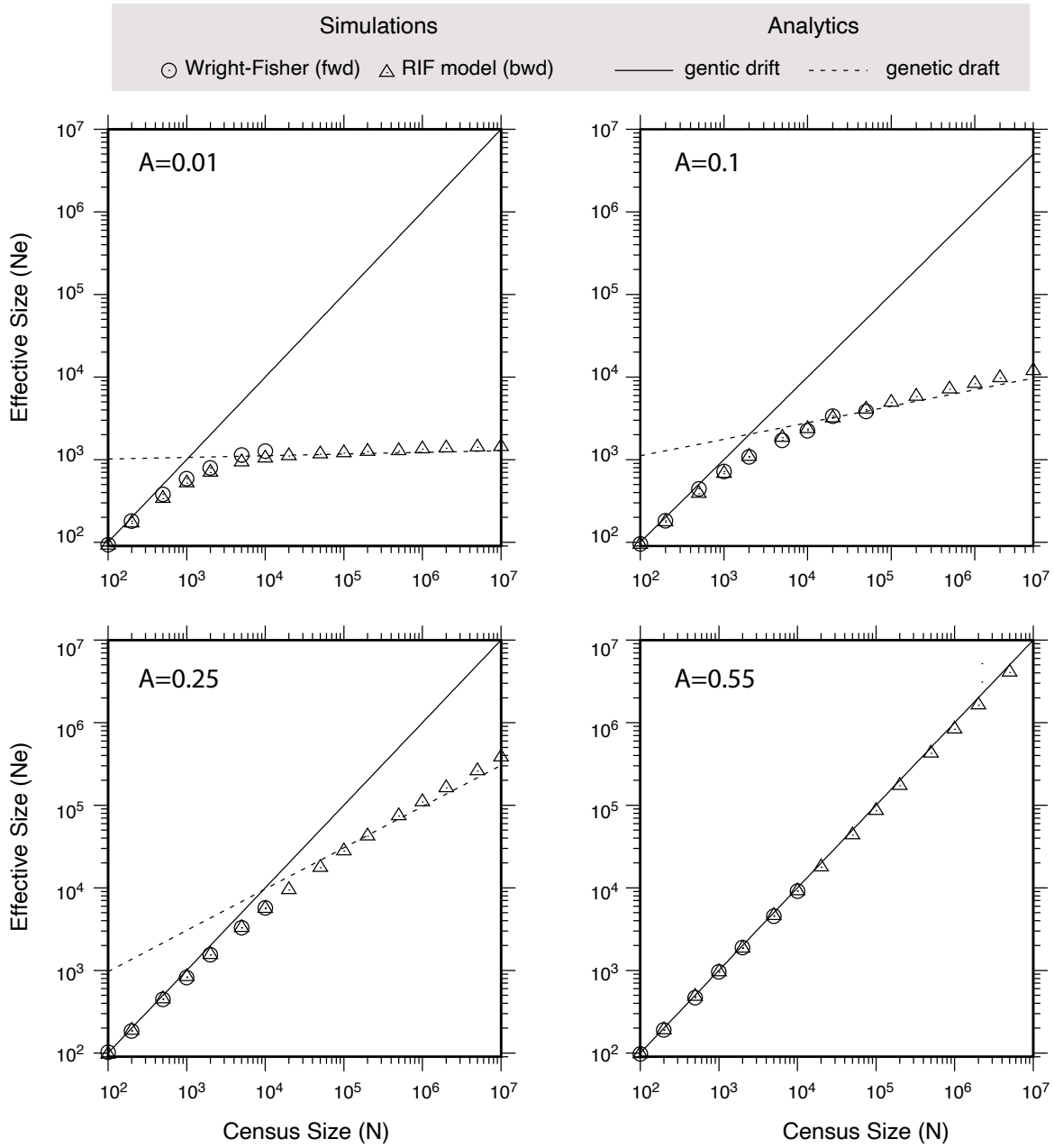

Figure 5: Average T_2 is a power function of N: $\hat{N}_e \propto N^{2A}$, for $A < 0.5$. The solid line represents the neutral case $N_e = N$, and the dashed line represents analytical predictions (eq. 3.15). The circles are averages of 10^3 forward individual based Wright-Fisher simulations. The triangles are averages of 10^4 backward RIF simulations. A line in this log-log plot signifies a power law. Each panel is a different value for $A = \{0.01, 0.1, 0.25, 0.55\}$. The other parameters are $\lambda = 0.001$, $N\mu = 0.1$ and $s = 0.05$. On average, a sweep lasts $\mathbb{E}[t^*_{fix}] = 20(\ln(N) - 4)$ generations while the environment changes every $1/\lambda = 1,000$ generations.

coalescent, once time is accelerated by N_e. In particular, T_k the first coalescence time among k lineages is then approximately distributed exponentially with rate $k(k-1)/2N_e$.

To assess the convergence of the weak draft selective regime to a Kingman coalescent, we numerically compute the mean ($\bar{T}_k$) and the variance (V_{T_k}) of T_k for $k = 2, \cdots, 7$. If the tree is a k-Kingman, $\bar{T}_k$ and $\sqrt{V_{T_k}}$ should both converge to $2N_e/k(k-1)$. As a visual test, we compare $\bar{T}_k \times (k-1)/2$ and $\sqrt{V_{T_k}} \times k(k-1)/2$ to see whether it converges to a constant (that should be N_e).

Results (Figure 6) shows that means and standard deviations of T_k for $k \in [2, 7]$ are usually close numerically supporting the convergence to exponential distributions. Furthermore, they are usually also in fairly good agreement with our theoretical expectations (genetic drift for small sizes, e.g. $N = 100$ and genetic draft for large population sizes $N > 10^5$), visually supporting the convergence to an approximately k-Kingman tree.

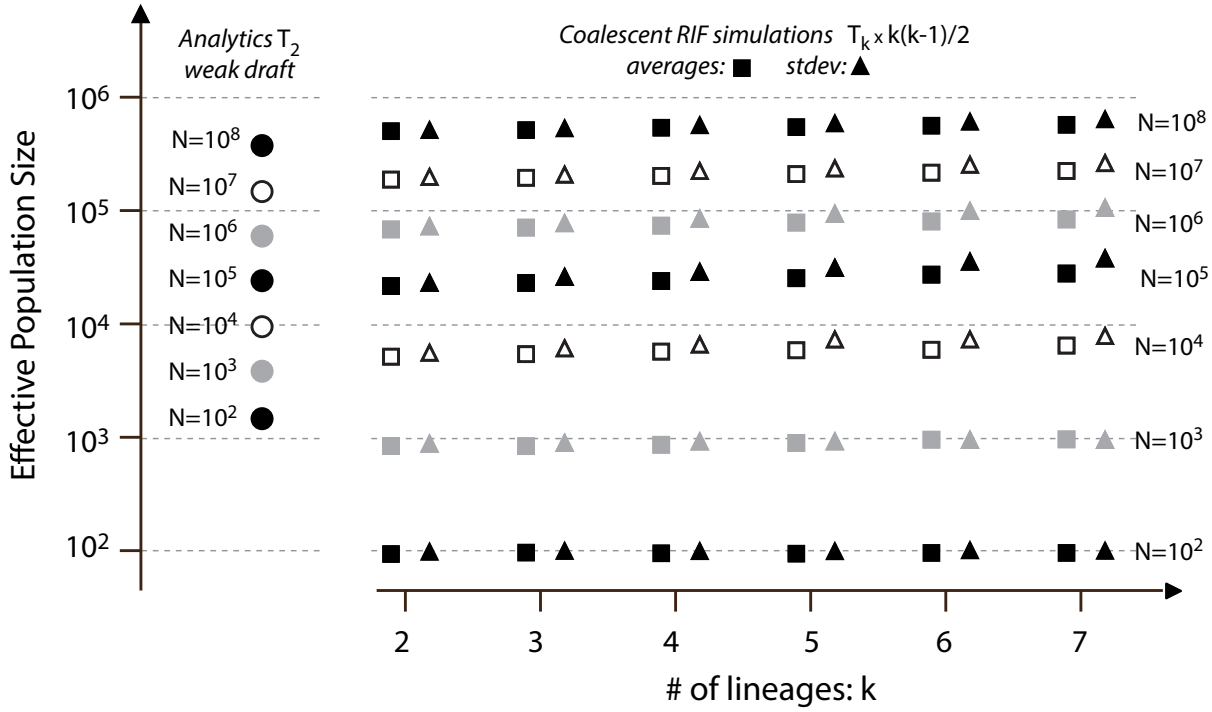

Figure 6: Time scale (N_e) of each stage of the coalescent tree. For large sizes, $N > 10^4$, the averages and their associated standard deviations follow approximately a Kingman coalescent distribution, but with a highly reduced N_e.

This implies that under weak draft regime, the coalescent tree likely converges to a neutral Kingman tree, with an overall reduced time scale.

3.3.3 Convergence to the Kingman SFS

An important feature of the SNM is the Site Frequency Spectrum (SFS) that is the distribution of frequencies among the variants. Under the SNM, assuming a an infinite site model, the expected SFS is given by $E[\xi_i] = 2N\mu/i$ (Fu, 1995). Using the Monte-Carlo backward simulator, we computed the average sample SFS for several values A for a population of 10^6 experiencing recurrent selective sweeps at rate $\lambda = 10^{-4}$. The SFS are reported first (top row, ξ_i) in branch length of coalescent time scale (N generations), disregarding the mutations themselves and then (bottom row, , $\varphi_i = i \times \xi_i / \sum \xi_i$) after normalization and transformation (Achaz, 2009) that results in a flat horizontal line for the case of Kingman coalescent.

Results (Figure 7) clearly shows that for $A = 0$, the expected SFS is monotonically rapidly decreasing, with almost exclusively singletons. If one assumes that no coalescence occur before the first selective sweep, one expects 20×10^4 generations in external branches, resulting in $E[\xi_1] = 0.2$ for the singletons. At the other extreme, in the Kingman regime ($A \to \infty$), one expects $E[\xi_1] = 2.0$ for the singletons (Fu, 1995).

Both draft regimes are well illustrated by the height and shape of the SFS. In the strong draft regime ($0 < A < 0.1$), the SFS exhibits a strong depletion of variants together with a strong U-shaped SFS (see the case of $A = 0.01$). For the weak draft regime ($0.1 < A < 0.5$), the SFS exhibits a clear depletion of variants but the shape of the SFS is Kingman-like (see $A = 0.3$). In this latter case, a neutrality test based on the SFS such as Tajima's D (Tajima, 1989) or any other SFS-based neutrality test (Achaz, 2009) will hardly reject the SNM.

4 Discussion

In this article, we have studied the effect of recurrent positive selection on neutral loci located at a given map distance A, expressed in Morgans per unit of s. We have compared theoretical derivations with individual based forward simulations and with backward coalescent simulations. We have shown that the trajectory of the frequency of a selected allele for a single selective event in finite population is very well approximated by a semi-deterministic model (Martin and Lambert, 2015) that we have name the RIF-model. This RIF model approximation is extremely efficient in computation time both in forward and in backward versions.

We have derived two new approximations for the coalescent probability under a strong selective sweep ($Ns \gg 1$) that are notably better than previous ones for "large" values of A (e.g. $A \in [0.1, 0.45]$) (see Figure 3). Based on these analytical derivations, we characterized the expected coalescent rate in a regime of recurrent sweeps (genetic draft regime) that result in strongly reduced diversity $N_e = O(N^{2A}) \ll N$, where A is the recombination/selection ratio. Importantly, under the draft regime, the relationship between N_e and N is not anymore linear, potentially explaining the still unexplained Lewontin's paradox, but certainly pointing the importance of deriving models with no proportionality between genetic diversity and population size N. As the effect of a single selective event is weak for large enough values of A, we named this regime weak genetic draft. This regime of weak genetic draft produces patterns that mimic neutrality and thus can

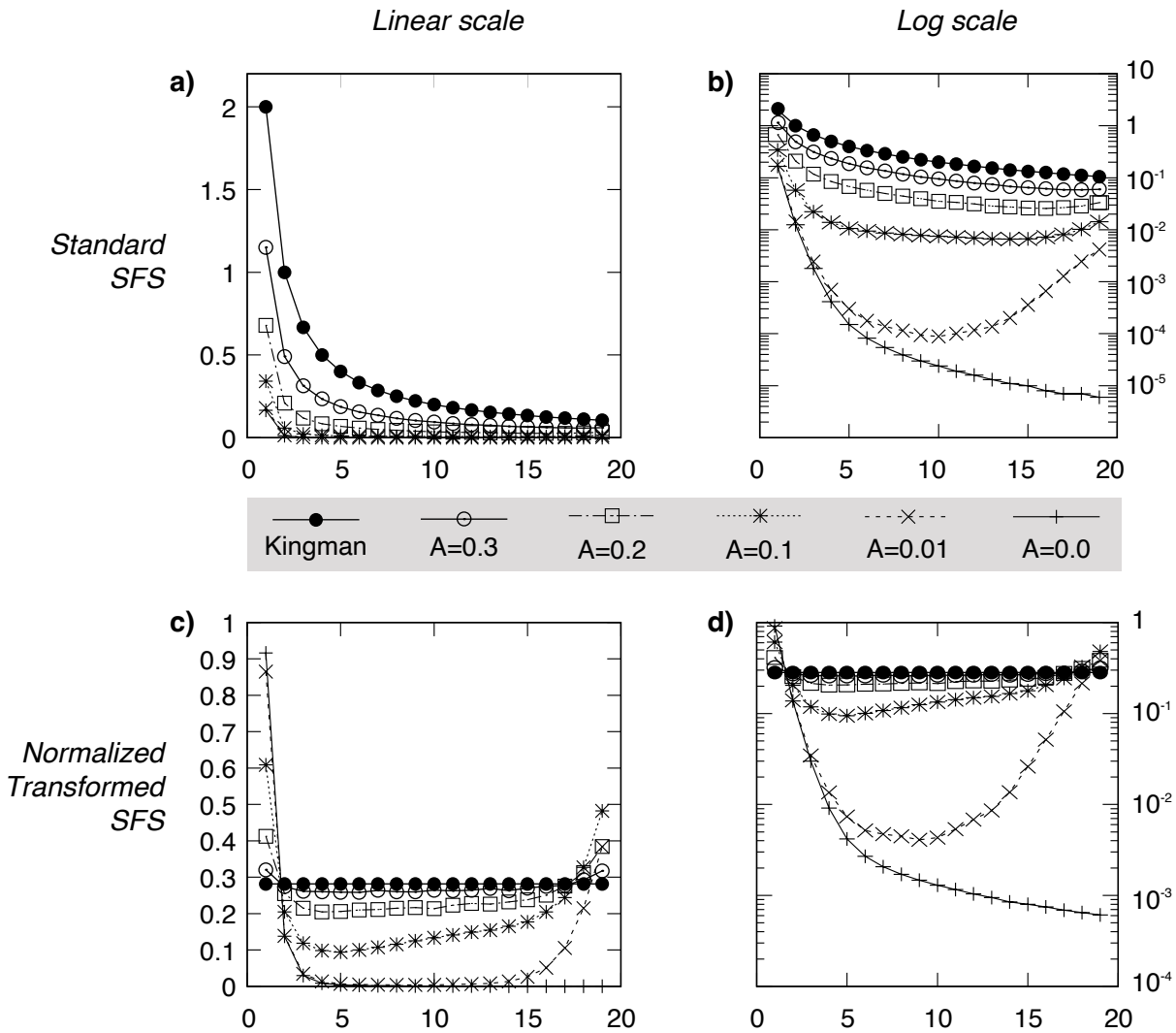

Figure 7: Average Site Frequency Spectrum (SFS) for various distance to the selected site. Averages were obtained using the backward coalescent simulator, with $N = 10^6$, $s = 0.01$, $n = 20$, $\lambda = 10^{-4}$, 10^5 replicates and $A \in \{0, 0.01, 0.1, 0.2, 0.3\}$. The case of Kingman ($A \to \infty$) is computed analytically. In the top row (a and b), the standard SFS (ξ_i) is computed in units of branch length (in a coalescent time scale of N generations). In the bottom row (c and d), the SFS (φ_i) is normalized and then transformed: $\varphi_i = i \times \xi_i / \sum \xi_i$, resulting in an horizontal line for a Kingman coalescent.

can be therefore framed in terms of *emergent neutrality* (Schiffels *et al.*, 2011; Neher and Shraiman, 2011). Importantly, at the limit, the mimicking is perfect in the weak draft as it generates an n-Kingman tree, where nodes are entirely driven by the selective events.

Under weak genetic draft, the recombination/selection ratio A values are large enough so that the emergent tree is a neutral-like Kingman tree. As a consequence, a neutrality test would not reject a standard neutral

model, even though the nodes of the tree are entirely caused by selective events. As a consequence, the weak draft regime has a large impact on the effective population size but interestingly the predictions of neutral theory are preserved.

Our predictions differ somehow from previous models of linked selection that have focused on different types of regime for the effect of linked selection. A number of studies (Etheridge *et al.*, 2006; Durrett and Schweinsberg, 2004; Weissman and Barton, 2012; Coop and Ralph, 2012; Elyashiv *et al.*, 2016) have focused on the impact of strong genetic draft, for which the distance to the selected locus is small ($A \ll 1$). Let us now emphasize one important difference between strong and weak genetic drafts. In the strong regime ($A \ll 1$), selection is so large as compared to recombination that the coalescence probability at every sweep is non-negligible. As a consequence, the time to coalescence is dictated by the time scale of the time elapsed between two consecutive sweeps.

Following up on our earlier discussion on recursive sweeps, we can explore a scenario where sweeps are triggered by environmental changes followed by rapid adaptations. As before, if we denote by λ the rate of environmental change, the time to coalescence is in strong draft typically of order $O(1/\lambda)$, *regardless of the population size*. Thus, in a strong genetic draft scenario, theory predicts that genetic draft could potentially lead to a reduction of N_e, but this reduction is problematic since it completely decorrelates N_e from N. This phenomenon is illustrated on the first panel of Figure 5 where N_e is almost insensitive to N. This important limitation of the strong draft hypothesis was already raised in previous works (Coop and Ralph, 2012; Buffalo, 2021; Charlesworth and Jensen, 2022) and has been advocated as one argument against the hitchhiking effect as a potential explanation for the observed discrepancy between $N_e^{(\pi)}$ and N.

Our calculation of the probability of coalescence takes into consideration the potential occurrence of multiple recombination events among lineages before they merge. Previous studies overlooked this particular type of coalescence event (Weissman and Barton, 2012; Elyashiv *et al.*, 2016), with some exception (Barton, 1998). This omission resulted in accurate approximations for scenarios with low recombination/selection ratios. However, our new estimations demonstrates that disregarding this event type leads to inaccurate predictions when A becomes large.

One contribution of our work is to show that a weak hitchhiking effect (weak draft) overcomes this limitation and generates a power law relation between N and N_e. For $0.1 < A < 0.5$, this is illustrated on Figure 5 for both simulations and our analytical predictions. This dependence arises from the dependence in N of the pairwise coalescence probability during a single sweep. In the strong draft regime ($A \ll 1$), this probability is barely sensitive to N. As the coalescence probability is $\propto \frac{1}{N^{2A}}$ so that when $A = O(1)$ the effective population size is $N_e^{(\pi)} \propto N^{2A}$.

We would like to emphasize that this seemingly technical point has lead to some confusion in the past. The weak draft regime has been alluded in an influential article by Coop and Ralph (2012) where it is

stated that the expected pairwise difference satisfies (see equation (18) therein)

$$\mathbb{E}(\pi) \;=\; \frac{4N\mu}{1+2\lambda N p_c} \tag{4.1}$$

where μ is the mutation rate at the neutral locus and p_c is a parameter which encapsulates the pairwise coalescence probability during a single sweep. Based on (4.1), Coop and Ralph concludes that "the key parameter is the population-scaled rate of sweep $2\lambda N$". In particular, this lead to the conclusion (see e.g. Buffalo (2021); Charlesworth and Jensen (2022)) that the weak recursive selective sweep would generate an effective population size N_e independent of N as along as $2N\lambda \gg 1$. An important difference between the approach in Coop and Ralph (2012) and the present work lies in the assumptions on the coalescence probability p_c. In Coop and Ralph (2012), p_c is considered as an exogenous parameter reflecting some hidden sweep mechanism (e.g., random selection coefficient, changing environment during the sweep). This is in contrast with the present work where we start from an explicit individual based model and we show that the coalescence probability p_c has an implicit dependence in N, i.e. $p_c = \frac{\bar{C}_{A,s}}{N^{2A}}$. Plugging this expression into (4.1), we find now that for $A < 0.5$ (neglecting the first term of genetic drift in the numerator):

$$\mathbb{E}(\pi) \;\approx\; N^{2A}\frac{2\mu}{\lambda \bar{C}_{A,s}}.$$

Again, this leads to an estimation of the effective population $N_e^{(\pi)} \propto N^{2A}$ which is consistent with our previous estimation of $\mathbb{E}(T_2)$ in eq. (3.15). To summarize, our approach is consistent with the one of Coop and Ralph, but we provide an explicit coalescence probability as a function of the population size N. In turn, this leads to a drastically different estimation of the effective population size.

In a strong draft scenario, the linkage is so strong that typically, several lineages coalesce at each selective sweep, resulting in processes that are known as Multiple Merger Coalescents (MMC) (Berestycki, 2009; Tellier and Lemaire, 2014) . At the limit of $A = 0$, the genealogical tree of the neutral is entirely driven by the last selective event. If this event was strong $Ns \gg 1$, the genealogy reassembles a radiation where all lineages coalesce very rapidly. For slightly larger values of A, only a part of the lineages coalesce at each selective events resulting in trees such as Bolthausen-Snitzman trees or more generally other MMC such as $\Lambda - coalescent$ or $\Xi - coalescent$. In contrast, we showed that the limiting genealogy in a weak draft regime is indistinguishable from a neutral genealogy.

On figure 8, we have pictured the different regimes of evolution as a function of the distance to the selected loci.

- At small distance ($A < 0.1$), the neutral loci are highly impacted by the selected events (strong draft regime) : (a) their genealogy is described by a Multiple-Merger Coalescent (MMC) tree (Berestycki, 2009; Tellier and Lemaire, 2014) and (b) their genetic diversity is extremely reduced and N_e is insensitive to the underlying population size N ($N_e \propto 1/\lambda$).

- At intermediate distance ($0.1 < A < 0.5$), the neutral loci are moderately impacted by the selected events (weak draft regime) : (a) their genealogy is described by a Kingman (Kgm) tree although (b) their genetic diversity is largely reduced ($N_e \propto N^{2A}$).

- At large distance ($A > 0.5$), neutral loci are weakly impacted by selective events (genetic drift regime): (a) their genealogy is described by a Kingman (Kgm) tree and (b) diversity $N_e^{(\pi)} \propto N$.

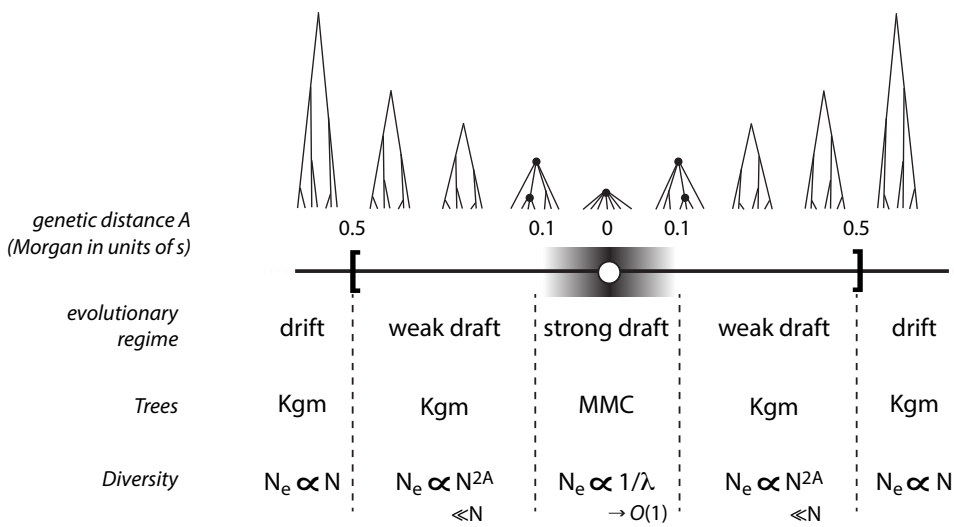

Figure 8: Overview of the different regimes of evolution as a function of the distance to the selected locus. **Strong draft**: at small distance ($A \approx 0$), the tree is best described by a Multiple Merger Coalescent and the diversity is insensitive to population size: $N_e^{(\pi)} \propto 1/\lambda$. **Weak draft**: at larger distance ($A \in [0.1, 0.5[$), the tree is best described by a Kingman tree but diversity is highly reduced $N_e^{(\pi)} \propto N_e^{2A}$. **Genetic drift**: at even larger distance ($A \geq 0.5$), the tree is a Kingman tree and diversity is given by the population size: $N_e^{(\pi)} \propto N$.

Intuitively, if one assumes that beneficial mutations arise frequently and are scattered throughout the genome, the average genomic pattern will likely be the result of a mix of the different scenarios, where the strongly affected regions likely encompass only a small proportion of the genomes. It is then tempting to postulate that the average patterns of diversity such as genome-wide SFS would be better fitted by Multiple Merger Coalescent models than by a pure Kingman coalescent model, an observation recently supported for a wide range of organisms (Freund *et al.*, 2023; Árnason *et al.*, 2023).

Although the model discussed here is framed in terms of fluctuating environment that triggers selective sweeps, it is also compatible with a regime of evolution by compensation, where deleterious substitutions accumulate in the population until a compensating mutation appears and then sweeps to restore the original fitness (Kimura, 1985). The role of compensatory mutations in molecular evolution has since then been extensively supported by data analysis (e.g. (Kondrashov *et al.*, 2002; Szamecz *et al.*, 2014; Latrille

et al., 2024)). Compensatory evolution differs from the regime of background selection (Charlesworth *et al.*, 1993) by the recurrent triggering of new selective sweeps.

Therefore loci which are immune to selected events should not be in the segment of size A (0.5 on both size) of the selected loci. To have crude sense of theses distances, if we consider selected events of $s = 0.01$ (1.01 relative advantage), the impacted area of a selected events is 1 cM (about 1Mb of the human genome and more generally on the order of 1% of a chromosome). Stronger selected events will impact the chromosome on a larger region. At any generation, many selective events can occur without driving the species to extinction (hundreds for Haldane (1957), much more for Maynard Smith (1968)). Therefore, it is likely a large fraction of the genomes is affected by linked selection at any generation, an hypothesis that has been proposed based on different arguments (Pouyet *et al.*, 2018).

As discussed above, at least two different regimes of genetic draft can be studied. In a first case, each new selective sweep occurs at a randomly chosen site in the chromosome. In this regime, the strongest diversity reduction for a neutral site, will be driven by rare events where the selected sites is closely linked to the studied neutral locus. The weak draft regime in this scenario only occurs transiently until a beneficial variant appears in the vicinity of the neutral locus. In the second case, potentially selected sites have fixed positions and thus neutral locus can well be always at a appreciable map distance of the selected site. For these neutral loci, the regime of weak genetic draft will entirely drives their evolution.

One particular restrictive assumption underlying the current study is that selection sweeps are non-overlapping. Under this assumption, we showed that the weak draft regime only had a sizable effect in a restricted region around the selected locus, i.e., when the recombination-selection ratio A is less than 0.5. One first step is to combine the effect of different selective sweeps occurring in the same time window but at different sites as in Weissman and Barton (2012); Elyashiv *et al.* (2016). However, since the weak draft regime is mainly driven by a high sweep frequency of sweeps, it is of fundamental importance to understand the weak draft regime in a scope of parameters where several beneficial mutations are in clonal interference. Whether the conclusion remains consistent with the present study (emerging neutrality together with a highly reduced time scale) and whether weak sweeps at higher frequency could potentially impact a larger scope of the genome (*i.e.* a recombination-selection ratio A beyond 0.5) requires to go beyond the current theoretical framework and is the subject of a further study.

Although, we do not pretend that this simple weak genetic draft model is realistic enough to solve Lewontin's paradox of variation, it nonetheless is an interesting step forward in understanding the effect of sparsed linked selection and in proposing models where diversity does not scale linearly with population size but as a power law that linear in log-log, an observation strongly supported by the most recent data analysis (Buffalo, 2021). Although the model we have consider here is simplistic as it assumes no demography, no structure and fixed values of c and s, it paves the way for future more realistic models. It now remains to explore all relaxations of the model such as having overlapping selective sweeps, random c and s, soft selective sweeps, incomplete sweeps to appreciate how it tunes diversity as a function of population size.

Acknowledgments

We would like to thank F. Ged and N. Barton for providing constructive feedbacks on this manuscript.

References

Achaz, G., 2009 Frequency spectrum neutrality tests: one for all and all for one. Genetics **183**: 249–58.

Aguade, M., N. Miyashita, and C. H. Langley, 1989 Reduced variation in the yellow-achaete-scute region in natural populations of drosophila melanogaster. Genetics **122**: 607–15.

Árnason, E., J. Koskela, K. Halldórsdóttir, and B. Eldon, 2023 Sweepstakes reproductive success via pervasive and recurrent selective sweeps. Elife **12**.

Barton, N., 1998 The effect of hitch-hiking on neutral genealogies. Genet. Res. Camb. **72**: 123–133.

Barton, N. H., 2000 Genetic hitchhiking. Philos Trans R Soc Lond B Biol Sci **355**: 1553–62.

Barton, N. H., and A. M. Etheridge, 2004 The effect of selection on genealogies. Genetics **166**: 1115–31.

Begun, D. J., and C. F. Aquadro, 1992 Levels of naturally occurring dna polymorphism correlate with recombination rates in d. melanogaster. Nature **356**: 519–20.

Berestycki, N., 2009 Recent progress in coalescent theory. arXiv preprint arXiv:0909.3985 .

Buffalo, V., 2021 Quantifying the relationship between genetic diversity and population size suggests natural selection cannot explain lewontin's paradox. Elife **10**.

Charlesworth, B., 2009 Effective population size and patterns of molecular evolution and variation. Nature Reviews Genetics **10**: 195–205.

Charlesworth, B., and J. D. Jensen, 2022 How can we resolve lewontin's paradox? Genome Biol Evol **14**.

Charlesworth, B., M. T. Morgan, and D. Charlesworth, 1993 The effect of deleterious mutations on neutral molecular variation. Genetics **134**: 1289–303.

Coop, G., and R. C. Griffiths, 2004 Ancestral inference on gene trees under selection. Theoretical population biology **66**: 219–232.

Coop, G., and P. Ralph, 2012 Patterns of neutral diversity under general models of selective sweeps. Genetics **192**: 205–24.

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.

Durrett, R., and J. Schweinsberg, 2004 Approximating selective sweeps. Theoretical population biology **66**: 129–138.

Ellegren, H., and N. Galtier, 2016 Determinants of genetic diversity. Nat Rev Genet **17**: 422–33.

Elyashiv, E., S. Sattath, T. T. Hu, A. Strutsovsky, G. McVicker, *et al.*, 2016 A genomic map of the effects of linked selection in drosophila. PLoS Genet **12**: e1006130.

Etheridge, A., 2011 *Some Mathematical Models from Population Genetics: École D'Été de Probabilités de Saint-Flour XXXIX-2009*, volume 2012. Springer Science & Business Media.

Etheridge, A., P. Pfaffelhuber, and A. Wakolbinger, 2006 An approximate sampling formula under genetic hitchhiking .

Evans, S. N., and N. O'Connell, 1994 Weighted occupation time for branching particle systems and a representation for the supercritical superprocess. Canadian Mathematical Bulletin **37**: 187–196.

Ewing, G., and J. Hermisson, 2010 MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. Bioinformatics **26**: 2064–2065.

Fay, J. C., and C. I. Wu, 2000 Hitchhiking under positive darwinian selection. Genetics **155**: 1405–13.

Freund, F., E. Kerdoncuff, S. Matuszewski, M. Lapierre, M. Hildebrandt, *et al.*, 2023 Interpreting the pervasive observation of u-shaped site frequency spectra. PLoS Genet **19**: e1010677.

Fu, Y. X., 1995 Statistical properties of segregating sites. Theor Popul Biol **48**: 172–97.

Galtier, N., and M. Rousselle, 2020 How much does ne vary among species? Genetics **216**: 559–572.

Gillespie, J. H., 2000a Genetic drift in an infinite population. the pseudohitchhiking model. Genetics **155**: 909–919.

Gillespie, J. H., 2000b The neutral theory in an infinite population. Gene **261**: 11–18.

Haldane, J. B. S., 1957 The cost of natural selection. Journal of Genetics **55**: 511–524.

Hein, J., M. H. Schierup, and C. Wiuf, 2005 *Gene Genealogies, Variation and Evolution*. Oxford University Press.

Hermisson, J., and P. S. Pennings, 2005 Soft sweeps: molecular population genetics of adaptation from standing genetic variation. Genetics **169**: 2335–52.

Hudson, R. R., and N. L. Kaplan, 1988 The coalescent process in models with selection and recombination. Genetics **120**: 831–40.

Kaplan, N. L., R. R. Hudson, and C. H. Langley, 1989 The "hitchhiking effect" revisited. Genetics **123**: 887–99.

Kimura, M., 1968 Evolutionary rate at the molecular level. Nature **217**: 624–6.

Kimura, M., 1983 *The Neutral Theory of Molecular Evolution*. Cambridge University Press.

Kimura, M., 1985 The role of compensatory neutral mutations in molecular evolution. Journal of Genetics **64**: 7–19.

Kimura, M., and T. Ohta, 1971 On the rate of molecular evolution. J Mol Evol **1**: 1–17.

King, J. L., and T. H. Jukes, 1969 Non-darwinian evolution. Science **164**: 788–98.

Kingman, J., 1982 The coalescent. Stochastic Process. Appl **13**: 235–248.

Kondrashov, A. S., S. Sunyaev, and F. A. Kondrashov, 2002 Dobzhansky-muller incompatibilities in protein evolution. Proc Natl Acad Sci U S A **99**: 14878–83.

Lambert, A., 2008 Population dynamics and random genealogies. Stochastic models **24**: 45–163.

Latrille, T., J. Joseph, D. A. Hartasánchez, and N. Salamin, 2024 Estimating the proportion of beneficial mutations that are not adaptive in mammals. PLoS Genet **20**: e1011536.

Leffler, E. M., K. Bullaughey, D. R. Matute, W. K. Meyer, L. Ségurel, *et al.*, 2012 Revisiting an old riddle: what determines genetic diversity levels within species? PLoS Biol **10**: e1001388.

Lewontin, R. C., 1974 *The genetic basis of evolutionary change*, volume 560. New York: Columbia University Press.

Malécot, G., 1948 *Les Mathématiques de l'Hérédité*. Masson et Cie. Paris.

Martin, G., and A. Lambert, 2015 A simple, semi-deterministic approximation to the distribution of selective sweeps in large populations. Theor Popul Biol **101**: 40–6.

Maynard Smith, J., 1968 "haldane's dilemma"and the rate of evolution. Nature **219**: 1114–1116.

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

Neher, R. A., and B. I. Shraiman, 2011 Genetic draft and quasi-neutrality in large facultatively sexual populations. Genetics **188**: 975–96.

OHTA, T., 1973 Slightly deleterious mutant substitutions in evolution. Nature **246**: 96–98.

Ohta, T., and M. Kimura, 1971 On the constancy of the evolutionary rate of cistrons. J Mol Evol **1**: 18–25.

Patwa, Z., and L. M. Wahl, 2008 The fixation probability of beneficial mutations. J R Soc Interface **5**: 1279–89.

Pennings, P. S., and J. Hermisson, 2006a Soft sweeps ii–molecular population genetics of adaptation from recurrent mutation or migration. Mol Biol Evol **23**: 1076–84.

Pennings, P. S., and J. Hermisson, 2006b Soft sweeps iii: the signature of positive selection from recurrent mutation. PLoS Genet **2**: e186.

Pitman, J., 1999 Coalescents with multiple collisions. Annals of Probability **27**: 1870–1902.

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**.

Robertson, A., 1961 Inbreeding in artificial selection programmes. Genet Res **2**: 189–194.

Romiguier, J., P. Gayral, M. Ballenghien, A. Bernard, V. Cahais, *et al.*, 2014 Comparative population genomics in animals uncovers the determinants of genetic diversity. Nature **515**: 261–3.

Sagitov, S., 1999 The general coalescent with asynchronous mergers of ancestral lines. Journal of Applied Probability **36**: 1116–1125.

Schiffels, S., G. J. Szöllosi, V. Mustonen, and M. Lässig, 2011 Emergent neutrality in adaptive asexual evolution. Genetics **189**: 1361–75.

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.

Stephan, W., and C. H. Langley, 1989 Molecular genetic variation in the centromeric region of the x chromosome in three drosophila ananassae populations. i. contrasts between the vermilion and forked loci. Genetics **121**: 89–99.

Szamecz, B., G. Boross, D. Kalapis, K. Kovács, G. Fekete, *et al.*, 2014 The genomic landscape of compensatory evolution. PLoS Biol **12**: e1001935.

Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by dna polymorphism. Genetics **123**: 585–95.

Tellier, A., and C. Lemaire, 2014 Coalescence 2.0: a multiple branching of recent theoretical developments and their applications. Molecular ecology **23**: 2637–2652.

Wakeley, J., 2009 *Coalescent theory, an Introduction*. Roberts and Company.

Waples, R. S., 2022 What is ne, anyway? J Hered **113**: 371–379.

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

Wiehe, T. H., and W. Stephan, 1993 Analysis of a genetic hitchhiking model, and its application to dna polymorphism data from drosophila melanogaster. Mol Biol Evol **10**: 842–54.

Wright, S., 1931 Evolution in mendelian populations. Genetics **16**: 97–159.

Zuckerkandl, E., 1962 Molecular disease, evolution, and genic heterogeneity. Horizons in biochemistry : 189–225.

A Appendices

A.1 Coalescence probability under the RIF model

Consider now two lineages starting in state 0. Let T_c be the time to coalescence which is possibly infinite if the two lineages do not coalesce. Our aim is to compute:

$$\begin{aligned}\mathbb{P}^{RIF}_{\text{coal2}} &= \mathbb{P}(T_c \leq \tau) \\ &\approx \frac{1}{(2Ns)^{2A}}\left(\frac{\pi A}{\sin(\pi A)}\right)^2\left[2^{2A}\int_0^\infty \gamma(1-2A,\frac{2}{x})e^{-x}dx \;+\; \int_0^\infty x^{2A}e^{-x-\frac{2}{x}}dx\right] \qquad (A.1)\end{aligned}$$

The derivation of this formula is decomposed below in several steps.

A single lineage. We first consider a single lineage starting in state 0 at time 0. We first show that the probability of a single lineage to be in state 0 close to time τ (time $\tau - t$ close to the onset of the sweep forward in time) is of the order $1/(Ns)^A$. We recall that $\hat{x}_t = 1 - x_{\tau-t}$ is the reversed RIF trajectory as defined in Section 3.2.1. We consider the conditional probability of being linked to a type 0, $u_t := \mathbb{P}(L_t = 0|\mathcal{E},\mathcal{E}')$. Then

$$\frac{du_t}{dt} = -A\hat{x}_t u_t + A(1-\hat{x}_t)(1-u_t), \quad \text{with } u_0 = 1.$$

A direct computation around τ leads to

$$\mathbb{P}(L_{\tau-t} = 0|\mathcal{E},\mathcal{E}') = \underbrace{A\int_t^\tau \exp(-A(v-t))x_v dv}_{(i)} + \underbrace{\exp\left(-A(\tau-t)\right)}_{(ii)}. \qquad (A.2)$$

We first evaluate the term (i)

$$A\int_t^\tau \exp(-A(v-t))x_v dv = A(\frac{e^t}{Ns})^A \int_{\frac{e^t}{Ns}}^{\frac{e^\tau}{Ns}} \frac{1}{w^{A+1}} x_{\ln(Ns)+\ln(w)} dw \qquad (A.3)$$

where we used the change of variable $w = e^v/Ns$. From the explicit solution of the logistic equation for x, this yields the limit

$$\begin{aligned}\lim_{Ns\to\infty}(2Ns)^A\ (i) &= A(2e^t)^A\int_0^\infty \frac{\frac{\mathcal{E}}{2}dw}{w^A(1+\frac{\mathcal{E}}{2}w)} \\ &= (\mathcal{E}e^t)^A\frac{\pi A}{\sin(\pi A)} \qquad (A.4)\end{aligned}$$

We now show that the second term (ii) is negligible compared to (i) as $Ns \gg 1$. This is indeed the case since, as τ is given by (3.11), we have

$$\lim_{Ns\to\infty}(Ns)^A \exp\left(-A(\tau-t)\right) = \lim_{Ns\to\infty}(Ns)^A \times \frac{1}{(Ns)^{2A}}(\frac{e^t}{4}\mathcal{E}\mathcal{E}')^A = 0.$$

This implies that $(ii) = o(i)$. Gathering the previous estimates yield

$$\mathbb{P}(L_{\tau-t} = 0|\mathcal{E},\mathcal{E}') \approx \left(\frac{\mathcal{E}e^t}{2Ns}\right)^A \frac{\pi A}{\sin(\pi A)}. \tag{A.5}$$

Two lineages. We will now use the previous estimate to evaluate the coalescence probability of two lineages during a sweep. Each lineage starts at backward time 0 with a type $L_t^1 = L_t^2 = 0$ and can switch to state 1 by recombination at rate $A\hat{x}_t$. It switches back to type 0 at rate $A(1-\hat{x}_t)$. Two lineages of type 0 coalesce at rate $\frac{1}{Ns(1-\hat{x}_t)}$ and they coalesce at rate $\frac{1}{Ns\hat{x}_t}$ when both are type 1. Figure 9 provides an illustration of the notations of the considered process.

We will now show that under the RIF model, the probability of coalescence during the selective sweep is well approximated by a sum of two terms:

I) the probability of coalescence between two types 0 at $t = \tau - O(1)$, close to the onset of the selective sweep.

II) the probability that both lineages are type 0 and reach τ because they had no opportunity for coalescence before. They immediately coalesce in the ancestor of all derived alleles.

To get our approximation of the coalescence probability, we construct the pair of *dependent* coalescing lineages $(\tilde{L}_t^1, \tilde{L}_t^2; t \le \tau)$ from two *independent* lineage processes $(L_t^1, L_t^2; t \le \tau)$. This is done as follows. Given a realisation of the reversed fixation trajectory $(\hat{x}_t; t \in [0,\tau])$, we define two independent Poisson Point Processes $\hat{\mathcal{P}}_{0,0}$ (potential coalescence events between two types 0) and $\hat{\mathcal{P}}_{1,1}$ (potential coalescence events between two types 1) with respective intensity measures:

$$\frac{1}{Ns(1-\hat{x}_t)} \quad \text{and} \quad \frac{1}{Ns\hat{x}_t}.$$

To enforce the constraint that $(0,0)$-lineages coalesce at time τ, we add an extra point to $\hat{\mathcal{P}}_{0,0}$ at time τ.

Next, we define T_c as the time at which a potential coalescent event ($\hat{\mathcal{P}}_{0,0}$ or $\hat{\mathcal{P}}_{1,1}$) coincides with two lineages having the correct type ($L_t^1 = L_t^2 = 0$ if $t \in \hat{\mathcal{P}}_{0,0}$ or $L_t^1 = L_t^2 = 1$ for $t \in \hat{\mathcal{P}}_{1,1}$). At such time, the two lineages must coalesce. We call such a potential event an effective coalescence event. Formally:

$$T_c \;:=\; \inf\{t \in \hat{\mathcal{P}}_{0,0} : L_t^1 = L_t^2 = 0\} \vee \inf\{t \in \hat{\mathcal{P}}_{1,1} : L_t^1 = L_t^2 = 1\}$$

with the possibility that it never occurs. Finally, we define a pair of *dependent* coalescing lineages as identical to the *independent* ones before T_c, while they both equate after T_c. Formally:

$$\bar{L}_t^1 \;=\; L_t^1 \quad \text{and} \quad \bar{L}_t^2 \;=\; \mathbf{1}_{(T_c<t)}L_t^1 \;+\; \mathbf{1}_{(T_c\ge t)}L_t^2$$

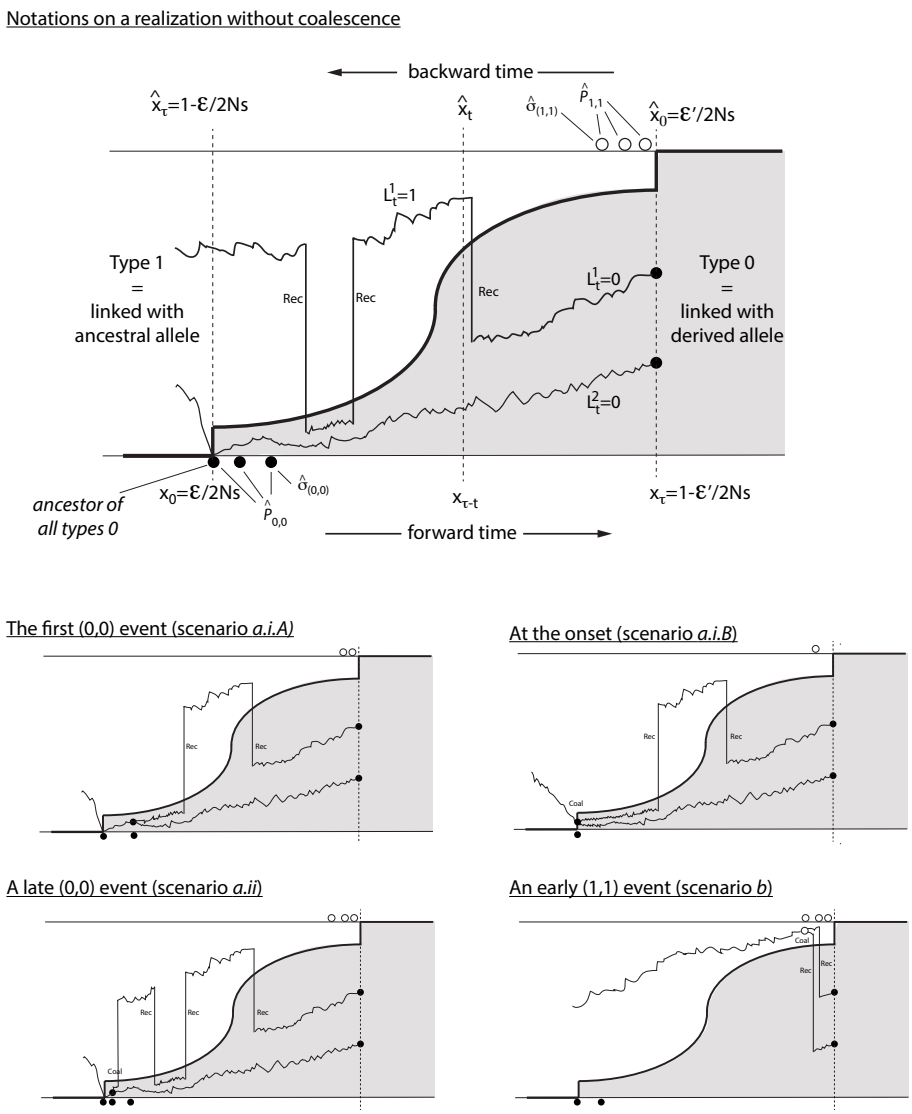

Figure 9: A sketch of the coalescent process under the RIF model. The top panel presents all notations used in the text. The four bottom panels reports all possible scenarios for an effective coalescence. The top panels represent the most likely events, namely when the two lineages are type 0 at the first $(0,0)$ potential coalescence event $\hat{\sigma}_{0,0}$. **scenario** $a.i.A$: the coalescence occurs at time $\hat{\sigma}_{0,0} > 0$. **scenario** $a.i.B$: no *positive* potential $(0,0)$-coalescence times. Coalescence occurs at the onset of the sweep at backward time τ. (part B of (A.1)). **scenario** $a.ii$: There is a $(0,0)$-effective event occurring after the first potential $(0,0)$-event $\hat{\sigma}_{0,0}$. In this scenario, a lineage transitions from 1 to 0 on the time interval $[\hat{\sigma}_{0,0}, \tau]$. **scenario** b effective $(1,1)$ coalescence event. The two lineages must transition from 0 to 1 before the occurrence of the last $(1,1)$ potential coalescence event $\hat{\sigma}_{(1,1)}$.

Using this construction, we will now argue that the coalescence probability of two lineages is approximately the probability that L^1 and L^2 are in state 0 at the first potential coalescence event $\hat{\mathcal{P}}_{0,0}$. According to the previous construction, coalescence can be decomposed into 3 scenarii (see *a.i*, *a.ii* and *b*). Figure 9 provides an illustration of all possible coalescence events.

scenario *a*. The effective coalescence is a $(0,0)$-event. This scenario is itself decomposed into two sub-scenarii (see below). First, let $\hat{\sigma}_{(0,0)} = \inf\{t : t \in \hat{\mathcal{P}}_{0,0}\}$ be the first potential $(0,0)$-event. Regarding $\hat{\sigma}_{(0,0)}$, an effective $(0,0)$-event can only occur in two ways:

scenario *a.i*. The independent lineages (L^1, L^2) are in state $(0,0)$ at time $\hat{\sigma}_{(0,0)}$. Either they coalesce at $\hat{\sigma}_{(0,0)} < \tau$ (scenario *a.i.A* of Figure 9), when the first (0,0) event is before the onset on the sweep, or at $\hat{\sigma}_{(0,0)} = \tau$ when no (0,0) event exist before the onset (scenario *a.i.B*).

scenario *a.ii*. The coalesce occurs at some time $t > \hat{\sigma}_{(0,0)}$, that is at subsequent point $\hat{\mathcal{P}}_{0,0}$, closer to τ, the onset of the sweep.

scenario *b*. The coalescence is a $(1,1)$-event. Both lineages recombine with resident alleles very early and coalesce at one of the (1,1)-event.

We will now show that only scenario *a.i* contribute significantly to the probability of coalescence and that the scenarii *a.ii* and *b* can be neglected.

Scenario *a.i*: This probability can be decomposed into two parts:

$$\mathbb{P}(a.i) = \frac{1}{(2Ns)^{2A}}\left[I + J\right] \tag{A.6}$$

where

$$\begin{aligned} I &:= (2Ns)^{2A}\mathbb{E}\left[\int_0^\tau \underbrace{\frac{1}{Ns(1-\hat{x}_t)}\exp(-\int_t^\tau \frac{du}{Ns(1-\hat{x}_u)})}_{\text{density of the } 1^{th} \text{ point in } \hat{\mathcal{P}}_{0,0}} \mathbb{P}\left(L_t = 0|\mathcal{E},\mathcal{E}'\right)^2 dt\right] \\ J &:= (2Ns)^{2A}\mathbb{E}\left[\underbrace{\exp(-\int_0^\tau \frac{du}{Ns(1-\hat{x}_u)})}_{\text{no point in } \hat{\mathcal{P}}_{0,0}} \mathbb{P}\left(L_\tau = 0|\mathcal{E},\mathcal{E}'\right)^2\right] \end{aligned}$$

Here, J is the (rescaled by $(2Ns)^A$) probability that the only potential $(0,0)$-event is time τ (onset of the sweep forward in time); and I is the (rescaled) probability that there is at least one potential $(0,0)$ event different from τ, and that the independent lineages are in state 0 at the first potential $(0,0)$ event.

To estimate I, we use that for a fixed time t,

$$\begin{aligned} I &= (2Ns)^{2A}\mathbb{E}\left[\int_0^\tau \frac{1}{Ns(1-\hat{x}_t)}\exp(-\int_t^\tau \frac{du}{Ns(1-\hat{x}_u)})\,\mathbb{P}\left(L_t = 0|\mathcal{E},\mathcal{E}'\right)^2\right] \\ &= \mathbb{E}\left[\int_0^\tau \frac{1}{Nsx_t}\exp(-\int_t^\tau \frac{du}{Nsx_u})\left((2Ns)^A\mathbb{P}\left(L_{\tau-t} = 0|\mathcal{E},\mathcal{E}'\right)\right)^2 du\right]. \end{aligned} \tag{A.7}$$

We have $Nsx_t \approx \frac{\mathcal{E}e^t}{2}$ when $Ns \to \infty$. Thus

$$\begin{aligned} I &\approx \left(\frac{\pi A}{\sin(\pi A)}\right)^2 2^{2A}\mathbb{E}\left[\int_0^\infty \frac{2}{\mathcal{E}e^t}\exp(-\int_t^\infty \frac{2du}{e^u\mathcal{E}})(\frac{\mathcal{E}}{2}e^t)^{2A}dt\right] \\ &= \left(\frac{\pi A}{\sin(\pi A)}\right)^2 \int_0^\infty 2^{2A}\gamma(1-2A,\frac{2}{x})e^{-x}dx, \end{aligned} \tag{A.8}$$

where γ is the lower gamma function. The second line is obtained by the change of variable $\bar{t} = \tau - t$. In the third line, we used the approximation $Nsx_t \approx \frac{\mathcal{E}e^t}{2}$ By a similar computation, we estimate J and further show that:

$$\mathbb{P}(a.i) \approx \frac{1}{(2Ns)^{2A}}\left(\frac{\pi A}{\sin(\pi A)}\right)^2\left[2^{2A}\int_0^\infty \gamma(1-2A,\frac{2}{x})e^{-x}dx + \int_0^\infty x^{2A}e^{-x-\frac{2}{x}}dx\right]. \tag{A.9}$$

To derive our final result (A.1), we now show that scenarii $(a.ii)$ and (b) are negligible compared to scenario $(a.i)$. Let us first estimate the probability of scenario b, that is a $(1,1)$-coalescence event. For a $(1,1)$-effective event, at least one of the two independent lineages transitions from state 0 to 1 in the time interval $[0, \hat{\sigma}_{(1,1)}]$, where $\hat{\sigma}_{(1,1)} = \sup\{t \in \hat{\mathcal{P}}_{1,1}\}$. This entails that

$$\mathbb{P}(b) \leq 2 \times \mathbb{E}\left(\int_0^\tau \frac{1}{Ns\hat{x}_t}e^{-\int_0^t \frac{1}{Ns\hat{x}_u}du}q_t dt\right).$$

where $q_t = \mathbb{P}_0(\exists u \in [0,t],\ L_u = 1|\mathcal{E},\mathcal{E}')$. Now,

$$q_t = 1 - e^{-A\int_0^t \hat{x}_u du} \leq At\hat{x}_t.$$

It follows that

$$\mathbb{P}(b) \leq \tfrac{2A}{Ns}\mathbb{E}\left(\int_0^\tau t dt\right) = \frac{A}{Ns}\mathbb{E}[(\tau)^2]$$

Since $\tau = O(\ln(Ns))$, it follows that

$$\mathbb{P}(b) = O\left(\frac{\ln(Ns)^2}{Ns}\right) \ll \mathbb{P}(a.i)$$

.

It remains to estimate scenario (a.ii) that coalescence occured at a (0,0) event after $\hat{\sigma}_{(0,0)}$. Under this event, one of the two lineages transitions from state 1 to 0 on the time interval $[\hat{\sigma}_{0,0}, \tau]$. By reversibility of $\hat{x}$, it is not difficult to see that the probability of latter event is identical in distribution to our upper bound for (b), so that

$$\mathbb{P}(a.ii) = O\left(\frac{\ln(Ns)^2}{Ns}\right) \ll \mathbb{P}(a.i)$$

Gathering all the previous estimates, this yields the coalescence probability as written in (A.1).

A.2 Coalescence probability under the Diffusion model

A similar reasoning on the diffusion model shows that the pairwise coalescent probability can be approximated in the following way

$$\mathbb{P}^{\text{diff.}}_{\text{coal2}} \approx \frac{1}{(2Ns)^{2A}}\left(\frac{\pi A}{\sin(\pi A)}\right)^2 \frac{\Gamma(2(A+1))}{1-2A}. \tag{A.10}$$

We compare this expression to the approximation under the RIF model. Figure 3 shows that both expressions lead to slightly different but numerically very similar outcomes. The rational behind this result is a combination our the approach developed in the RIF model together with the method developed in Etheridge *et al.* (2006). We decompose the outline of the proof into 4 steps.

Step 1. As in the RIF model, we start from the observation that the probability of coalescence can be decomposed approximately as in (A.6) where the term J is now 0 since the process is continuous and starts at 0. For the term I, the approximation (A.7) derived for RIF is also valid in the diffusion case, that is, one can show along the same lines that

$$\mathbb{P}^{\text{diff}}_{\text{coal2}} \approx \frac{1}{(2Ns)^A} I$$

where

$$I := \mathbb{E}\left[\int_0^\tau \frac{1}{Nsz_t} \exp(-\int_t^\tau \frac{du}{Nsz_u}) \left((2Ns)^A \mathbb{P}\left(L_{\tau-t} = 0 \mid z\right)\right)^2 du\right]$$

where z is solution of the Stochastic Differential Equation (3.4). As in RIF, this is justified by the observation that the probability of coalescence is well approximated by the probability of finding the two independent ancestral lineages in the 0 population at the first point of potential coalescence time in the point process $\hat{\mathcal{P}}_{0,0}$. We will not repeat the arguments here and refer the reader to Figure 9 for reference.

Step 2. Recall that for a fixed time t, the process Nsz_t is well approximated by the process y_t as defined in (3.7). As in (3.8), we define $\mathcal{E}$ as the stochastic correction to the exponential growth of the population process y

$$\lim_{u\to\infty} e^{-u} y_u = \frac{\mathcal{E}}{2} \quad \text{almost surely.} \tag{A.11}$$

By arguing analogously to the RIF model (see the reasoning yielding (A.5)), we can approximate Nsz_t by y_t in the previous integral and

$$\mathbb{P}(L_{\tau-t} = 0|z) \approx \left(\frac{\mathcal{E}e^t}{2Ns}\right)^A \frac{\pi A}{\sin(\pi A)}.$$

The arguments are completely analogous to the RIF case and will be skipped for the sake of conciseness. This yields

$$I \approx \left(\frac{\pi A}{\sin(\pi A)}\right)^2 \mathbb{E}\left[\int_0^\infty \frac{(e^t\mathcal{E})^{2A}}{y_t} \exp(-\int_t^\infty \frac{du}{y_u})dt\right] \tag{A.12}$$

Step 3. We are now left with the previous expectation whose integrand only depends on the population process y. Recall that y can be thought of as the continuum limit of a near-critical branching process conditioned on non-extinction. In order to compute the latter quantity, we will rely on the genealogical structure of the diffusion y that we now describe. To make this genealogical structure explicit, we start from a Yule process $\mathcal{Y}$ starting from a single individual with branching rate 1. We think of $\mathcal{Y}$ as describing the infinite lines of descent in the conditioned supercritical continuum branching process y. In Evans and O'Connell (1994), it is demonstrated that the process y can be constructed by allowing "mass" to emigrate from the ancestral lineages. A visual illustration of how the process y can be reconstructed from the ancestral

process $\mathcal{Y}$ is provided in Fig. 10. Intuitively, the mass that departs from the genealogy $\mathcal{Y}$ can be interpreted as the descendants of a particular infinite ancestral line.

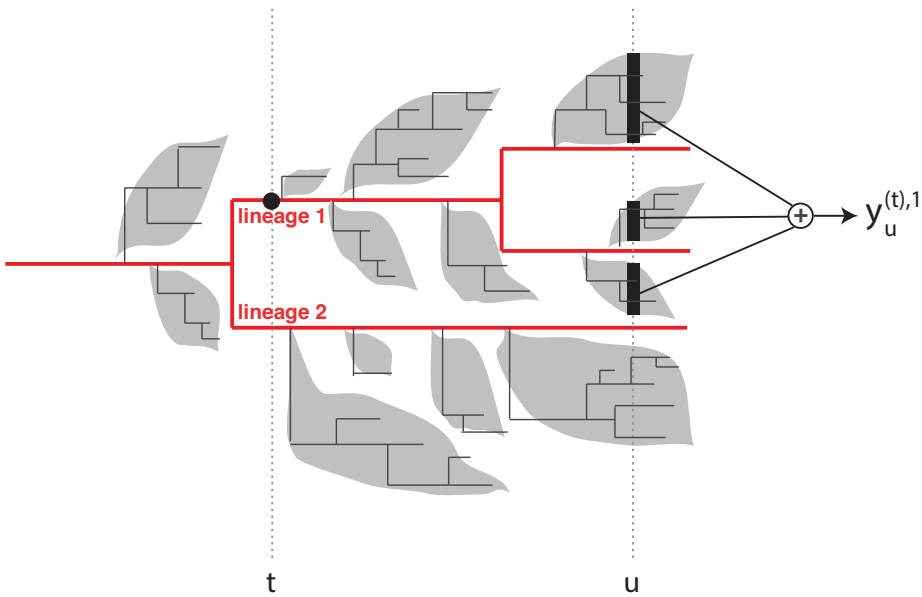

Figure 10: The Yule process is depicted in red. Mass is *emigrated* from the lineages, represented as grey bubbles. At the discrete level, these "mass" processes correspond to the contributions of subcritical branching processes grafted onto the immortal lineages. For each $t \geq 0$, the total population y_t is obtained by summing the contributions of all the grafted sub-populations. Finally, for a fixed time t, the quantity $y_u^{(t),i}$ denotes the contribution at time u of the ith immortal particle alive at time t. In the illustration, this is represented by the sum of all black intervals at time u.

In the following, we won't need to go into the details of the ancestral construction of y, but we will only rely on a few fact that we now briefly recall. Let L_t be the number of immortal lineages present at time t (that is the number of lineages present at time t in the Yule process $\mathcal{Y}$). For i^{th} immortal lineage present at time t, and for $u \geq t$, we let $y_u^{(t),i}$ be the mass emigrated from this lineage at time u. See again Fig. 10 for a discrete visualization.

By the Markov property, the time shifted process $(y_{u+t}^{(t),i}; u \geq 0)$ is identical in distribution to the original process y. According to (3.8), this implies the existence of a standard exponential random variable $\mathcal{E}^{(t),i}$ such that

$$\lim_{u\to\infty} e^{-(u-t)} y_u^{(t),i} = \frac{\mathcal{E}^{(t),i}}{2} \quad \text{a.s.} \tag{A.13}$$

Further, conditional on L_t, the random variables $\mathcal{E}^{(i),t}$ are independent. Finally, we note that (A.11) and (A.13) imply the identity

$$\mathcal{E} \;=\; e^{-t} \sum_{i=1}^{L_t} \mathcal{E}^{(t),i}. \tag{A.14}$$

Step 4. Having those facts at hands, we are now ready to compute the expeactation in (A.12). We consider the expected value

$$I' := \mathbb{E}\left[\int_0^\infty \frac{(e^t\mathcal{E})^{2A}}{y_t}\exp(-\int_t^\infty \frac{du}{y_u})du\right]$$

Let us denote by T_c the coalescing time of two lineages sampled at ∞. Since two lineages coalesce at a rate which is inversely proportional to the inverse population size, the previous quantity can be interpreted as

$$I' = \mathbb{E}\left[(\mathcal{E}e^{T_c})^{2A}\right].$$

(Formally, and as in noted in Lemma 4.5. in Etheridge *et al.* (2006), this is a variant of Perkin's desintegration result.) From Etheridge *et al.* (2006), L_{T_c+} has the following distribution

$$\forall k \geq 2, \quad \mathbb{P}(L_{T_c+} = k) \;=\; \frac{2}{k} - \frac{2}{k+1}.$$

Decomposing the expected value I' along the values for the number of immortal lineages right before coalescence, we have

$$I' = \sum_{k=2}^{\infty}\mathbb{E}\left[\left(\sum_{i=1}^{k}\mathcal{E}^{(T_c+),i}\right)^{2A}\right]\left(\frac{2}{k} - \frac{2}{k+1}\right).$$

From the previous discussion, the sum $\sum_{i=1}^{k}\mathcal{E}^{(T_c+),i}$ is distributed as a Gamma$(1,k)$. This yields

$$I' = \sum_{k=2}^{\infty}\frac{\Gamma(2A+k)}{\Gamma(k)}\left(\frac{2}{k} - \frac{2}{k+1}\right).$$

Next, a straightforward computation using properties of the Gamma functions shows that

$$\sum_{k=2}^{n}\frac{\Gamma(k+2A)}{\Gamma(k)}(\frac{2}{k} - \frac{2}{k+1}) \;=\; \frac{\Gamma(2(A+1))}{1-2A} \;-\; \frac{2\Gamma(2A+n+1)}{(1-2A)\Gamma(n+1)(n+1)}$$

where the second term on the right hand side goes to 0 for $n \to \infty$. The result then follows after combining the previous steps.