J. Math. Biol. (2011) 63:399–431
DOI 10.1007/s00285-010-0382-4

Mathematical Biology

Selective sweeps for recessive alleles and for other modes of dominance

G. Ewing · J. Hermisson · P. Pfaffelhuber · J. Rudolf

Received: 30 May 2010 / Revised: 13 October 2010 / Published online: 13 November 2010

Abstract A selective sweep describes the reduction of linked genetic variation due to strong positive selection. If s is the fitness advantage of a homozygote for the beneficial allele and h its dominance coefficient, it is usually assumed that $h = 1/2$, i.e. the beneficial allele is co-dominant. We complement existing theory for selective sweeps by assuming that h is any value in [0, 1]. We show that genetic diversity patterns under selective sweeps with strength s and dominance $0 < h < 1$ are similar to co-dominant sweeps with selection strength $2hs$. Moreover, we focus on the case $h = 0$ of a completely recessive beneficial allele. We find that the length of the sweep, i.e. the time from occurrence until fixation of the beneficial allele, is of the order of $\sqrt{N/s}$ generations, if N is the population size. Simulations as well as our results show that genetic diversity patterns in the recessive case $h = 0$ greatly differ from all other cases.

Keywords Genetic hitchhiking · Selective sweep · Beneficial mutation · Recessive allele · Genealogy

Mathematics Subject Classification (2000) 92D15 · 60J70 · 60K35

G. Ewing · J. Hermisson
University of Vienna, Vienna, Austria
e-mail: greg.ewing@univie.ac.at

J. Hermisson
e-mail: joachim.hermisson@univie.ac.at

P. Pfaffelhuber (✉) · J. Rudolf
Albert-Ludwigs University of Freiburg, Freiburg, Germany
e-mail: p.p@stochastik.uni-freiburg.de

1 Introduction

The model of selective sweeps (also called genetic hitchhiking) predicts a reduction in sequence diversity at a neutral locus closely linked to a beneficial allele (Maynard Smith and Haigh 1974). Most analysis of this model assumes that the beneficial allele is co-dominant. Accordingly, genome scans for the evidence of recent positive selection test a neutral model against (strong) selection for a co-dominant allele. Many such methods use information about the expected site frequency spectrum under a hitchhiking model to detect signatures of positive selection (e.g. Kim and Stephan 2002; Nielsen et al. 2005; Jensen et al. 2005). Simpler approaches use test statistics such as sample heterozygosity (usually called π), or Tajima's D to reject a standard neutral model. Simulation results by Teshima and Przeworski (2006) and Teshima et al. (2006) show that the false-discovery and false-negative rates of such methods increase if selection acts on a recessive rather than a co-dominant allele.

Although adaptations are often assumed to be rather dominant than recessive (Charlesworth 1998), also the case of recessive beneficial alleles is well documented in the empirical literature. Many cases that have been described concern resistance alleles. Here, the loss of a function of a gene conveys resistance to a pathogen. Often only the homozygous mutant is resistant, leading to a recessive sweep. Examples include: (i) The Duffy blood group locus in humans, where the homozygous null-allele (FY-0) confers complete resistance to vivax malaria. Hamblin and Di Rienzo (2000) report that the FY-0-genotype is at or near fixation in most sub-Saharan African populations but is very rare outside Africa, which suggests that it is locally under strong positive selection. (ii) Resistance to the yellow mosaic virus disease in barley has been mapped to several recessive resistance genes (Ordon et al. 2004). (iii) The plant gene eIF4E (present e.g. in pepper, pea and tomato) is a factor involved in basic cellular processes and can be used by viruses to complete their life cycle. Only if the function of both gene copies is compromised, the plant is resistant and positive selection can act (Cavatorta et al. 2008). (iv) The yellow fever mosquito *Aedes aegypti* is resistant to the drug permethrin, if both copies carry a replacement mutation in the gene para, as shown in Garcia et al. (2009).

Starting with the original publication by Maynard Smith and Haigh (1974), an extensive body of analytical theory has been established for the hitchhiking model (e.g. Kaplan et al. 1989; Stephan et al. 1992; Barton 1998; Durrett and Schweinsberg 2004; Etheridge et al. 2006). In addition to results on reduced diversity and the frequency spectrum, linkage disequilibrium has been studied by Stephan et al. (2006), McVean (2007), Pfaffelhuber and Studeny (2007) and Leocard and Pardoux (2010). Moreover, the model was extended to the case of multiple origins of the beneficial allele due to mutation or from standing genetic variation (Pennings and Hermisson 2006a,b; Hermisson and Pennings 2005; Hermisson and Pfaffelhuber 2008). All these results are built around the simplest possible scenario for adaptation, where positive selection acts on a single locus without dominance. Despite its empirical importance, dominance was only studied quite recently. Teshima and Przeworski (2006) use computer simulations to demonstrate the impact of intermediate dominance on the most important summary statistics for the frequency spectrum, see also Teshima et al. (2006). Explicit analytical results are even more sparse and only exist for the fixation time (duration of the sweep) of the beneficial allele (van Herwaarden and van der Wal 2002). The case of a completely recessive beneficial allele, in particular, has not been treated in any of these publications.

The goal of our investigation is the extension of previous analytical results to the case of arbitrary dominance. We focus, in particular, on the completely recessive case. While sweeps with intermediate dominance, $0 < h < 1$, and selection coefficient s produce diversity patterns similar to a co-dominant beneficial allele with selection coefficient $2hs$ (Theorem 4), our results show that recessive sweeps, $h = 0$, are qualitatively different (Theorem 2). Also, for the probability of fixation and the duration of the sweep, the recessive case is widely different from other modes of dominance. See Proposition 3.1 and Theorem 1 for the recessive case and Proposition 4.1 and Theorem 3 for $0 < h \leq 1$.

The paper is organized as follows: in Sect. 2 we introduce the model for selective sweeps with arbitrary dominance coefficient, both at the selected locus (Sect. 2.1) and at the neutral locus (Sect. 2.2). In Sect. 3, we give our results on sweeps of recessive alleles. Section 4 contains our results for sweeps in the cases $0 < h \leq 1$. In Sect. 5, we describe sequence diversity patterns under recessive sweeps using simulations and compare them with the case $0 < h < 1$. We conclude with the proofs in Sect. 6.

2 The model

We use discrete (Wright–Fisher) models as well as diffusion processes for modeling allelic frequency paths (Sect. 2.1). In order to study genetic diversity patterns, we use a structured coalescent (Sect. 2.2).

2.1 The allelic frequency path

Consider a one-locus Wright–Fisher model, consisting of N diploid (and hence $2N$ haploid) individuals. The beneficial mutant is B and the wildtype allele is denoted b. The (relative) fitness of diploids is given as follows:

Genotype	BB	Bb	bb
Relative fitness	$1+s$	$1+sh$	1

We are interested in the dynamics of $(X_t^N)_{t=0,1,2,\dots}$, where X_t^N is the frequency of the beneficial allele B in generation t. This process is a Markov chain and given $X_t^N = \frac{i}{2N}$, the transition probabilities are

$$\mathbb{P}[X_{t+1}^N = \tfrac{j}{2N} | X_t^N = \tfrac{i}{2N}] = \binom{2N}{j} \tilde{p}_i^j (1 - \tilde{p}_i)^{2N-j},$$

where

$$\tilde{p}_i = \frac{i^2(1+s) + i(2N-i)(1+sh)}{i^2(1+s) + 2i(2N-i)(1+sh) + (2N-i)^2}.$$

For $N \to \infty$, $s \to 0$ such that $2Ns \to \alpha$, and $X_0^N \Rightarrow X_0$, it is well-known (e.g. Ewens 2004) that $(X_{\lfloor Nt \rfloor}^N)_{t \geq 0} \Rightarrow (X_t)_{t \geq 0}$, where '$\Rightarrow$' means convergence in distribution (in the space of real-valued functions, equipped with uniform convergence on compacts), where $\mathcal{X} := (X_t)_{t \geq 0}$ is the diffusion, uniquely determined by the stochastic differential equation (SDE) given in (2.1).

Definition 2.1 (*Allelic frequency path*) The diffusion $\mathcal{X} = (X_t)_{t \geq 0}$ is the unique solution of the SDE

$$dX = \alpha(h + X(1 - 2h))X(1 - X)dt + \sqrt{X(1 - X)}dW, \tag{2.1}$$

where W is a standard Brownian motion. We set

$$T_0 := \inf\{t \geq 0 : X_t = 0\}, \qquad T_1 := \inf\{t \geq 0 : X_t = 1\}, \tag{2.2}$$

which are the times of loss and fixation of the beneficial allele, respectively. Moreover, $\mathcal{X}^* = (X_t^*)_{t \geq 0}$ is the process $\mathcal{X}$, conditioned on the event $\{T_1 < T_0\}$. We set

$$T^* := \inf\{t \geq 0 : X_t^* = 1\}, \tag{2.3}$$

which is the time of fixation of the beneficial allele. If not mentioned otherwise, we assume that $X_0 = 0$ and $\mathcal{X}^*$ arises as limit of conditioned processes which are started in ε as $\varepsilon \to 0$. For $\alpha, h \in \mathbb{R}$, we denote the distribution of $\mathcal{X}$, started in $X_0 = x$, by $\mathbb{P}_x^{\alpha,h}[.]$. Expectations and variances are denoted by $\mathbb{E}_x^{\alpha,h}[.]$ and by $\mathbb{V}_x^{\alpha,h}[.]$, respectively.

Remark 2.2 (Fixation probability for a single mutant) Let $\mathbb{P}_x^{N,s,h}[.]$ be the probability measure for the Wright–Fisher model with population size N, selection coefficient s and dominance h, started in $X_0^N = x$. Note that the weak convergence $(X_{\lfloor Nt \rfloor}^N)_{t \geq 0} \Rightarrow (X_t)_{t \geq 0}$ does not imply convergence for all interesting functionals of the Wright–Fisher model. In particular, convergence of fixation probabilities for a single mutant in the sense

$$\frac{\mathbb{P}_{1/(2N)}^{N,s_N,h}[X_\infty^N = 1]}{\mathbb{P}_{1/(2N)}^{\alpha,h}[X_\infty = 1]} \xrightarrow{N \to \infty} 1$$

with $Ns_N \xrightarrow{N \to \infty} \alpha$ has only been proved in the case $h = \frac{1}{2}$ (see Bürger and Ewens 1995, p. 565). For this reason, all our assertions are stated in the diffusion framework.

2.2 The structured coalescent

Consider a sample of size n, taken from the population at the time of fixation of the beneficial allele. From each of these individuals, consider homologous neutral loci, which are linked to the beneficial allele at recombination distance ρ (i.e. the probability that a recombination occurs between the beneficial and the neutral locus is r per

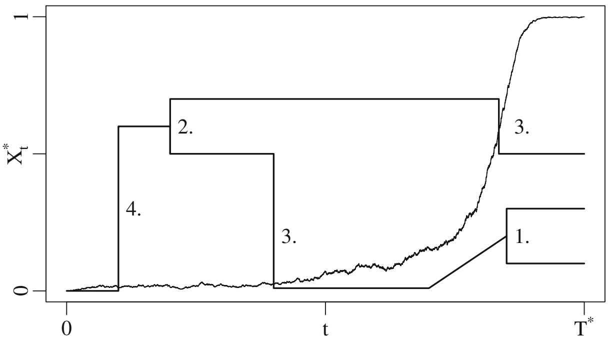

Fig. 1 Illustration of the structured coalescent in the case $h = 0$. Pairs of lines can coalesce within the beneficial background (event 1.), or within the wildtype background (event 2.). Changes from the beneficial to the wildtype background (event 3.) or vice versa (event 4.) may occur as well

generation in the Wright–Fisher model of diploid size N, and $2Nr \to \rho$ as $N \to \infty$). In order to study the genetic diversity within these neutral loci, we follow Kaplan et al. (1989) and Barton et al. (2004), and introduce the following coalescent process, which is conditioned on a path of $\mathcal{X}^*$.

Definition 2.3 (*The structured coalescent* $\mathcal{K}$) Let $\mathcal{X}^*$ be as in Definition 2.1. Conditioned on $\mathcal{X}^*$, we define a Markov process, $\mathcal{K} = (\mathcal{K}^b, \mathcal{K}^w) = (K^b_\beta, K^w_\beta)_{0\le\beta\le T^*}$ with $\beta = T^* - t$. Here, K^b_β and K^w_β are the number of lines in the beneficial and wild-type background at time β, respectively. Taking values in $\mathbb{Z}^2_+$, this process starts in $(K^b_0, K^w_0) = (n, 0)$ for some $n \in \mathbb{N}$. If $K_\beta = (k^b, k^w)$, then there are transitions to

1. $(k^b - 1, k^w)$ at rate $\binom{k^b}{2}\frac{1}{X^*_{T^*-\beta}}$
   (two lines in the beneficial background are merged to a single line)
2. $(k^b, k^w - 1)$ at rate $\binom{k^w}{2}\frac{1}{1-X^*_{T^*-\beta}}$
   (two lines in the wildtype background are merged to a single line)
3. $(k^b - 1, k^w + 1)$ at rate $k^b\rho(1 - X_{T^*-\beta})$
   (one line in the beneficial background changes to the wildtype background)
4. $(k^b + 1, k^w - 1)$ at rate $k^w\rho X_{T^*-\beta}$
   (one line in the wildtype background changes to the beneficial background).

See Fig. 1 for an illustration of $\mathcal{K}$.

3 Results on recessive alleles

In this section, we focus on the case $h = 0$, i.e. on properties of the diffusion $\mathcal{X} = (X_t)_{t\ge0}$ given by the SDE

$$dX = \alpha X^2(1 - X)dt + \sqrt{X(1 - X)}dW \tag{3.1}$$

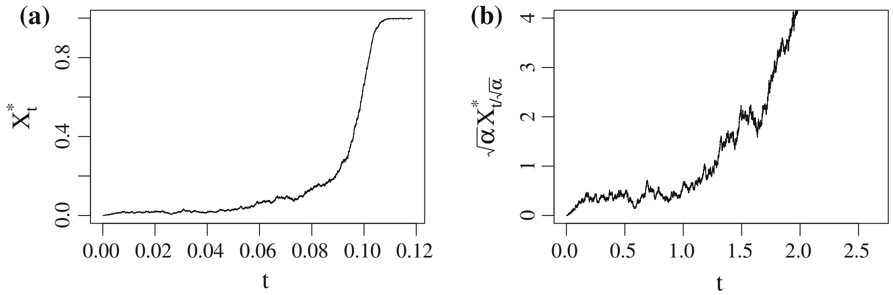

Fig. 2 **a** The frequency path of a beneficial recessive allele, conditioned on fixation, $\mathcal{X}^*$. We have used a Wright–Fisher model with $N = 5 \times 10^4$ and $s = 5 \times 10^{-3}$, i.e. $2Ns = 500$. **b** The rescaled frequency path $\left(\sqrt{\alpha}X^*_{t/\sqrt{\alpha}}\right)_{t\geq 0}$

and the corresponding diffusion $\mathcal{X}^*$, which is conditioned to hit 1 (recall from Definition 2.1). It is crucial to note that the process $\mathcal{X}^*$ spends most of its time at low frequencies for $h = 0$; see Fig. 2a. The reason is that for low frequency, most beneficial alleles are found in heterozygotes and selection can hence not be efficient. In order to make this statement more quantitative, we will show (see Lemma 6.3), that $(\sqrt{\alpha}X_{t/\sqrt{\alpha}})_{t\geq 0}$ converges to the diffusion $(Y_t)_{t\geq 0}$ given by $dY = Y^2dt + \sqrt{Y}dW$. This implies that $\mathcal{X}^*$ spends most of its time in frequencies of order $1/\sqrt{\alpha}$. See also Fig. 2b for an illustration of the process $(\sqrt{\alpha}X_{t/\sqrt{\alpha}})_{t\geq 0}$.

We give our main three results on fixation probability (Proposition 3.1)—actually already derived by Kimura (1962)—the duration of the recessive sweep (Theorem 1) and the structured coalescent (Theorem 2). All proofs are found in Sect. 6.1.

3.1 Fixation probability

The following is a classical fact, which can be read off from Kimura (1962), equation (14). It describes the fixation probability of the beneficial allele and is stated here for completeness.

Proposition 3.1 (Fixation probability) *Let $\mathcal{X}$ be as in Definition 2.1, $\alpha \to \infty$ and ε_α be such that $\varepsilon_\alpha\sqrt{\alpha} \to 0$. Then,*[1]

$$\mathbb{P}^{\alpha,h=0}_{\varepsilon_\alpha}[T_1 < T_0] \overset{\alpha\to\infty}{\approx} 2\varepsilon_\alpha\sqrt{\frac{\alpha}{\pi}}. \tag{3.2}$$

Remark 3.2 (Fixation probability in a finite population) Consider a finite population of diploid size N. Assume that $X_0 = \frac{1}{2N}$, meaning that there is only a single copy of

---

[1] For two functions a_α and b_α we write $a_\alpha \overset{\alpha\to\infty}{\approx} b_\alpha$ iff $\lim_{\alpha\to\infty} \frac{a_\alpha}{b_\alpha} = 1$.

the beneficial allele at time 0. For $\alpha = 2Ns$, we find that, as $N \to \infty$,

$$\mathbb{P}^{\alpha,h=0}_{1/(2N)}[T_1 < T_0] \approx \frac{1}{N}\sqrt{\frac{2Ns}{\pi}} = \sqrt{\frac{2s}{\pi N}}.$$

(This is exactly equation (15) in Kimura 1962). Hence, a new recessive beneficial allele has a chance to be fixed in the population, which is much larger than $1/2N$, its chance if it was neutral, and much smaller than $2hs$, its chance if it would not be recessive (compare with Remark 4.2). However, note that it is not shown that the fixation probability for a single copy of the beneficial allele in a Wright–Fisher model has the same limit behavior; compare with Remark 2.2.

3.2 Length of the recessive sweep

Now, we come to the results on the duration of the recessive sweep.

Theorem 1 (Length of the recessive sweep) *Let* T^* *be as in* (2.3). *Then,*

$$\mathbb{E}^{\alpha,h=0}_0[T^*] - \frac{4c_{cat}}{\sqrt{\pi}} \cdot \frac{1}{\sqrt{\alpha}} \overset{\alpha\to\infty}{\approx} \frac{3\log\alpha}{2\alpha}, \tag{3.3}$$

$$\mathbb{V}^{\alpha,h=0}_0[T^*] \overset{\alpha\to\infty}{\approx} \frac{c}{\alpha}, \tag{3.4}$$

where $c_{cat} \approx 0.916$ *is Catalan's constant and some* $0 < c < \infty$.

Remark 3.3 (Further investigation of T^*)

1. We find from (3.3) that

$$\mathbb{E}^{\alpha,h=0}_0[T^*] \approx \frac{2.067}{\sqrt{\alpha}} + \frac{3\log\alpha}{2\alpha}$$

   for large values of α. For a finite population of size N and a beneficial recessive allele with selection coefficient s, this means that it takes $2.067\sqrt{N/s} + 3\log(2Ns)/2s$ generations until fixation of the beneficial allele on average. As simulations show, this gives accurate numerical results; see Fig. 3a. In addition, a numerical investigation of (6.11) reveals that

$$\mathbb{V}^{\alpha,h=0}_0[T^*] \approx \frac{0.6362}{\alpha} \tag{3.5}$$

   for large α. This value also fits well with simulations; see Fig. 3b.
2. Our calculations reveal not only limits for the expected total duration of the sweep, but also of the duration of the allele being in low (below some $\varepsilon > 0$, not depending on α), intermediate (between ε and $1-\varepsilon$) and in high (above $1-\varepsilon$) frequency. These results are collected in Table 1.

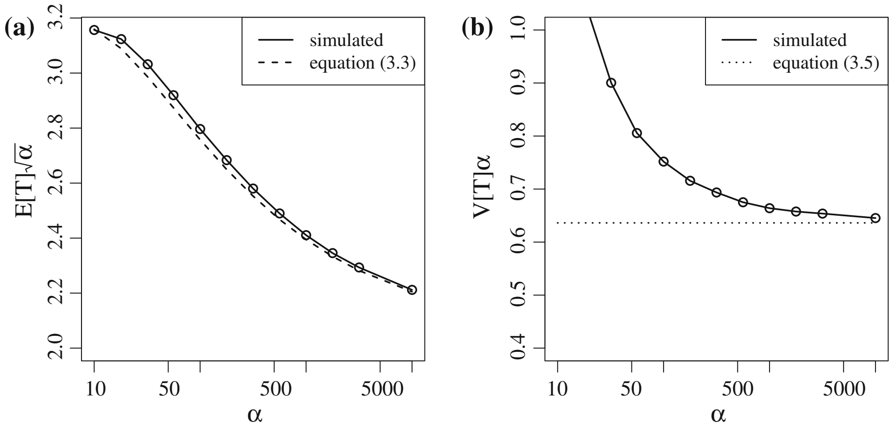

Fig. 3 The duration of the sweep of a beneficial recessive allele, is plotted against $\alpha = 2Ns$. We have used a Wright–Fisher model, with $N = 10^6$. **a** Average length of the duration of the recessive sweep. **b** Empirical variance of the duration of the recessive sweep

Table 1 The leading terms for the expected duration of the time when the frequency of the beneficial allele is low (i.e. smaller than some $\varepsilon > 0$), intermediate (between ε and $1-\varepsilon$) and high (above $1-\varepsilon$)

	$0 \le X_t \le \varepsilon$	$\varepsilon < X_t \le 1-\varepsilon$	$1-\varepsilon < X_t \le 1$
$h = 0$	$\frac{4c_{\text{cat}}}{\sqrt{\pi\alpha}} + \frac{\log\alpha}{2\alpha}$	$\mathcal{O}\left(\frac{1}{\alpha}\right)$	$\frac{\log\alpha}{\alpha}$
$0 < h < 1$	$\frac{\log\alpha}{h\alpha}$	$\mathcal{O}\left(\frac{1}{\alpha}\right)$	$\frac{\log\alpha}{(1-h)\alpha}$
$h = 1$	$\frac{\log\alpha}{\alpha}$	$\mathcal{O}\left(\frac{1}{\alpha}\right)$	$\frac{\pi^{3/2}}{2\sqrt{\alpha}} + \frac{\log\alpha}{2\alpha}$

In case the leading term can only be determined up to a factor, we write $\mathcal{O}(.)$

Remark 3.4 (Deterministic approach) In the co-dominant case, the fixation process is often approximated by a logistic increase of the beneficial allele; e.g. Kaplan et al. (1989) and Stephan et al. (1992). Formally, this approximation replaces the SDE (2.1) by the corresponding ODE. Calculating the duration of the sweep as the time to reach $1 - \frac{1}{2N}$ from $\frac{1}{2N}$ under the logistic results is a rough, but acceptable first approximation for the fixation time. This is different for the completely recessive case, where the corresponding ODE reads

$$\dot{y} = \alpha y^2(1-y), \qquad y_0 = \tfrac{1}{2N}.$$

The time t^* to reach $y_{t^*} = 1 - \frac{1}{2N}$ from an initial frequency $\frac{1}{2N}$ is easily derived to be

$$t^* = \frac{1}{\alpha}\left(\frac{4N(N-1)}{2N-1} + 2\log(2N-1)\right) \overset{N\to\infty}{\approx} \frac{2N}{\alpha}.$$

Note that the asymptotic behavior for $\alpha \to \infty$ of this expression differs considerably from the scaling $T^* \sim \alpha^{-1/2}$ in (3.3) that we have found from the full stochastic equation.

3.3 The structured coalescent

In this section, we study the structured coalescent from Sect. 2.2 for $h = 0$.

Theorem 2 (The structured coalescent for $h = 0$) *Let $\mathcal{K} = (\mathcal{K}^b_\beta, \mathcal{K}^w_\beta)_{0\leq\beta\leq T^*}$ be as in Definition 2.3, and $\rho = \rho_\alpha$ such that $\rho/\sqrt{\alpha} \to \lambda$ for some $0 < \lambda < \infty$ as $\alpha \to \infty$. Then,*

$$\mathbb{P}_0^{\alpha,h=0}[\mathcal{K}_{T^*} = (1,0)|\mathcal{K}_0 = (1,0)] \overset{\alpha\to\infty}{\approx} \mathbb{E}_0^{\alpha,h=0}[e^{-\lambda(\sqrt{\alpha}\cdot T^*)}], \tag{3.6}$$

$$\mathbb{P}_0^{\alpha,h=0}[K^b_{T^*} + K^w_{T^*} = k|\mathcal{K}_0 = (n,0)] \overset{\alpha\to\infty}{\approx} c_{\lambda,n,k} \tag{3.7}$$

for some $0 < c_{\lambda,n,k} < 1$, not depending on α.

Remark 3.5 (Approximation of the genealogy for $h = 0$) It is important to note that only the scaling of the recombination rate ρ by $\sqrt{\alpha}$ leads to a non-trivial limit result in (3.6) and (3.7). In applications, however, finite values of α and ρ must be assumed. In order to use the above Theorem for large α, we set $\lambda = \rho/\sqrt{\alpha}$. In this case, the Theorem implies that every single line changes from the beneficial to the wildtype background approximately at rate $\rho = \lambda\sqrt{\alpha}$—see (3.6)—during the sweep. A first naive guess about the genealogy at the neutral locus is a star-like approximation as in the case of co-dominant alleles (Durrett and Schweinsberg 2004, see also Sect. 4.3). This means that all lines recombine to the wild-type background (transition 3.) independently at rate ρ and those which did not recombine coalesce at the beginning of the sweep. However, $\mathcal{X}^*$ spends most time in frequencies of order $1/\sqrt{\alpha}$, leading to an increased rate of coalescence during the sweep. We suggest that the process $\mathcal{K}$ can be approximated by a coalescence process (which is not conditioned on $\mathcal{X}^*$) where two lines merge at rate $\sqrt{\alpha}$ and each line escapes the sweep at rate ρ. As we will see in Sect. 5, this approximation (in contrast to the star-like approximation) produces reasonable numerical results and helps to explain some of the qualitative differences in frequency spectra of recessive and co-dominant sweeps. Note, however, that the approximation is not a formal convergence result. (Due to the fluctuations in $\mathcal{X}^*$ around $1/\sqrt{\alpha}$ such a formal result is not easily obtained).

Remark 3.6 (Reduction of diversity and $c_{\lambda,2,2}$) The most prominent attribute of a selective sweep is the reduction in sequence diversity, which can be quantified by the average reduction in heterozygosity due to the sweep. Let $H_2(t)$ be the heterozygosity at a neutral locus linked to the selected site. Then (see equation (16) in Kaplan et al. 1989)

$$\mathbb{P}_0^{\alpha,h=0}[H_2(T^*)] = p_{22} \cdot \mathbb{P}[H_2(0)],$$

for

$$p_{22} := \mathbb{P}_0^{\alpha,h=0}[K^b_{T^*} + K^w_{T^*} = 2 | \mathcal{K}_0 = (2,0)] \approx c_{\lambda,2,2}.$$

Indeed, assuming that there are no new mutations in $[0, T^*]$, two neutral loci picked at the end of the selective sweep are different if the ancestors at the beginning of the sweep are different (probability p_{22}) and if these carry different alleles (probability $\mathbb{P}[H_2(0)]$). In particular, p_{22} captures the reduction of diversity within recessive sweeps. Using the approximate genealogy suggested in the last remark and a competing Poisson process argument, we find

$$p_{22} \approx \frac{2\rho}{\sqrt{\alpha} + 2\rho} \approx \frac{2\lambda}{1 + 2\lambda}. \tag{3.8}$$

We will see in Sect. 5 that this approximation produces a reasonable fit to simulations; see Fig. 6a. Note that the star-like approximation (described above) would lead to $p_{22} \approx \mathbb{E}[1 - \exp(-2\rho T^*)]$, and hence would predict a reduction in sequence diversity which is weaker than seen in simulations (not shown, but compare with Fig. 6a).

4 Results on incompletely and completely dominant alleles

In this section we give approximations for the fixation probability (Proposition 4.1) and the fixation time (Theorem 3) in the case $0 < h \leq 1$.

4.1 Fixation probability

The next result is the complement of Proposition 3.1 for the case $0 < h \leq 1$.

Proposition 4.1 (Fixation probability) *Let* $0 < h \leq 1$, *and* ε_α *be such that* $\varepsilon_\alpha \alpha \to 0$ *for* $\alpha \to \infty$. *Then,*

$$\mathbb{P}^{\alpha,h}_{\varepsilon_\alpha}[T_1 < T_0] \overset{\alpha\to\infty}{\approx} 2h\alpha\varepsilon_\alpha. \tag{4.1}$$

Remark 4.2 (Fixation probability in a finite population) For a finite population of haploid size $2N$, and if the homozygote for the beneficial allele has selective advantage $\alpha = 2Ns$, the above result means that

$$\mathbb{P}^{\alpha,h}_{1/(2N)}[T_1 < T_0] \approx 2hs.$$

Hence, the fixation probability of an incompletely dominant allele with dominance h and advantage s is approximately the same as for a co-dominant allele with advantage $2hs$.

4.2 Length of the sweep

Next, we give our results on the duration of the sweep for $0 < h \leq 1$.

Theorem 3 (Length of the sweep)

1. *Let* $0 < h < 1$ *and* T^* *be as in* (2.3). *Then,*

$$\mathbb{E}_0^{\alpha,h}[T^*] - \frac{\log(2\alpha)}{\alpha h(1-h)} \overset{\alpha\to\infty}{\approx} \frac{1}{\alpha h(1-h)} \times (\gamma + (2-3h)\log h + (3h-1)\log(1-h)) \tag{4.2}$$

where γ *is Euler's constant, and*

$$\mathbb{V}_0^{\alpha,h}[T^*] \overset{\alpha\to\infty}{\approx} \frac{1}{\alpha^2}\left(\frac{c'}{h^2} + \frac{c''}{(1-h)^2}\right)$$

for some $0 < c', c'' < \infty$, *not depending on* h.

2. *For* $h = 1$,

$$\mathbb{E}_0^{\alpha,h=1}[T^*] - \frac{\pi^{3/2}}{2\sqrt{\alpha}} \overset{\alpha\to\infty}{\approx} \frac{3\log\alpha}{2\alpha}. \tag{4.3}$$

and

$$\mathbb{V}_0^{\alpha,h=1}[T^*] \overset{\alpha\to\infty}{\approx} \frac{c'''}{\alpha}$$

for some $0 < c''' < \infty$.

Remark 4.3 (Further investigation of T^*)

1. The approximate expected duration of a sweep for $0 < h < 1$ from (4.2) has already been obtained in van Herwaarden and van der Wal (2002) (using other methods) and we only give these results for completeness. Comparison with numerical results (Fig. 4) show that (4.2) is accurate as long as $h\alpha \gg 1$ and $(1-h)\alpha \gg 1$, i.e. dominance is intermediate. For nearly recessive alleles ($h\alpha$ of order 1) or nearly dominant alleles ($(1-h)\alpha$ of order 1), however, T^* is much better approximated by the formulas for $h = 0$ (3.3) or $h = 1$ (4.3).
2. Although (4.2) is symmetric under the exchange of h and $(1-h)$, this does not reflect an exact identity of the fixation time. As noted by van Herwaarden and van der Wal (2002), an exact symmetry of the fixation process under the Wright–Fisher diffusion is obtained under the map $(\alpha, h) \mapsto (-\alpha, 1-h)$, i.e. if fixation of a dominant beneficial allele is compared with a recessive deleterious one. For beneficial alleles, the approximate symmetry under $h \mapsto (1-h)$ becomes much worse for nearly recessive or dominant alleles, as already observed by Teshima and Przeworski (2006). From our formulas for the fixation times of completely

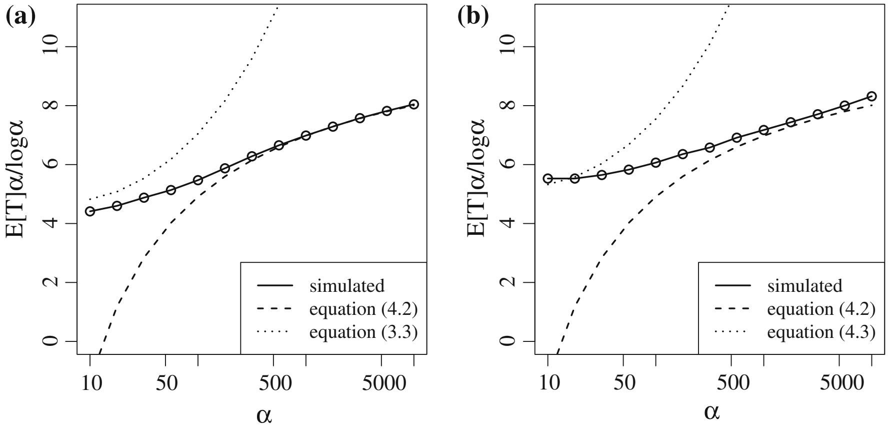

Fig. 4 The duration of the sweep of a beneficial allele is plotted against α. We have used the same parameters (except for dominance) as in Fig. 3. **a** $h = 0.1$, **b** $h = 0.9$. For $h\alpha < 10$ (in **a**, the nearly recessive case) and $(1-h)\alpha < 10$ (in **b**, the nearly dominant case), the simulation curve crosses over to the predicted sweep times for $h = 0$ (3.3) and $h = 1$ (4.3), respectively

recessive and dominant alleles, (3.3) and (4.3), this asymmetry is most obvious. In particular, the expected length of a dominant sweep ($h = 1$)

$$\mathbb{E}_0^{\alpha,h=1}[T^*] \approx \frac{2.7841}{\sqrt{\alpha}} + \frac{3\log\alpha}{2\alpha}.$$

is much longer than any sweep of an allele with the same homozygous advantage but $1 > h \geq 0$. Most of this time is spent near frequency 1. For the variance, numeric integration in the last line of (6.19) results in

$$\mathbb{V}_0^{\alpha,h=1}[T^*] \approx \frac{1.818}{\alpha}, \tag{4.4}$$

which is also larger than for $h = 0$. See Fig. 5 for comparison with simulations.

3. As for the recessive case, our calculations also reveal the expected times for three phases of the sweep, with the allele being in low, intermediate and high frequency (cf. Table 1).

4.3 The structured coalescent

The approximation of the coalescent at the neutral locus in a sweep region by a star-like genealogy is widely used. A formal convergence result for the co-dominant case is due to Durrett and Schweinsberg (2004). Using our results from Sect. 6, this result can be extended to the general non-recessive case with $0 < h < 1$.

Theorem 4 (The structured coalescent for $0 < h < 1$) *Let* $0 < h < 1$, $\mathcal{K} = (\mathcal{K}_\beta^b, \mathcal{K}_\beta^w)_{0\leq\beta\leq T^*}$ *be as in Definition* 2.3, *and* $\rho = \rho_\alpha$ *such that* $\rho\log\alpha/\alpha \to \lambda$

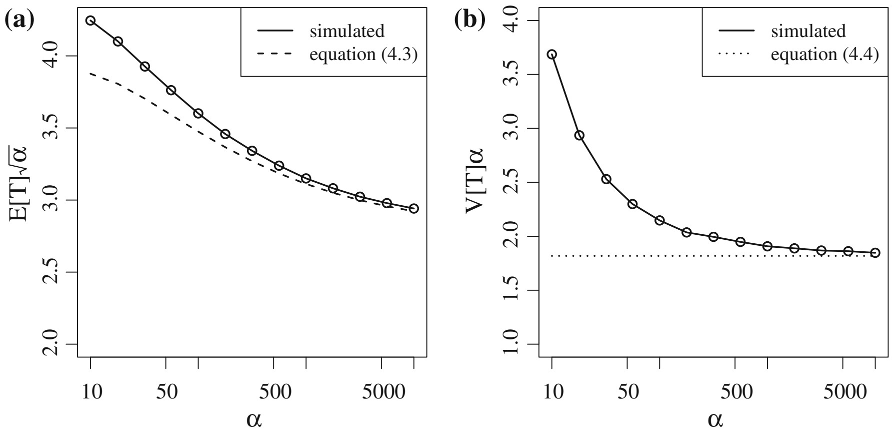

Fig. 5 The duration of the sweep of a beneficial dominant allele, $h = 1$, is plotted against $\sqrt{\alpha}$. We have used the same parameters (except for dominance) as in Fig. 3. **a** Expectation. **b** Variance

for some $0 < \lambda < \infty$ *as* $\alpha \to \infty$. *Then,*

$$\mathbb{P}^{\alpha,h}[\mathcal{K}_{T^*} = (1,0)|\mathcal{K}_0 = (1,0)] \overset{\alpha\to\infty}{\approx} e^{-\lambda/h}, \tag{4.5}$$

$$\mathbb{P}^{\alpha,h}[K^b_{T^*} + K^w_{T^*} = 2|\mathcal{K}_0 = (2,0)] \overset{\alpha\to\infty}{\approx} 1 - e^{-2\lambda/h}. \tag{4.6}$$

Remark 4.4 (Star-like genealogy, scaling and dominant alleles)

1. Again, it is crucial to note that only by the scaling of ρ by $\alpha/\log\alpha$ a non-trivial limit for the structured coalescent arises.
2. The Theorem must be read as follows: any single line recombines out of the sweep (i.e. experiences a transition 3., compare with Definition 2.3) with approximate probability $1-e^{-\lambda/h}$. Moreover, any two lines behave independently and coalesce if and only if both lines do not recombine. In particular, this interpretation shows the star-like approximation for the genealogy at the neutral locus.
3. It is now possible to compare the reduction in diversity of a beneficial allele with dominance h and selection coefficient α_h and a co-dominant beneficial allele with selection coefficient $\alpha_{1/2}$. As the Theorem shows, the reduction in diversity only depends on $\rho\log\alpha/(\alpha h)$. Hence, ignoring logarithmic terms, the diversity reduction is approximately the same if $2h\alpha_h = \alpha_{1/2}$. In other words, the effect of a beneficial allele with selection coefficient α and dominance h is similar as for a co-dominant one with selection coefficient $2h\alpha$.
4. We can approximate the expected reduction in heterozygosity at the end of a sweep in the case $0 < h < 1$. Using the same notation as in Remark 3.6, with $\rho\alpha/\log\alpha \to \lambda$ as in the above result, we find that

$$\begin{aligned}\mathbb{P}_0^{\alpha,h}[H_2(T^*)] &\approx \left(1 - e^{-2\lambda/h}\right) \cdot \mathbb{P}_0^{\alpha,h}[H_2(0)] \\ &\approx \left(1 - e^{-2\rho\log\alpha/(\alpha h)}\right) \cdot \mathbb{P}_0^{\alpha,h}[H_2(0)].\end{aligned} \tag{4.7}$$

5. The proof uses (and establishes) the fact that for $\alpha \to \infty$ recombination and coalescence can only happen while the frequency of the beneficial allele is small ($X_t < \epsilon$, first phase in Table 1). Note that the result uses the scaling of the recombination rate $\rho \sim \alpha/\log\alpha$. This is in stark contrast to the recessive case where ρ must be of order $\sqrt{\alpha}$.
6. The above theorem does not cover the case $h = 1$. For such a dominant sweep, recall from Theorem 3 that the duration of the sweep is of the order of $1/\sqrt{\alpha}$ in the diffusion time scale. However, this order is due to the final phase of the fixation process, where the frequency of the beneficial allele is near 1. During this final phase, it is possible that additional recombination events to the wild-type background occur, plus back-recombination events to the beneficial background. In particular, (i) in the proof of Theorem 4 is not true in the case $h = 1$.
7. Our result shows that methods, which rely on the star-like approximation (such as Nielsen et al. 2005), can easily be adapted to the case of intermediate dominance. In contrast, they are expected to fail for recessive sweeps, i.e. $h\alpha$ of order 1, even if selection is very strong. Note that the scaling result for the star-like approximation uses the limit $\alpha \to \infty$ for constant h and should therefore only be applied to cases where $h\alpha \gg 1$.

5 Simulations

In this section, we show simulation results indicating the implications of our findings for data analysis. All simulations were done using the program *msms* (Ewing and Hermisson 2010). This program is able to simulate samples of arbitrary size of homologous genetic material of any sequence length for arbitrary selection scenarios on a single bi-allelic locus, including temporally or spatially heterogeneous selection and dominance.

We concentrate our simulations on the reduction of sequence diversity at the neutral locus. In particular, we compare the effects of selective sweeps for recessive and co-dominant alleles. Since the same value for the selection strength α will lead to a much weaker genetic footprint for the recessive case, $h = 0$, as compared to $h = 0.5$, we give a heuristic argument in Sect. 5.1 that describes which values of α correspond to a footprint with a given strength for recessive and co-dominant alleles. The resulting sequence diversity is described in Sect. 5.2. A power analysis of Tajima's D, a common test statistic to reject the standard neutral model, is given Sect. 5.3.

5.1 Comparison of recessive and co-dominant sweeps

The difference of recessive and co-dominant sweeps is best seen from the approximations of the genealogy at the neutral locus; compare Remarks 3.6 and 4.4. In particular, we have seen that a star-like approximation can be used for strong selection and $h > 0$ (see (4.7)), while the case of $h = 0$ is best described by competing Poisson processes for recombination and coalescence with constant rates; see (3.8). Moreover, note that (4.7) and (3.8) already give first-order estimates for the reduction of heterozygosity due to a selective sweep.

We can ask, which selection strength α_0 for a recessive allele is needed to obtain the same expected reduction in heterozygosity as a co-dominant sweep with a given selection coefficient $\alpha_{1/2}$. Equating (3.8) and (4.7),

$$\frac{2\rho}{\sqrt{\alpha}_0 + 2\rho} = 1 - e^{-4\rho \log \alpha_{1/2}/\alpha_{1/2}},$$

we obtain the following condition for small ρ,

$$\alpha_0 \approx \left(\frac{\alpha_{1/2}}{2 \log \alpha_{1/2}}\right)^2. \tag{5.1}$$

We note that this relation produces sweeps of almost identical total length. Indeed, equating the leading order terms for the fixation times (3.3) and (4.3) leads to a selection coefficient for the recessive case of $\tilde{\alpha}_0 \approx 1.07\alpha_0$, with α_0 given in (5.1). In our simulations, we will use the particular pair $\alpha_{1/2} = 1000$ and $\alpha_0 = 5300$, which approximately fulfill (5.1).

5.2 The reduction in sequence diversity

A commonly used measure for sequence diversity is the nucleotide diversity, which we denote by $\widehat{\theta}_T$. For a sample of n homologous sequences of neutral loci,

$$\widehat{\theta}_T = \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} H_{ij}.$$

Here, H_{ij} is the number of differences between the ith and jth locus. Recall that $\widehat{\theta}_T$ is an unbiased estimator for the population mutation rate θ under the standard neutral model (Tajima 1983).

Note that $\mathbb{E}[\widehat{\theta}_T] = \mathbb{E}[H_2]$, where H_2 is the heterozygosity in the population. Hence, we can use (4.7) and (3.8) for a prediction of $\widehat{\theta}_T$ at the end of the selective sweep. We have tested these predictions using coalescent simulations under the infinite sites model of a stretch of DNA for a sample of size $n = 50$ for several pairs of selection coefficients according to (5.1). Results for the particular pair $\alpha_{1/2} = 1000$ and $\alpha_0 = 5300$ are shown in Fig. 6. The match for the reduction in heterozygosity is good as long as $\rho \ll \alpha$ (compare the slope for $\widehat{\theta}_T$ near $\rho = 0$ for $h = 0$ and $h = 0.5$). For larger values of ρ, there are several competing forces and all approximations show deviations. For $h = 0.5$, the star-like approximation overestimates the length of the genealogy and gives a too small valley of reduced heterozygosity. For $h = 0$, our Poisson process approximation assumes that we can neglect fluctuations of the frequency of the beneficial allele around $1/\sqrt{\alpha}$. The simulations show that a too broad valley of reduced heterozygosity is predicted (see Fig. 6a). Note that the star-like approximation for $h = 0.5$ and the Poisson process approximation for $h = 0$, give errors of comparable magnitude.

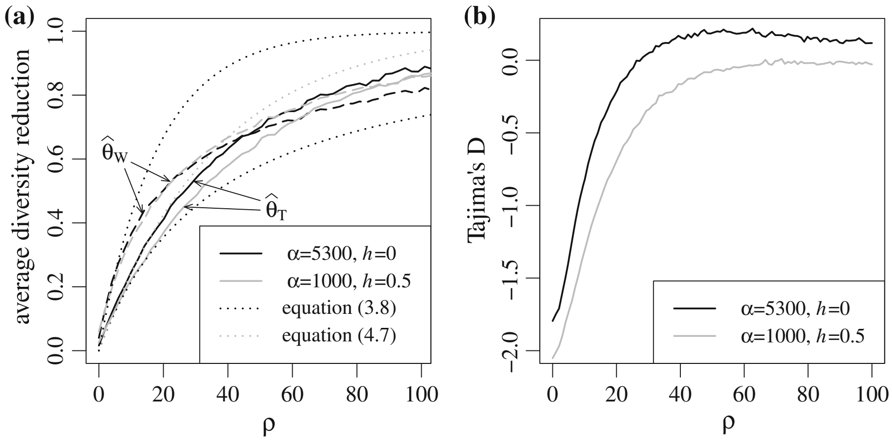

Fig. 6 **a** The average reduction of nucleotide diversity $\widehat{\theta}_T$ and Watterson's estimator $\widehat{\theta}_W$ for the population mutation rate θ after a sweep. **b** Tajima's D as a function of the recombination distance, ρ, for co-dominant and recessive sweeps

5.3 The power of Tajima's D

Another unbiased estimator for the population mutation rate θ is given by Watterson's θ,

$$\widehat{\theta}_W = \frac{S}{\sum_{i=1}^{n-1} \frac{1}{i}},$$

where S is the total number of single nucleotide polymorphisms (SNPs) in all n lines (Watterson 1975). In order to reject the neutral model, we consider the frequently used statistics Tajima's D (Tajima 1989). This statistics was one of the first test statistics for neutrality tests and is proportional to $\widehat{\theta}_T - \widehat{\theta}_W$. It is a classical result that Tajima's D is negative for selective sweeps.

The simulations show a more shallow increase for $\widehat{\theta}_W$ in the recessive case (see Fig. 6a). This implies a smaller difference of $\widehat{\theta}_W$ and $\widehat{\theta}_T$ and hence a less negative value for Tajima's D relative to sweep with $h = 0.5$, as seen in Fig. 6b. Also, this finding is easily understood from the approximated coalescent histories. With a star-like genealogy, only single lines of descent recombine out of the sweep, leading to an excess of singletons (more generally: low frequency polymorphisms) relative to the neutral expectation. This bias in the frequency spectrum is indicated by a shift to negative values of Tajima's D. In contrast, coalescence events will often occur prior to recombination events under a competing Poisson process scheme, valid for recessive sweeps. As a consequence, lines of descent that recombine out of the sweep will often have multiple descendants among the sequences in the sample. The shift towards low frequency polymorphisms is therefore less pronounced and Tajima's D less negative. Similar results have previously been reported for moderately recessive sweeps ($h = 0.1$) by Teshima and Przeworski (2006).

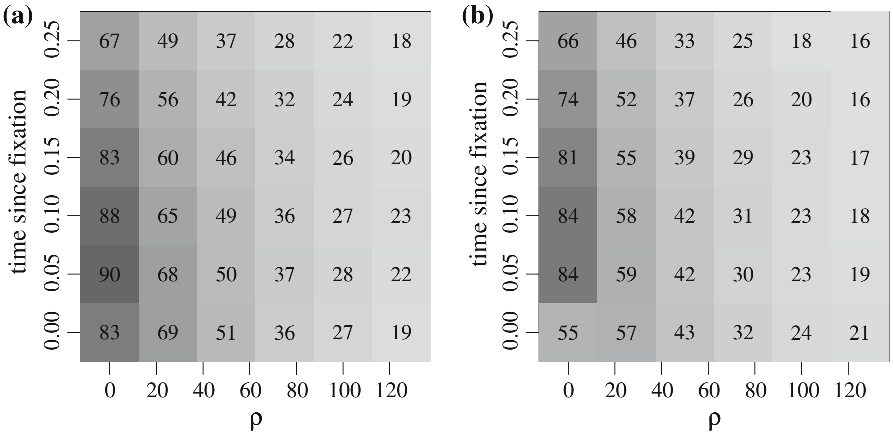

Fig. 7 The power of Tajima's D, depending on dominance, recombination distance and time since completion of the selective sweep. **a** Co-dominant and **b** recessive, parameters as in Fig. 6b

The approximate genealogy with constant coalescence and recombination rates is equivalent to a model of a structured population with constant migration among two islands representing the beneficial and the ancestral background. Depending on the migration rate (and the sampling scheme) both a surplus ($D < 0$) and a deficit ($D > 0$) of low frequency polymorphisms can result under this scenario. For the recessive sweep, we see that the expected Tajima's D indeed turns positive at a larger recombination distance to the selected site.

Since the deviation of Tajima's D from zero (in either direction) is relatively smaller for recessive sweeps, we expect that also the power to detect selection in a simple test based on D is reduced. This is confirmed by the power table in Fig. 7, consistent with previous results on partially recessive sweeps by Teshima et al. (2006).

6 Proofs

All our assertions are dealing with the case $\alpha \to \infty$. Hence, all '$\approx$' in our proofs are to be read as $\overset{\alpha\to\infty}{\approx}$. Frequently, we make use of a sequence $(c_\alpha)_{\alpha>0}$ of real numbers with $c_\alpha \to \infty$ slowly with $\alpha \to \infty$. For example, we write

$$\int_0^\alpha \frac{1-e^{-\xi}}{\xi} d\xi \approx \int_{c_\alpha}^\alpha \frac{1-e^{-\xi}}{\xi} d\xi \approx \int_{c_\alpha}^\alpha \frac{1}{\xi} d\xi \approx \log \alpha.$$

Our proofs are based on classical one-dimensional diffusion theory; see e.g. Karlin and Taylor (1981) and Ewens (2004). The solutions of the SDE as given in (2.1) have infinitesimal mean and variance

$$\mu(x) = \alpha(h + x(1-2h))x(1-x), \qquad \sigma^2(x) = x(1-x).$$

The scale function is

$$S(x) = \int_0^x \exp\left(-\int_0^\xi \frac{2\mu(\eta)}{\sigma^2(\eta)} d\eta\right) d\xi$$
$$= \int_0^x \exp\left(-2h\alpha\xi - \alpha\xi^2(1-2h)\right) d\xi. \tag{6.1}$$

Since 0 and 1 are exit boundaries for the diffusion, and for T_0 and T_1as defined in (2.2),

$$\mathbb{P}_x^{\alpha,h}[T_1 < T_0] = \frac{S(x)}{S(1)} = \frac{\int_0^x e^{-2h\alpha\eta-\alpha\eta^2(1-2h)} d\eta}{\int_0^1 e^{-2h\alpha\eta-\alpha\eta^2(1-2h)} d\eta}. \tag{6.2}$$

The Green function for X^*, the diffusion which is conditioned to hit 1 and started in x, is given by

$$G^*(x,\xi) = \begin{cases} \dfrac{2(S(1)-S(\xi))}{S(1)} \dfrac{S(\xi)}{\sigma^2(\xi)S'(\xi)}, & x \le \xi, \\[2ex] \dfrac{2(S(1)-S(x))S(\xi)}{S(1)} \dfrac{S(\xi)}{\sigma^2(\xi)S'(\xi)S(x)}, & x > \xi \end{cases}$$
$$= \begin{cases} 2\dfrac{\int_0^\xi e^{-2h\alpha\eta-\alpha\eta^2(1-2h)} d\eta \int_\xi^1 e^{-2h\alpha\eta-\alpha\eta^2(1-2h)} d\eta}{\xi(1-\xi)e^{-2h\alpha\xi-\alpha\xi^2(1-2h)} \int_0^1 e^{-2h\alpha\eta-\alpha\eta^2(1-2h)} d\eta}, & x \le \xi, \\[2ex] 2\dfrac{\int_x^1 e^{-2h\alpha\eta-\alpha\eta^2(1-2h)} d\eta}{\xi(1-\xi)e^{-2h\alpha\xi-\alpha\xi^2(1-2h)} \int_0^1 e^{-2h\alpha\eta-\alpha\eta^2(1-2h)} d\eta} \\ \qquad \cdot \dfrac{\left(\int_0^\xi e^{-2h\alpha\eta-\alpha\eta^2(1-2h)} d\eta\right)^2}{\int_0^x e^{-2h\alpha\eta-\alpha\eta^2(1-2h)} d\eta}, & x > \xi. \end{cases} \tag{6.3}$$

Here, $G^*(x,\xi)d\xi$ is the average time the conditioned diffusion spends in $(\xi, \xi + d\xi)$ before hitting 1. In particular, we will use that

$$\mathbb{E}_0[T^*] = \int_0^1 G^*(0,\xi) d\xi, \qquad \mathbb{V}_0[T^*] = 2\int_0^1 \int_0^\xi G^*(0,\xi)G^*(\xi,\eta) d\eta d\xi. \tag{6.4}$$

While the first identity is a classical result in diffusion theory, the second identity can e.g. be read off from Dawson et al. (2001, equation 2.1.1).

6.1 Proofs from Sect. 3

In this section, we fix $h = 0$. We will start with our key Lemma 6.3, which states that a scaling of the diffusion $\mathcal{X}$ immediately gives our result that the sweep length is of order $1/\sqrt{\alpha}$. Afterwards we provide proofs for Proposition 3.1, and Theorems 1 and 2.

A key lemma

In order to study the diffusion (3.1) and the structured coalescent, we use a rescaling argument. We need a definition first.

Definition 6.1 The diffusion $\mathcal{Y} = (\mathcal{Y}_\tau)_{\tau \geq 0}$, which takes values in $[0, \infty]$, is the unique solution of the SDE

$$dY = Y^2 d\tau + \sqrt{Y} dW,$$

for which $Y_\tau = \infty$ is a trap. We set

$$T_{\mathcal{Y}} := \inf\{\tau \geq 0 : Y_\tau = \infty\}.$$

The process $\mathcal{Y}^* = (Y^*_\tau)_{\tau \geq 0}$ is given as $\mathcal{Y}$, conditioned on the event $\{T_{\mathcal{Y}} < \infty\}$. We write $T^*_{\mathcal{Y}}$ for $T_{\mathcal{Y}}$, also conditioned on $\{T_{\mathcal{Y}} < \infty\}$.

Remark 6.2 (The diffusion $\mathcal{Y}^*$) The diffusion $\mathcal{Y}^*$ will play a crucial role in our proofs. Note that the scale function of this diffusion is

$$S^{\mathcal{Y}}(x) = \int_0^x e^{-\xi^2} d\xi.$$

For the Green function of $\mathcal{Y}^*$,

$$G^{\mathcal{Y},*}(x, \xi) = \begin{cases} \dfrac{4}{\sqrt{\pi}} \dfrac{\displaystyle\int_\xi^\infty e^{-\eta^2} d\eta \int_0^\xi e^{-\eta^2} d\eta}{\xi e^{-\xi^2}}, & x \leq \xi, \\ \dfrac{4}{\sqrt{\pi}} \dfrac{\displaystyle\int_x^\infty e^{-\eta^2} d\eta \left(\int_0^\xi e^{-\eta^2} d\eta\right)^2}{\xi e^{-\xi^2} \displaystyle\int_0^x e^{-\eta^2} d\eta}, & x > \xi. \end{cases} \tag{6.5}$$

Lemma 6.3 (Key lemma) *Let* T^*, $\mathcal{X}^*$ *be as in Definition 2.1 and* $T_{\mathcal{Y}}$, $T^*_{\mathcal{Y}}$, $\mathcal{Y}^*$ *be as in Definition 6.1, such that* $\mathcal{X}^*$ *and* $\mathcal{Y}^*$ *are started in* 0*. Then, for* $\widetilde{\mathcal{X}}^* := (\widetilde{X}^*_\tau)_{\tau\geq 0}$, $\widetilde{X}^*_\tau := \sqrt{\alpha}X^*_{\tau/\sqrt{\alpha}}$,

$$\widetilde{\mathcal{X}}^* \Rightarrow \mathcal{Y}^* \ \ \text{as } \alpha \to \infty. \tag{6.6}$$

Moreover,

$$\begin{aligned}\mathbb{E}_0^{\alpha,h=0}[T^*] - \tfrac{1}{\sqrt{\alpha}}\mathbb{E}_0[T^*_{\mathcal{Y}}] &\overset{\alpha\to\infty}{\approx} \frac{3\log\alpha}{2\alpha}, & (6.7)\\ \mathbb{E}_0[T^*_{\mathcal{Y}}] &= \frac{4}{\sqrt{\pi}}c_{cat}, & (6.8)\end{aligned}$$

where $c_{cat} \approx 0.916$ *is Catalan's constant. In particular,*

$$\sqrt{\alpha}\cdot T^* \Rightarrow T^*_{\mathcal{Y}} \qquad \text{as } \alpha \to \infty. \tag{6.9}$$

Proof For (6.6), we start by a change of variables

$$\widetilde{X}_t = \sqrt{\alpha}X_{t/\sqrt{\alpha}}.$$

By changing the time scale to $d\tau = \sqrt{\alpha}dt$, we obtain by Itô's formula that

$$\begin{aligned} d\widetilde{X} = \sqrt{\alpha}dX &= \sqrt{\alpha}\widetilde{X}^2(1-\widetilde{X}/\sqrt{\alpha})dt + \sqrt{\alpha\tfrac{\widetilde{X}}{\sqrt{\alpha}}(1-\widetilde{X}/\sqrt{\alpha})}dW\\ &= \widetilde{X}^2(1-\widetilde{X}/\sqrt{\alpha})d\tau + \sqrt{\widetilde{X}(1-\widetilde{X}/\sqrt{\alpha})}dW. \end{aligned}$$

Since $\alpha \to \infty$, we see that

$$\widetilde{\mathcal{X}} \Rightarrow \mathcal{Y} \quad \text{as } \alpha \to \infty$$

which also implies (6.6).

For (6.8), with a little help from MATHEMATICA to evaluate the last integral,

$$\mathbb{E}_0[T^*_{\mathcal{Y}}] = \int_0^\infty G^{\mathcal{Y},*}(0,\xi)d\xi = \frac{4}{\sqrt{\pi}}\int_0^\infty \frac{\int_\xi^\infty e^{-\eta^2}d\eta\int_0^\xi e^{-\eta^2}d\eta}{\xi e^{-\xi^2}}d\xi = \frac{4c_{cat}}{\sqrt{\pi}}. \tag{6.10}$$

For (6.7), recall the Green function for the diffusion $\mathcal{X}^*$ from (6.3). We use

$$\int_\xi^\infty e^{-\eta^2}d\eta \overset{\xi\to\infty}{\approx} \frac{e^{-\xi^2}}{2\xi}, \quad \int_0^\infty e^{-\eta^2}d\eta = \frac{\sqrt{\pi}}{2}, \quad \frac{1}{\xi(1-\xi/\sqrt{\alpha})} = \frac{1}{\xi} + \frac{1}{\sqrt{\alpha}-\xi}$$

and (6.10) to obtain

$$\begin{aligned}\mathbb{E}_0^{\alpha,h=0}[T^*] &= \int_0^1 G^{\mathcal{X},*}(0,\xi)d\xi \approx \frac{4}{\sqrt{\pi\alpha}} \int_0^{\sqrt{\alpha}} \frac{\int_\xi^{\sqrt{\alpha}} e^{-\eta^2}d\eta \int_0^\xi e^{-\eta^2}d\eta}{\xi(1-\xi/\sqrt{\alpha})e^{-\xi^2}}d\xi \\ &\approx \tfrac{1}{\sqrt{\alpha}}\mathbb{E}_0[T^*_{\mathcal{Y}}] + \tfrac{4}{\sqrt{\pi\alpha}}\left(\Delta_1+\Delta_2+\Delta_3+\Delta_4\right)\end{aligned}$$

with (recall the sequence c_α going to infinity slowly enough)

$$\begin{aligned}
\Delta_1 &:= -\int_{\sqrt{\alpha}}^{\infty} \frac{\int_\xi^{\infty} e^{-\eta^2}d\eta \int_0^\xi e^{-\eta^2}d\eta}{\xi e^{-\xi^2}}d\xi \approx -\tfrac{\sqrt{\pi}}{4}\int_{\sqrt{\alpha}}^{\infty}\frac{1}{\xi^2}d\xi = -\tfrac{\sqrt{\pi}}{4\sqrt{\alpha}}, \\
\Delta_2 &:= -\int_0^{\sqrt{\alpha}} \frac{\int_{\sqrt{\alpha}}^{\infty} e^{-\eta^2}d\eta \int_0^\xi e^{-\eta^2}d\eta}{\xi e^{-\xi^2}}d\xi \approx -\tfrac{1}{2\sqrt{\alpha}}\int_0^{\sqrt{\alpha}} \frac{e^{\xi^2-\alpha}\int_0^\xi e^{-\eta^2}d\eta}{\xi}d\xi \\
&\approx -\tfrac{1}{2\sqrt{\alpha}}\int_{\sqrt{\alpha}/2}^{\sqrt{\alpha}} \frac{e^{\xi^2-\alpha}\int_0^\xi e^{-\eta^2}d\eta}{\xi}d\xi \approx -\tfrac{\sqrt{\pi}}{4\sqrt{\alpha}}e^{-\alpha}\int_{\sqrt{\alpha}/2}^{\sqrt{\alpha}}\frac{e^{\xi^2}}{\xi}d\xi \approx -\frac{\sqrt{\pi}}{8\alpha^{3/2}}, \\
\Delta_3 &:= \int_0^{\sqrt{\alpha}-c_\alpha} \frac{\int_\xi^{\sqrt{\alpha}} e^{-\eta^2}d\eta \int_0^\xi e^{-\eta^2}d\eta}{(\sqrt{\alpha}-\xi)e^{-\xi^2}}d\xi \approx \int_{c_\alpha}^{\sqrt{\alpha}-c_\alpha} \frac{\int_\xi^{\infty} e^{-\eta^2}d\eta \int_0^\xi e^{-\eta^2}d\eta}{(\sqrt{\alpha}-\xi)e^{-\xi^2}}d\xi \\
&\approx \tfrac{\sqrt{\pi}}{4}\int_{c_\alpha}^{\sqrt{\alpha}-c_\alpha}\frac{1}{\xi(\sqrt{\alpha}-\xi)}d\xi = \tfrac{\sqrt{\pi}}{4\sqrt{\alpha}}\int_{c_\alpha}^{\sqrt{\alpha}-c_\alpha}\frac{1}{\xi}+\frac{1}{\sqrt{\alpha}-\xi}d\xi \\
&\approx \tfrac{\sqrt{\pi}}{2\sqrt{\alpha}}\log(\sqrt{\alpha}) = \tfrac{\sqrt{\pi}}{4\sqrt{\alpha}}\log\alpha, \\
\Delta_4 &:= \int_{\sqrt{\alpha}-c_\alpha}^{\sqrt{\alpha}} \frac{\int_\xi^{\sqrt{\alpha}} e^{-\eta^2}d\eta \int_0^\xi e^{-\eta^2}d\eta}{(\sqrt{\alpha}-\xi)e^{-\xi^2}}d\xi \approx \tfrac{\sqrt{\pi}}{2}\int_{\sqrt{\alpha}-c_\alpha}^{\sqrt{\alpha}} \frac{\int_\xi^{\sqrt{\alpha}} e^{-(\eta-\xi)(\eta+\xi)}d\eta}{\sqrt{\alpha}-\xi}d\xi \\
&\approx \tfrac{\sqrt{\pi}}{2}\int_{\sqrt{\alpha}-c_\alpha}^{\sqrt{\alpha}} \frac{\int_\xi^{\sqrt{\alpha}} e^{-(\eta-\xi)2\sqrt{\alpha}}d\eta}{\sqrt{\alpha}-\xi}d\xi = \tfrac{\sqrt{\pi}}{4\sqrt{\alpha}}\int_{\sqrt{\alpha}-c_\alpha}^{\sqrt{\alpha}} \frac{1-e^{-(\sqrt{\alpha}-\xi)2\sqrt{\alpha}}}{\sqrt{\alpha}-\xi}d\xi \\
&= \tfrac{\sqrt{\pi}}{4\sqrt{\alpha}}\int_0^{c_\alpha}\frac{1-e^{-\xi 2\sqrt{\alpha}}}{\xi}d\xi = \tfrac{\sqrt{\pi}}{4\sqrt{\alpha}}\int_0^{\sqrt{\alpha}c_\alpha}\frac{1-e^{-2\xi}}{\xi}d\xi \approx \tfrac{\sqrt{\pi}}{8\sqrt{\alpha}}\log\alpha.
\end{aligned}$$

The result (6.7) follows.

Proof of Proposition 3.1 Due to Lemma 6.3, we can compute, by classical diffusion theory,

$$\begin{aligned}\mathbb{P}^{\alpha,h}_{\varepsilon_\alpha}[T_1 < T_0] &\approx \mathbb{P}_{\sqrt{\alpha}\cdot\varepsilon_\alpha}[T_{\mathcal{Y}} < \infty] = \frac{S^{\mathcal{Y}}(\sqrt{\alpha}\cdot\varepsilon_\alpha)}{S^{\mathcal{Y}}(\infty)}\\ &= \frac{\int_0^{\sqrt{\alpha}\cdot\varepsilon_\alpha} e^{-\eta^2}d\eta}{\int_0^\infty e^{-\eta^2}d\eta} \approx \sqrt{\alpha}\cdot\varepsilon_\alpha\cdot\frac{2}{\sqrt{\pi}}\end{aligned}$$

and we are done.

Proof of Theorem 1 Clearly, (3.3) is a combination of (6.7) and (6.8). For (3.4) we use (6.4) and write

$$\begin{aligned}\alpha\mathbb{V}^{\alpha,h=0}_0[T^*] &= 8\alpha\int_0^1\int_0^\xi \frac{\left(\int_\xi^1 e^{-\alpha\zeta^2}d\zeta\right)^2\left(\int_0^\eta e^{-\alpha\zeta^2}d\zeta\right)^2}{\left(\int_0^1 e^{-\alpha\zeta^2}d\zeta\right)^2\xi(1-\xi)\eta(1-\eta)e^{-\alpha\xi^2}e^{-\alpha\eta^2}}d\eta d\xi\\ &\approx 8\int_0^\infty\int_0^\xi \frac{\left(\int_\xi^\infty e^{-\zeta^2}d\zeta\right)^2\left(\int_0^\eta e^{-\zeta^2}d\zeta\right)^2}{\left(\int_0^\infty e^{-\zeta^2}d\zeta\right)^2\xi\eta e^{-\xi^2}e^{-\eta^2}}d\eta d\xi < \infty. \qquad (6.11)\end{aligned}$$

Proof of Theorem 2 We use the same notation as in Lemma 6.3 (in particular $\sqrt{\alpha}dt = d\tau$) and transform the structured coalescent to a structured coalescent conditioned on the process $\mathcal{Y}$. For (3.7), note that the coalescence rate at time t is given by

$$\frac{1}{X_t}dt = \frac{\sqrt{\alpha}}{\sqrt{\alpha}\widetilde{X}_\tau}d\tau \Rightarrow \frac{1}{Y_\tau}d\tau \text{ as } \alpha\to\infty$$

Recombinations occur at rate

$$\rho(1-X_t)dt = \frac{\rho}{\sqrt{\alpha}}\left(1-\frac{\widetilde{X}_\tau}{\sqrt{\alpha}}\right)d\tau \approx \lambda. \qquad (6.12)$$

For back recombination from the wild-type to the beneficial background,

$$\rho X_t dt = \frac{\rho}{\sqrt{\alpha}}\frac{\widetilde{X}_\tau}{\sqrt{\alpha}}d\tau \approx 0.$$

All three limits are valid for all $0 \le \tau < T^*_{\mathcal{Y}}$.

So, for the structured coalescent conditioned on $\mathcal{Y}$, we find that the coalescence rate is $1/Y_\tau d\tau$, and the recombination rate is λ. During times when Y_τ is of order 1, we see that both, coalescence and recombination happen at rate of order 1. In particular, we have shown (3.7). For (3.6), since back recombination can be ignored, by (6.12),

$$\mathbb{P}^{\alpha,h=0}_0[\mathcal{K}_{T^*} = (1,0)|\mathcal{K}_0 = (1,0)] = \mathbb{E}^{\alpha,h=0}_0[e^{-\lambda(\sqrt{\alpha}\cdot T^*)}].$$

6.2 Proofs from Sect. 4

We directly use (6.2) in order to prove Proposition 4.1, and (6.4) in order to prove Theorem 3. We will use several approximation results on integrals appearing in these equations. We will use

$$\begin{aligned}\int_0^\alpha e^{\xi^2}d\xi - \frac{e^{\alpha^2}}{2\alpha} &= e^{\alpha^2}\left(\int_0^\alpha e^{-2\xi\alpha}(e^{\xi^2}-1)d\xi - \int_\alpha^\infty e^{-2\xi\alpha}d\xi\right)\\ &\approx e^{\alpha^2}\int_0^\alpha \xi^2 e^{-2\alpha\xi}d\xi \approx \frac{e^{\alpha^2}}{4\alpha^3}, \qquad (6.13)\end{aligned}$$

as well as

$$\begin{aligned}\int_0^\xi e^{-2h\eta+(2h-1)\eta^2/\alpha}d\eta - \frac{1}{2h}(1-e^{-2h\xi}) &= \int_0^\xi e^{-2h\eta}(e^{(2h-1)\eta^2/\alpha}-1)d\eta\\ \approx \frac{2h-1}{\alpha}\int_0^\xi e^{-2h\eta}\eta^2 d\eta &\approx \frac{c(1-e^{-2h\xi})+de^{-2h\xi}}{\alpha} \qquad (6.14)\end{aligned}$$

for $\xi \leq \alpha$ and fixed $0 < h \leq 1$, for some $0 \leq c, d < \infty$. The last equation implies for $\xi = \alpha$

$$\int_0^\alpha e^{-2h\eta+(2h-1)\eta^2/\alpha}d\eta - \frac{1}{2h} \approx \frac{2h-1}{4h^3\alpha}. \qquad (6.15)$$

Proof of Proposition 4.1 For $0 < h \leq 1$, we write, using (6.2) and (6.15),

$$\mathbb{P}^{\alpha,h}_{\varepsilon_\alpha}[T_1 < T_0] = \frac{\int_0^{\alpha\varepsilon_\alpha} e^{-2h\eta+(2h-1)\eta^2/\alpha}d\eta}{\int_0^\alpha e^{-2h\eta+(2h-1)\eta^2/\alpha}d\eta} \approx 2h\alpha\varepsilon_\alpha.$$

Proof of Theorem 3 1. For $0 < h < 1$, we write with (6.15)

$$\begin{aligned}\mathbb{E}^{\alpha,h}_0[T^*] &= 2\int_0^\alpha \frac{\int_0^\xi e^{-2h\eta+(2h-1)\eta^2/\alpha}d\eta \int_\xi^\alpha e^{-2h\eta+(2h-1)\eta^2/\alpha}d\eta}{\xi(\alpha-\xi)e^{-2h\xi+(2h-1)\xi^2/\alpha}\int_0^\alpha e^{-2h\eta+(2h-1)\eta^2/\alpha}d\eta}d\xi\\ &\approx \frac{4h}{\alpha}(A_1 + A_2) \qquad (6.16)\end{aligned}$$

with

$$A_1 := \int\limits_0^\alpha \frac{\int_0^\xi e^{-2h\eta+(2h-1)\eta^2/\alpha}d\eta \int_\xi^\alpha e^{-2h\eta+(2h-1)\eta^2/\alpha}d\eta}{\xi e^{-2h\xi+(2h-1)\xi^2/\alpha}},$$

$$A_2 := \int\limits_0^\alpha \frac{\int_0^\xi e^{-2h\eta+(2h-1)\eta^2/\alpha}d\eta \int_\xi^\alpha e^{-2h\eta+(2h-1)\eta^2/\alpha}d\eta}{(\alpha-\xi) e^{-2h\xi+(2h-1)\xi^2/\alpha}}.$$

We use

$$\int\limits_0^\xi \frac{1-e^{-\eta}}{\eta}d\eta - \log\xi \approx \gamma$$

for $\xi \to \infty$ and, if $0 \le \xi \le \alpha$,

$$\begin{aligned}
&\frac{\int_\xi^\alpha e^{-2h\eta+(2h-1)\eta^2/\alpha}d\eta}{e^{-2h\xi+(2h-1)\xi^2/\alpha}} - \frac{1-e^{-(2h-(2h-1)2\xi/\alpha)(\alpha-\xi)}}{2h-2(2h-1)\xi/\alpha} \\
&= \int\limits_0^{\alpha-\xi} e^{-2h\eta+(2h-1)(2\eta\xi+\eta^2)/\alpha}d\eta - \int\limits_0^{\alpha-\xi} e^{-(2h-(2h-1)2\xi/\alpha)\eta}d\eta \\
&= \int\limits_0^{\alpha-\xi} e^{-(2h-2(2h-1)\xi/\alpha)\eta}\left(e^{(2h-1)\eta^2/\alpha}-1\right)d\eta \\
&\approx \frac{2h-1}{\alpha}\int\limits_0^{\alpha-\xi} e^{-(2h-2(2h-1)\xi/\alpha)\eta}\eta^2 d\eta \\
&= \frac{c(1-e^{-(2h-2(2h-1)\xi/\alpha)(\alpha-\xi)}) + de^{-(2h-2(2h-1)\xi/\alpha)(\alpha-\xi)}}{\alpha}
\end{aligned} \tag{6.17}$$

for some $c, d \ge 0$, not depending on α, in order to obtain

$$\begin{aligned}
&4hA_1 - \frac{1}{h}(\log(2h\alpha)+\gamma) \\
&\approx 4h\int\limits_0^\alpha \frac{\int_0^\xi e^{-2h\eta+(2h-1)\eta^2/\alpha}d\eta \int_\xi^\alpha e^{-2h\eta+(2h-1)\eta^2/\alpha}d\eta}{\xi e^{-2h\xi+(2h-1)\xi^2/\alpha}}d\xi \\
&-\frac{1}{h}\int\limits_0^{2h\alpha}\frac{1-e^{-\xi}}{\xi}d\xi
\end{aligned}$$

$$
\begin{aligned}
&\approx \frac{1}{h}\left(\int_0^\alpha \frac{1-e^{-2h\xi}}{\xi}\frac{2h}{2h-2(2h-1)\xi/\alpha}d\xi - \int_0^\alpha \frac{1-e^{-2h\xi}}{\xi}d\xi\right)\\
&= \frac{1}{h}\int_0^\alpha (1-e^{-2h\xi})\frac{2(2h-1)/\alpha}{2h-2(2h-1)\xi/\alpha}d\xi\\
&\approx \frac{1}{h}\log\left(2h-2(2h-1)\xi/\alpha\right)\Big|_{\xi=\alpha}^{\xi=0}\\
&= \frac{1}{h}\left(\log(2h)-\log(2(1-h))\right).
\end{aligned}
\tag{6.18}
$$

Here, we have used (6.14) and (6.17) for the second $\approx$. In addition,

$$
\begin{aligned}
&4hA_2 - \frac{1}{1-h}(\log(2(1-h)\alpha)+\gamma)\\
&\approx 4h\int_0^\alpha \frac{\int_0^\xi e^{-2h\eta+(2h-1)\eta^2/\alpha}d\eta \int_\xi^\alpha e^{-2h\eta+(2h-1)\eta^2/\alpha}d\eta}{(\alpha-\xi)e^{-2h\xi+(2h-1)\xi^2/\alpha}}d\xi\\
&\quad - \frac{1}{1-h}\int_0^{2(1-h)\alpha} \frac{1-e^{-\xi}}{\xi}d\xi\\
&\approx \frac{1}{1-h}\left(\int_0^\alpha \frac{1-e^{-(2h-(2h-1)2\xi/\alpha)(\alpha-\xi)}}{\alpha-\xi}\frac{2(1-h)}{2h-2(2h-1)\xi/\alpha}d\xi\right.\\
&\quad \left. - \int_0^\alpha \frac{1-e^{-2(1-h)\xi}}{\xi}d\xi\right)\\
&\overset{\xi\to\alpha-\xi}{\approx} \frac{1}{1-h}\left(\int_0^\alpha \frac{1-e^{-2(1-h)\xi}}{\xi}\frac{2(1-h)}{2(1-h)-2(2(1-h)-1)\xi/\alpha}d\xi\right.\\
&\quad \left. - \int_0^\alpha \frac{1-e^{-2(1-h)\xi}}{\xi}d\xi\right)
\end{aligned}
$$

The last line is exactly line (6.18) with h replaced by $1-h$, and hence we obtain

$$
\begin{aligned}
\mathbb{E}_0^{\alpha,h}[T^*] - \frac{\log(2\alpha)}{\alpha h(1-h)} &\approx \tfrac{1}{\alpha}\Big(\tfrac{1}{h}(\gamma + 2\log h - \log(1-h))\\
&\quad + \tfrac{1}{1-h}(\gamma + 2\log(1-h) - \log h)\Big)\\
&= \frac{1}{\alpha h(1-h)}\left(\gamma + (2-3h)\log h + (3h-1)\log(1-h)\right).
\end{aligned}
$$

The variance is computed by (6.4) and

$$\begin{aligned}
\mathbb{V}_0^{\alpha,h}[T^*] &= 8\int_0^1\int_0^\xi \frac{\left(\int_\xi^1 e^{-2h\alpha\zeta+(2h-1)\alpha\zeta^2}d\zeta\right)^2}{\xi(1-\xi)\eta(1-\eta)e^{-2h\alpha\xi+(2h-1)\alpha\xi^2}e^{-2h\alpha\eta+(2h-1)\alpha\eta^2}}\\
&\quad\times\frac{\left(\int_0^\eta e^{-2h\alpha\zeta+(2h-1)\alpha\zeta^2}d\zeta\right)^2}{\left(\int_0^1 e^{-2h\alpha\zeta+(2h-1)\alpha\zeta^2}d\zeta\right)^2}d\eta d\xi\\
&= 8\int_0^\alpha\int_0^\xi \frac{\left(\int_\xi^\alpha e^{-2h\zeta+(2h-1)\zeta^2/\alpha}d\zeta\right)^2}{\xi(\alpha-\xi)\eta(\alpha-\eta)e^{-2h\xi+(2h-1)\xi^2/\alpha}e^{-2h\eta+(2h-1)\eta^2/\alpha}}\\
&\quad\times\frac{\left(\int_0^\eta e^{-2h\zeta+(2h-1)\zeta^2/\alpha}d\zeta\right)^2}{\left(\int_0^\alpha e^{-2h\zeta+(2h-1)\zeta^2/\alpha}d\zeta\right)^2}d\eta d\xi\\
&\approx \frac{8}{\alpha^2}\left(\int_0^{c_\alpha}\int_0^\xi \frac{e^{-4h\xi}\left(\int_0^\eta e^{-2h\zeta}d\zeta\right)^2}{\xi\eta e^{-2h\xi}e^{-2h\eta}}d\eta d\xi\right.\\
&\quad+\int_{\alpha-c_\alpha}^\alpha\int_{\alpha-c_\alpha}^\xi \frac{1}{(\alpha-\xi)(\alpha-\eta)e^{-2h\xi+(2h-1)\xi^2/\alpha}}\\
&\quad\left.\times\frac{\left(\int_\xi^\alpha e^{-2h\zeta+(2h-1)\zeta^2/\alpha}d\zeta\right)^2}{e^{-2h\eta+(2h-1)\eta^2/\alpha}}d\eta d\xi\right)\\
&\approx \frac{8}{\alpha^2}\left(\int_0^{c_\alpha}\int_0^\xi \frac{e^{-2h(\xi-\eta)}\left(\int_0^\eta e^{-2h\zeta}d\zeta\right)^2}{\xi\eta}d\eta d\xi\right.\\
&\quad\left.+\int_0^{c_\alpha}\int_\xi^{c_\alpha} \frac{\left(\int_0^\xi e^{2(1-h)\zeta}d\zeta\right)^2}{\xi\eta e^{2(1-h)\xi}e^{2(1-h)\eta}}d\eta d\xi\right)\\
&\approx \frac{2}{\alpha^2}\left(\frac{1}{h^2}\int_0^\infty\int_0^\xi \frac{e^{-(\xi-\eta)}\left(\int_0^\eta e^{-\zeta}d\zeta\right)^2}{\xi\eta}d\eta d\xi\right.\\
&\quad\left.+\frac{1}{(1-h)^2}\int_0^\infty\int_\xi^\infty \frac{\left(\int_0^\xi e^{\zeta}d\zeta\right)^2}{\xi\eta e^\xi e^\eta}d\eta d\xi\right) \approx \frac{1}{\alpha^2}\left(\frac{c'}{h^2}+\frac{c''}{(1-h)^2}\right)
\end{aligned}$$

for some finite c', c''.

2. Now, we come to the case $h = 1$. Here, we use the scale function (which is up to a factor of $e^{-\alpha}$ the same as in (6.1) for $h = 1$)

$$S(x) = \int_0^x \exp\left(-2\alpha \int_1^\xi (1-\eta)d\eta\right) d\xi = \int_{1-x}^1 e^{\alpha\xi^2} d\xi.$$

Then,

$$\begin{aligned}
\mathbb{E}_0^{\alpha,h=1}[T^*] &= 2\int_0^1 \frac{\int_0^\xi e^{\alpha\eta^2}d\eta \int_\xi^1 e^{\alpha\eta^2}d\eta}{\xi(1-\xi)e^{\alpha\xi^2}\int_0^1 e^{\alpha\eta^2}d\eta} d\xi \\
&= \frac{2}{\sqrt{\alpha}}\int_0^{\sqrt{\alpha}} \frac{\int_0^\xi e^{\eta^2}d\eta \int_\xi^{\sqrt{\alpha}} e^{\eta^2}d\eta}{\xi(1-\xi/\sqrt{\alpha})e^{\xi^2}\int_0^{\sqrt{\alpha}} e^{\eta^2}d\eta} d\xi \\
&= \frac{2}{\sqrt{\alpha}}(A + \Delta_1 + \Delta_2 + \Delta_3 + \Delta_4),
\end{aligned}$$

such that (recall (6.13), and using $\int_\xi^{\sqrt{\alpha}} e^{\eta^2}d\eta = \int_0^{\sqrt{\alpha}} e^{\eta^2}d\eta - \int_0^\xi e^{\eta^2}d\eta$ and MATHEMATICA for the first term

$$\begin{aligned}
A &= \int_0^\infty \frac{\int_0^\xi e^{\eta^2}d\eta}{\xi e^{\xi^2}} d\xi = \frac{\pi^{3/2}}{4}, \\
\Delta_1 &= -\int_{\sqrt{\alpha}}^\infty \frac{\int_0^\xi e^{\eta^2}d\eta}{\xi e^{\xi^2}} d\xi \approx -\int_{\sqrt{\alpha}}^\infty \frac{1}{2\xi^2} d\xi = -\frac{1}{2\sqrt{\alpha}}, \\
\Delta_2 &= -\int_0^{\sqrt{\alpha}} \frac{\int_0^\xi e^{\eta^2}d\eta \int_0^\xi e^{\eta^2}d\eta}{\xi e^{\xi^2}\int_0^{\sqrt{\alpha}} e^{\eta^2}d\eta} d\xi \approx -\frac{2\sqrt{\alpha}}{e^\alpha}\int_{c_\alpha}^{\sqrt{\alpha}} \frac{e^{\xi^2}}{4\xi^3}\frac{8\xi^4}{8\sqrt{\alpha}^4} d\xi \approx -\frac{1}{4\sqrt{\alpha}^3}, \\
\Delta_3 &= \int_0^{\sqrt{\alpha}-c_\alpha} \frac{\int_\xi^{\sqrt{\alpha}} e^{\eta^2}d\eta \int_0^\xi e^{\eta^2}d\eta}{(\sqrt{\alpha}-\xi)e^{\xi^2}\int_0^{\sqrt{\alpha}} e^{\eta^2}d\eta} d\xi \approx \int_{c_\alpha}^{\sqrt{\alpha}-c_\alpha} \frac{1}{2\xi(\sqrt{\alpha}-\xi)} d\xi \\
&= \frac{1}{2\sqrt{\alpha}}\int_{c_\alpha}^{\sqrt{\alpha}-c_\alpha} \frac{1}{\xi} + \frac{1}{\sqrt{\alpha}-\xi} d\xi \approx \frac{\log\alpha}{2\sqrt{\alpha}},
\end{aligned}$$

$$
\begin{aligned}
\Delta_4 &= \int\limits_{\sqrt{\alpha}-c_\alpha}^{\sqrt{\alpha}} \frac{\int_\xi^{\sqrt{\alpha}} e^{\eta^2} d\eta \int_0^\xi e^{\eta^2} d\eta}{(\sqrt{\alpha}-\xi) e^{\xi^2} \int_0^{\sqrt{\alpha}} e^{\eta^2} d\eta} d\xi \approx \frac{1}{e^\alpha} \int\limits_{\sqrt{\alpha}-c_\alpha}^{\sqrt{\alpha}} \frac{\int_\xi^{\sqrt{\alpha}} \frac{2\eta}{2\sqrt{\alpha}} e^{\eta^2} d\eta}{\sqrt{\alpha}-\xi} d\xi \\
&= \frac{1}{2\sqrt{\alpha}} \int\limits_{\sqrt{\alpha}-c_\alpha}^{\sqrt{\alpha}} \frac{1-e^{-(\alpha-\xi^2)}}{\sqrt{\alpha}-\xi} d\xi = \frac{1}{2\sqrt{\alpha}} \int\limits_0^{c_\alpha} \frac{1-e^{-\xi(2\sqrt{\alpha}-\xi)}}{\xi} d\xi \\
&\approx \frac{1}{2\sqrt{\alpha}} \int\limits_0^{c_\alpha\sqrt{\alpha}} \frac{1-e^{-2\xi}}{\xi} d\xi \approx \frac{\log\alpha}{4\sqrt{\alpha}}.
\end{aligned}
$$

For the variance,

$$
\begin{aligned}
\alpha \mathbb{V}_0^{\alpha,h=1}[T^*] &= 8\alpha \int\limits_0^1 \int\limits_0^\xi \frac{\left(\int_0^{1-\xi} e^{\alpha\zeta^2} d\zeta\right)^2 \left(\int_{1-\eta}^1 e^{\alpha\zeta^2} d\zeta\right)^2}{\left(\int_0^1 e^{\alpha\zeta^2} d\zeta\right)^2 \xi(1-\xi)\eta(1-\eta) e^{\alpha(1-\xi)^2} e^{\alpha(1-\eta)^2}} d\eta d\xi \\
&= 8\alpha \int_0^{\sqrt{\alpha}} \int_\xi^{\sqrt{\alpha}} \frac{\left(\int_0^\xi e^{\zeta^2} d\zeta\right)^2 \left(\int_\eta^{\sqrt{\alpha}} e^{\zeta^2} d\zeta\right)^2}{\left(\int_0^{\sqrt{\alpha}} e^{\zeta^2} d\zeta\right)^2 \xi(\sqrt{\alpha}-\xi)\eta(\sqrt{\alpha}-\eta) e^{\xi^2} e^{\eta^2}} d\eta d\xi \\
&\approx 8 \int_0^\infty \int_0^\eta \frac{\left(\int_0^\xi e^{\zeta^2} d\zeta\right)^2}{\xi\eta e^{\xi^2} e^{\eta^2}} d\xi d\eta < \infty.
\end{aligned}
\tag{6.19}
$$

Proof of Theorem 4 The proof uses an approach similar to the proof of Proposition 3.4 in Etheridge et al. (2006). Recall the structured coalescent from Definition 2.3. We need to show the following assertions in the limit of large α:

(i) any line never undergoes a transition 4. (back recombination to the beneficial background).

(ii) any pair of lines never undergoes a transition 2. (coalescence in the wild-type background).

(iii) any pair of lines never makes a transition 1. (coalescence in the beneficial background) and the coalesced line then makes a transition 3. (recombination to the wild-type).

(iv) the probability for each line to stay in the beneficial background (i.e. no transition 3., recombination to the wildtype background) until the origin of the beneficial allele is $e^{-\lambda/h}$.

(v) if two lines do not coalesce (no transition 1.), recombination to the wildtype background (transition 3.) occurs for both lines independently.

Let us explain how these five assertions imply Theorem 4. Consider a single line, i.e. $\mathcal{K}_0 = (1,0)$. When (i),…,(v) are shown, (i) and (iv) immediately imply (4.5). Next, consider $\mathcal{K}_0 = (2,0)$. For (4.6), the event $K^b_{T^*} + K^w_{T^*} = 1$ means that an event 1. has occurred (since 2. does not occur by (ii)). In addition, the lines coalesce at $\beta = T^*$, the beginning of the sweep. Otherwise there would be the chance that the coalesced line recombines through a transition 3., which is not possible by (iii). In particular, the last two arguments show that $\mathcal{K}_{T^*} = (1,0)$. Hence, $\mathcal{K}_{T^*} = (1,0)$ iff both lines do not make a transition 3. if coalescence cannot occur. This probability is $e^{-2\lambda/h}$ by (v).

The main point in showing the star-like approximation is (iii), as this assertion exactly implies that the structured coalescent converges to a star-like tree. So we start with proving this. Since transitions 1. occur at rate $1/X_\beta$ and transitions 3. at rate $\rho(1-X_\beta)$, and using the Green function identity from Dawson et al. (2001, equation 2.1.1) in the third line,

$$\begin{aligned}
&\mathbb{P}[\text{event described in (iii) occurs}]\\
&= \mathbb{E}^{\alpha,h}\Bigg[\int_0^{T^*}\int_0^t \rho(1-X_s)\exp\left(-\int_s^{T^*}\rho(1+1_{r>t})(1-X_r)dr\right)\\
&\qquad \times \frac{1}{X_t}\exp\left(-\int_t^{T^*}\frac{1}{X_r}dr\right)dsdt\Bigg]\\
&\le \mathbb{E}^{\alpha,h}\left[\int_0^{T^*}\int_0^t \rho(1-X_s)\frac{1}{X_t}dsdt\right]\\
&= \rho\int_0^1\int_0^1 G^*(0,\xi)G^*(\xi,\eta)(1-\xi)\frac{1}{\eta}d\eta d\xi \approx \frac{4\lambda\alpha}{\log\alpha}(B_1+B_2),
\end{aligned}\tag{6.20}$$

where B_1 and B_2 are given below. For B_1, we have for some finite c, using (6.3),

$$\begin{aligned}
B_1 &= \frac{1}{4}\int_0^1\int_0^\xi G^*(0,\xi)G^*(\xi,\eta)(1-\xi)\frac{1}{\eta}d\eta d\xi\\
&= \int_0^1\int_0^\xi \frac{\left(\int_\xi^1 e^{-2h\alpha\zeta+(2h-1)\alpha\zeta^2}d\zeta\right)^2}{\xi\eta^2(1-\eta)e^{-2h\alpha\xi+(2h-1)\alpha\xi^2}e^{-2h\alpha\eta+(2h-1)\alpha\eta^2}}\\
&\qquad \times\frac{\left(\int_0^\eta e^{-2h\alpha\zeta+(2h-1)\alpha\zeta^2}d\zeta\right)^2}{\left(\int_0^1 e^{-2h\alpha\zeta+(2h-1)\alpha\zeta^2}d\zeta\right)^2}d\eta d\xi
\end{aligned}$$

$$\begin{aligned}
&= \int\limits_0^{2h\alpha}\int\limits_0^{\xi} \frac{\left(\int_{\xi}^{2h\alpha} e^{-\zeta+(2h-1)\zeta^2/(4h^2\alpha)}d\zeta\right)^2}{\xi\eta^2(2h\alpha-\eta)e^{-\xi+(2h-1)\xi^2/(4h^2\alpha)}e^{-\eta+(2h-1)\eta^2/(4h^2\alpha)}} \\
&\quad \times \frac{\left(\int_0^{\eta} e^{-\zeta+(2h-1)\zeta^2/(4h^2\alpha)}d\zeta\right)^2}{\left(\int_0^{2h\alpha} e^{-\zeta+(2h-1)\zeta^2/(4h^2\alpha)}d\zeta\right)^2} d\eta d\xi \\
&\approx \frac{1}{2h\alpha}\int\limits_0^{\infty}\int\limits_0^{\xi} \frac{e^{-(\xi-\eta)}\left(\int_0^{\eta} e^{-\zeta}d\zeta\right)^2}{\xi\eta^2} d\eta d\xi \approx \frac{c}{2h\alpha}. \qquad (6.21)
\end{aligned}$$

For B_2, we compute, with some finite c',

$$\begin{aligned}
B_2 &= \frac{1}{4}\int\limits_0^{1}\int\limits_{\xi}^{1} G^*(0,\xi)G^*(\xi,\eta)(1-\xi)\frac{1}{\eta}d\eta d\xi \\
&= \int\limits_0^{1}\int\limits_{\xi}^{1} \frac{\int_0^{\xi} e^{-2h\alpha\zeta+\alpha\zeta^2(2h-1)}d\zeta \int_{\xi}^{1}\cdots d\zeta \int_0^{\eta}\cdots d\zeta \int_{\eta}^{1}\cdots d\zeta}{\xi\eta^2(1-\eta)e^{-2h\alpha\xi+\alpha\xi^2(2h-1)}e^{-2h\alpha\eta+\alpha\eta^2(2h-1)}\left(\int_0^{1}\cdots d\zeta\right)^2} d\eta d\xi \\
&= \int\limits_0^{2h\alpha}\int\limits_{\xi}^{2h\alpha} \frac{\int_0^{\xi} e^{-\zeta+\zeta^2(2h-1)/(4h^2\alpha)}d\zeta \int_{\xi}^{2h\alpha}\cdots d\zeta \int_0^{\eta}\cdots d\zeta}{\xi\eta^2(2h\alpha-\eta)e^{-\xi+\xi^2(2h-1)/(4h^2\alpha)}e^{-\eta+\eta^2(2h-1)/(4h^2\alpha)}} \\
&\qquad \times \frac{\int_{\eta}^{2h\alpha}\cdots d\zeta}{\left(\int_0^{2h\alpha}\cdots d\zeta\right)^2} d\eta d\xi \\
&\approx \frac{1}{2h\alpha}\int\limits_0^{c_\alpha}\int\limits_{\xi}^{c_\alpha} \frac{\int_0^{\xi} e^{-\zeta}d\zeta \int_{\xi}^{2h\alpha} e^{-\zeta}d\zeta \int_0^{\eta} e^{-\zeta}d\zeta \int_{\eta}^{2h\alpha} e^{-\zeta}d\zeta}{\xi\eta^2 e^{-\xi}e^{-\eta}\left(\int_0^{\infty} e^{-\zeta}d\zeta\right)^2} d\eta d\xi \\
&\approx \frac{1}{2h\alpha}\int\limits_0^{c_\alpha}\int\limits_0^{\eta} \frac{\int_0^{\xi} e^{-\zeta}d\zeta \int_0^{\eta} e^{-\zeta}d\zeta}{\xi\eta^2} d\xi d\eta = \frac{c'}{2h\alpha} \qquad (6.22)
\end{aligned}$$

where each integrand in $\int \cdots d\zeta$ is the same as the first integrand in the respective line. Combining (6.20), (6.21) and (6.22) shows that

$$\mathbb{P}[\text{event described in (iii) occurs}] \xrightarrow{\alpha\to\infty} 0$$

which proves (iii). Similar calculations show that

$$\mathbb{P}[\text{event described in (i) occurs}] \leq \rho^2 \int\limits_0^1 \int\limits_0^1 G^*(0,\xi)G^*(\xi,\eta)\xi(1-\eta)d\eta d\xi \xrightarrow{\alpha\to\infty} 0,$$

$$\mathbb{P}[\text{event described in (ii) occurs}] \leq \rho \int\limits_0^1 \int\limits_0^1 G^*(0,\xi)G^*(\xi,\eta)\frac{1}{1-\xi}(1-\eta)d\eta d\xi \xrightarrow{\alpha\to\infty} 0,$$

and (i), (ii) follow. For (iv) and (v), note that the probability that a line does not recombine at all is

$$\mathbb{P}[\text{event described in (iv) occurs}] = \mathbb{E}\left[\exp\left(-\rho\int\limits_0^{T^*}(1-X_s)ds\right)\right]$$

and (v) is equivalent to

$$\mathbb{P}[\text{neither of two lines recombines}] = \mathbb{E}\left[\exp\left(-2\rho\int\limits_0^{T^*}(1-X_s)ds\right)\right] \approx e^{-2\lambda/h}.$$

Hence, (iv) and (v) are proved once we show that

$$\mathbb{E}\left[\exp\left(-\rho\int\limits_0^{T^*}(1-X_s)ds\right)\right] \approx e^{-\lambda/h}. \tag{6.23}$$

In order to see this, we show that

$$\rho\int\limits_0^{T^*}(1-X_s)ds \xrightarrow{\alpha\to\infty} \frac{\lambda}{h}$$

in L^2. Since L^2-convergence implies convergence in distribution, and $x \mapsto e^{-x}$ is bounded on $\mathbb{R}_+$ and continuous, (6.23) then follows. We compute

$$\begin{aligned}\mathbb{E}\left[\rho\int_0^{T^*}(1-X_s)ds\right] &= \rho\int_0^1 G^*(0,\xi)(1-\xi)d\xi \\ &= \frac{2\rho}{\alpha}\int_0^\alpha \frac{\int_0^\xi e^{-2h\eta-(1-2h)\eta^2/\alpha}d\eta \int_\xi^\alpha e^{-2h\eta-(1-2h)\eta^2/\alpha}d\eta}{\xi e^{-2h\xi-(1-2h)\xi^2/\alpha}\int_0^\alpha e^{-2h\eta-(1-2h)\eta^2/\alpha}d\eta}d\xi \\ &\approx \frac{4h\rho}{\alpha}A_1 \approx \frac{\rho\log\alpha}{h\alpha} \approx \frac{\lambda}{h} \qquad (6.24)\end{aligned}$$

with A_1 from (6.16), where we have used (6.18). Next,

$$\begin{aligned}\mathbb{V}\left[\rho^2\int_0^{T^*}(1-X_s)ds\right] &= 8\rho^2\int_0^1\int_0^\xi G^*(0,\xi)G^*(\xi,\eta)(1-\xi)(1-\eta)d\eta d\xi \\ &\leq 8\rho^2\mathbb{V}[T^*] \xrightarrow{\alpha\to\infty} 0 \qquad (6.25)\end{aligned}$$

by Theorem 3. Combining (6.24) and (6.25), L^2 convergence follows and we have proved (6.23) as well as Theorem 4.

Acknowledgments We thank Cornelia Borck for fruitful discussion, two anonymous referees for various clarifying comments and Reinhard Bürger for bringing the contents of Remark 2.2 to our knowledge. The work was made possible with financial support by the Vienna Science and Technology Fund (WWTF) to JH and GE and by the Federal Ministry of Education and Research, Germany (BMBF) to PP (Kennzeichen 0313921). JH and PP acknowledge support from the DFG Forschergruppe 1078 *Natural selection in structured populations*.

References

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

Barton N, Etheridge A, Sturm A (2004) Coalescence in a random background. Ann Appl Probab 14(2):754–785

Bürger R, Ewens W (1995) Fixation probabilities of additive alleles in diploid populations. J Math Biol 33:557–575

Cavatorta J, Savage A, Yeam I, Gray S, Jahn M (2008) Positive Darwinian selection at single amino acid sites conferring plant virus resistance. J Mol Evol 67:551–559

Charlesworth B (1998) Adaptive evolution: the struggle for dominance. Curr Biol 8(14):R502–R504

Dawson D, Gorostiza L, Wakolbinger A (2001) Occupation time fluctuations in branching systems. J Theor Probab 14:729–796

Durrett R, Schweinsberg J (2004) Approximating selective sweeps. Theor Popul Biol 66:129–138

Etheridge A, Pfaffelhuber P, Wakolbinger A (2006) An approximate sampling formula under genetic hitchhiking. Ann Appl Probab 16:685–729

Ewens W (2004) Mathematical Population Genetics, I. Theoretical introduction. 2nd edn. Springer, New York

Ewing G, Hermisson J (2010) MSMS: a coalescent simulation program including recombination, demographic structure, and selection at a single locus. Bioinformatics 26(16):2064–2065

Garcia G, Flores A, Fernandez-Salas I, Saavedra-Rodriguez K, Reyes-Solis G, Lozano-Fuentes S, Guillermo Bond J, Casas-Martinez M, Ramsey J, Garcia-Rejon J, Dominguez-Galera M, Ranson H, Hemingway J, Eisen L, Black WIV (2009) Recent rapid rise of a permethrin knock down resistance allele in *Aedes aegypti* in Mexico. PLoS Negl Trop Dis 3:e531

Hamblin M, Di Rienzo A (2000) Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. Am J Hum Genet 66:1669–1679

Hermisson J, Pennings P (2005) Soft sweeps: molecular population genetics of adaptation from standing genetic variation. Genetics 169(4):2335–2352

Hermisson J, Pfaffelhuber P (2008) The pattern of genetic hitchhiking under recurrent mutation. Electron J Probab 13(68):2069–2106

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

Kaplan N, Hudson R, Langley C (1989) The ‘Hitchhiking effect’ revisited. Genetics 123:887–899

Karlin S, Taylor H (1981) A Second Course in Stochastic Processes. Academic Press, London

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

Kimura M (1962) On the probability of fixation of mutant genes in a population. Genetics 47:713–719

Leocard S, Pardoux E (2010) Evolution of the ancestral recombination graph along the genome in case of selective sweep. J Math Biol 61(6):819–841

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

McVean G (2007) The structure of linkage disequilibrium around a selective sweep. Genetics 175:1395–1406

Nielsen R, Williamson S, Kim Y, Hubisz M, Clark A, Bustamante C (2005) Genomic scans for selective sweeps using SNP data. Genome Res 15:1566–1575

Ordon F, Friedt W, Scheurer K, Pellio B, Werner K, Neuhaus G, Huth W, Habekuss A, Graner A (2004) Molecular markers in breeding for virus resistance in barley. J Appl Genet 45:145–159

Pennings P, Hermisson J (2006a) Soft sweeps II—molecular population genetics of adaptation from recurrent mutation or migration. Mol Biol Evol 23(5):1076–1084

Pennings P, Hermisson J (2006b) Soft sweeps III—the signature of positive selection from recurrent mutation. PLoS Genet 2:e186

Pfaffelhuber P, Studeny A (2007) Approximating genealogies for partially linked neutral loci under a selective sweep. J Math Biol 55:299–330

Stephan W, Wiehe T, Lenz M (1992) The effect of strongly selected substitutions on neutral polymorphism: analytical results based on diffusion theory. Theor Popul Biol 41:237–254

Stephan W, Song Y, Langley C (2006) The hitchhiking effect on linkage disequilibrium between linked neutral loci. Genetics 172:2647–2663

Tajima F (1983) Evolutionary relationship of DN sequences in finite populations. Genetics 105:437–460

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595

Teshima K, Przeworski M (2006) Directional positive selection on an allele of arbitrary dominance. Genetics 172:713–718

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

van Herwaarden O, van der Wal N (2002) Extinction time and age of an allele in a large finite population. Theor Popul Biol 61:311–318

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