*Adv. Appl. Prob.* **41,** 731–764 (2009)
*Printed in Northern Ireland*
© *Applied Probability Trust* 2009

# SELECTIVE SWEEP AND THE SIZE OF THE HITCHHIKING SET

STEPHANIE LEOCARD,* *Université de Provence*

## Abstract

Just after the fixation of an advantageous allele in the population (this spread is called a selective sweep), the neutral genes close to the site under selection tend to have the same ancestor as the gene under selection. However, some recombinations may occur during the selective sweep and break the link, which reduces the number of hitchhiking alleles. We consider a large selection coefficient $\alpha$ and extend the results of Etheridge, Pfaffelhuber and Wakolbinger (2006) and the work of Pfaffelhuber and Studeny (2007) about genetic hitchhiking, where the recombination rate scales with $\alpha/\log\alpha$. We first describe the genealogy at an arbitrary number of partially linked neutral loci, with an order of accuracy of $\mathcal{O}(1/(\log\alpha)^2)$ in total variation. Then, we use this framework to obtain an approximate distribution for the size of the hitchhiking set at the end of the selective sweep, with the same accuracy.

*Keywords:* Coalescence; recombination; selective sweep; hitchhiking allele

2000 Mathematics Subject Classification: Primary 92D15
Secondary 60J80; 60J85; 60K37; 92D10

## 1. Introduction

We study the allelic diversity of neutral genes close to a site under selection, immediately after a selective sweep, when an advantageous allele quickly spreads in the population until its fixation. Selection tends to reduce the diversity of these genes. Indeed, if no recombination event occurred in the region separating these genes from the selected gene during the selective sweep, all the individuals in the population would carry the same alleles as the ancestral individual (the original carrier of the beneficial mutation) at the end of the selective sweep. However, because of recombinations, the loss of diversity is less and less radical as the distance from the selected locus increases. This concept of genetic hitchhiking has been highlighted by Maynard-Smith and Haigh [10] and is the topic of a number of studies [2], [16], [17], [11]. In particular, the complexity of this phenomenon induces a need for approximations that yield quantitative information [2], [14], [16].

These considerations motivate us to introduce the concept of a *hitchhiking set*. We define the hitchhiking set as the set of contiguous alleles located close to the selected site that share the same ancestor as the site under selection (a more precise definition will be given in Section 3.2). The knowledge of the joint genealogy of the neutral genes allows us to reconstruct the hitchhiking set of individuals at the end of the selective sweep. The measure of the size of the hitchhiking set is a very interesting tool for phylogenetic studies: since the fixation of the advantageous allele is very quick, few recombinations happen during the selective sweep and, consequently, the hitchhiking set is bigger than under neutrality. Thus, under the hypothesis of constant

Received 5 November 2008; revision received 18 May 2009.

* Postal address: CMI, Technopôle Château-Gombert, 39 rue Joliot Curie, 13453 Marseille cedex 13, France.
Email address: leocard@cmi.univ-mrs.fr

https://doi.org/10.1239/aap/1253281062 Published online by Cambridge University Press

recombination probabilities along the chromosome (for example, the considered region does not contain a recombination hotspot, which is a small portion with highly elevated recombination rate [11]), detecting large hitchhiking sets can be a way to detect positive selection. With similar arguments, Sabeti *et al.* [15] used the evolution of extended haplotype homozygosity along the genome to deduce that two genes involved in resistance to malaria are under selection. Depaulis and Veuille [1], Hanchard *et al.* [6], Hudson *et al.* [7], and Wang *et al.* [18] also based their statistical tests of detection on haplotypes.

We consider a model with a single selected site and we ignore the effects of mutation (at the selected site or elsewhere) during the selective sweep. In case of strong selection and when the recombination rate scales with $\alpha/(\log\alpha)$, where $\alpha$ is the rescaled selective advantage, Etheridge *et al.* [2] recently obtained an approximate formula for the genealogy at a single partially linked neutral locus, with an order of accuracy of $\mathcal{O}(1/(\log\alpha)^2)$ compared to the Wright–Fisher model. Pfaffelhuber and Studeny [13] used some of the results of that work to describe the genealogy at two partially linked neutral loci, with the same order of accuracy.

The present work begins with an extension of the results established by Pfaffelhuber and Studeny. With the same accuracy of $\mathcal{O}(1/(\log\alpha)^2)$, we obtain the full description of the joint genealogy of a number, $2m$, of partially linked neutral loci ($m \geq 1$), $m$ located at one side and $m$ at the other side of the site under strong selection. In this part, our analysis is based on the methodology of [13]. However, not only do we correct some imprecisions, but we need novel observations and calculations necessary to cover the case of several neutral genes in the proof of Proposition 13 as well. We then use these results as well as some results established in [2] to calculate an approximate distribution for the size of the hitchhiking set, with the same accuracy. We made simulations under the exact model of evolution to see how well our approximations perform. The results are presented in Section 3.3. We can see that the error due to approximations is very small when the size of the sample is small.

The paper is organized as follows. In Section 2 we present the evolutionary model, describing the evolution at the selected site and the modeling of coalescence and recombination events. In Section 3 we present the two main results of this work and some numerical comparisons. In Section 4 we construct the approximate model obtained by successive approximations, neglecting events whose probability is $\mathcal{O}(1/(\log\alpha)^2)$, and we give the proof of Theorem 1. Section 5 is dedicated to the proof of Theorem 2. Section 6 contains the technical proof of a proposition needed in Section 4.

## 2. Model of evolution

### 2.1. Notation and hypotheses

The population is assumed to be haploid and of constant size $2N$.

We consider the genome (or its subregion) as a single chromosome with a succession of genes. We make the hypothesis that one of these genes, denoted by $R_0$, is under a selective sweep (see Section 2.2 for precisions) and that its location is known. We use it to initialize the location of all the genes, denoting by 0 the locus of this gene. Let $m$ be the number of genes to the right and to the left of $R_0$. We denote by $R_1, \dots, R_m$ the genes to the right of $R_0$ and by $R_{-m}, \dots, R_{-1}$ those genes to its left, as shown in Figure 1.

Recombinations may arise along the genome. The rate of recombination between $R_{a-1}$ and $R_a$, $-m + 1 < a < m$, is denoted by $\rho_a$, and is assumed to be constant in time. These rates depend on $a$ because they depend on the length of the DNA between $R_{a-1}$ and $R_a$. Indeed, the larger this gap, the more recombinations may happen within this gap.

https://doi.org/10.1239/aap/1253281062 Published online by Cambridge University Press

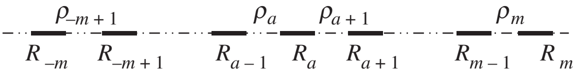


FIGURE 1: Model for the genome. The gene under selection is $R_0$. There are $m$ neutral genes to the right and to the left of $R_0$. The recombination rate between loci $a-1$ and $a$ is denoted by $\rho_a$, $-m+1 \le a \le m$.

**Remark 1.** The results can be easily extended to a continuous chromosome on $[-1, 1]$ with a continuum of loci. In this case, recombinations occur during the evolution process according to a Poisson process on $\mathbb{R}_+ \times [-1, 1]$ with intensity $\rho\,\mathrm{d}t \times \mathrm{d}x$.

### 2.2. Evolution at the selected site

First of all, let us describe precisely the evolution at locus 0. We assume that reproduction in the population follows the Wright–Fisher model. At the beginning, all the individuals have the wild-type allele $b$ of the gene $R_0$. At $t = 0$, a new allele $B$ appears in one individual. Compared to $b$, this allele has a selective advantage denoted by $s$. We make two assumptions: no other mutation occurs at this locus and $B$ is fixed in the population after a finite time $T$, so that nobody carries allele $b$ after time $T < \infty$. This process is called a *selective sweep* and $T$ is its duration.

Let $X_t$ be the proportion of $B$ in the whole population at time $t \ge 0$. With this notation, $T = \inf\{t \ge 0;\ X_t = 1\}$.

To model the evolution of $X_t$, we make the approximation that the size of the population is infinite and we assume that $\lim_{N\to\infty} 2sN = \alpha$, where $\alpha \in (0, +\infty)$. Rescaling time so that $2N$ generations become one unit of time and conditioned upon fixation of the allele $B$, $(X_t)_{t\ge 0}$ is the solution of the following stochastic differential equation:

$$\mathrm{d}X_t = \alpha X_t(1 - X_t)\coth\left(\frac{\alpha}{2}X_t\right)\mathrm{d}t + \sqrt{2X_t(1 - X_t)}\,\mathrm{d}W_t, \qquad X_0 = 0, \qquad (1)$$

where $W$ is a realization of the standard Brownian motion (see [5]).

We are interested in the case where $1 \ll \alpha$.

### 2.3. The different ways of recombination and coalescence

We consider a sample of $n$ individuals taken at the end of the selective sweep (i.e. at $t = T$). We consider $2m + 1$ loci from each of them.

We denote by $r_{a,p}$ the allele of gene $R_a$ of individual $p$, $-m \le a \le m$, $1 \le p \le n$. For simplicity, we identify $r_{a,p}$ with the integer $an + p$, so that $an + 1, \dots, an + n$ correspond to the alleles of the sample for gene $R_a$ (see Figure 2).

Going back in time from $t = T$ to $t = 0$, coalescence and recombination events occur. Note that until the end, we will always look back in time, so that if two events $A_1$ and $A_2$ happen at time $t_1$ and $t_2$, respectively, and if $t_1 > t_2$, we will say that $A_1$ happens *before* $A_2$.

First, we describe the coalescence events. Because of the hypothesis that no mutation happens for the gene under selective sweep after the one which created the advantageous allele at $t = 0$, we think in terms of *structured coalescence* [8], where individuals carrying allele $b$ cannot coalesce with those carrying allele $B$. Thus, coalescence events are only allowed for two individuals carrying the same allele at locus 0.

Recall that $X_t$ is the proportion of $B$ in the population at time $t$.

https://doi.org/10.1239/aap/1253281062 Published online by Cambridge University Press

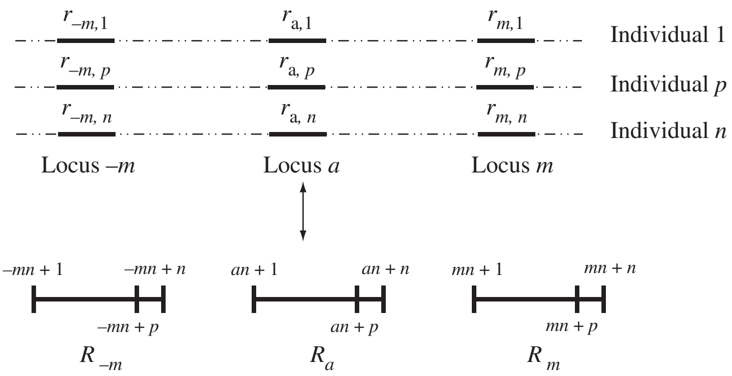


FIGURE 2: For $-m \leq a \leq m$ and $1 \leq p \leq n$, the allele $r_{a,p}$ of gene $R_a$ of individual $p$ is identified with the integer $an + p$.

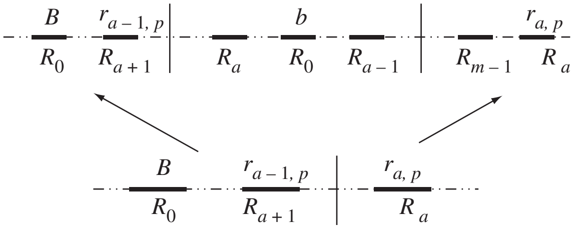


FIGURE 3: The event $(B \Rsh b, a)$ for $a > 0$. Genes $R_a, \ldots, R_m$ originate from an individual carrying $b$, whereas $R_{-m}, \ldots, R_{a-1}$ originate from an individual carrying $B$.

Two individuals carrying $B$ coalesce at rate $2/X_t$ (this event will be called coal$B$), and two individuals carrying $b$ coalesce at rate $2/(1 - X_t)$ (this event will be called coal$b$) [2].

We now focus on the recombination events, which are independent of the coalescence events. For any ancestral individual, we care about the alleles present at locus 0 and the genetic material that is present in the sample at the end of the selective sweep. Several kinds of recombination may occur, according to the place of the recombination and the allele which is present at locus 0. For $-m + 1 \leq a \leq m$, we define the events $(B \Rsh b, a)$ and $(B \Rsh B, a)$ if the individual carries $B$, and $(b \Rsh B, a)$ and $(b \Rsh b, a)$ if the individual carries $b$, as follows.

Consider the event $(b_1 \Rsh b_2, a)$, where $b_1$ and $b_2$ can be either $B$ or $b$. For an individual carrying $b_1$ at locus 0, a recombination happens between the loci $a-1$ and $a$, and the ancestor for gene $R_a$ carries $b_2$ at locus 0. In other words, if $a > 0$ or $a < 0$, genes $R_c$ for $-m \leq c \leq a - 1$ have to originate from an individual carrying $b_1$ or, respectively, $b_2$ and genes $R_c$ for $a \leq c \leq m$ come from an individual carrying $b_2$ or, respectively, $b_1$. An example is given in Figure 3.

For any individual carrying $B$ or $b$, the rate of the events of type $(B \Rsh b, a)$ or, respectively, $(b \Rsh b, a)$ is $\rho_a(1 - X_t)$ at time $t$. Indeed, the recombination occurs between the loci $a - 1$ and $a$ at rate $\rho_a$, and the ancestor for gene $R_a$ is chosen from the population carrying $b$, whose proportion is $(1 - X_t)$ at time $t$. Similarly, the rate of event $(B \Rsh B, a)$ or, respectively, $(b \Rsh B, a)$ is $\rho_a X_t$ at time $t$.

https://doi.org/10.1239/aap/1253281062 Published online by Cambridge University Press

We have to pay attention to the number of recombination events occurring during the selective sweep. If the number of events is negligible, we expect to find (almost) no variation in the data. If the number of events is excessive, we expect a neutral pattern. Since the duration of the selective sweep is of order $\mathcal{O}(\log \alpha/\alpha)$ (see [2, Lemma 3.1]), we choose the rates $\rho_a$ to be of the order $\alpha/\log\alpha$. Then the number of recombinations during the selective sweep is not trivial. We let $\rho_a$ depend upon $\alpha$ as follows: for each $-m+1 \le a \le m$, let $\gamma_a$ be a fixed positive number and let $\rho_a = \gamma_a\alpha/\log\alpha$.

Given all these events, we construct the genealogy of the sample at time $t$. The genealogy $\pi(t)$ of the alleles at time $t$ is a marked partition (an asterisk is added to denote marked blocks) of $\{-mn+1, \dots, (m+1)n\}$. In previous work [2], it was said that two alleles are in the same block if they have the same ancestral allele. Here we consider the evolution of several genes, so we say instead that two alleles are in the same block at time $t$ if their ancestral alleles are carried by the same individual at time $t$. We mark the blocks corresponding to ancestors carrying $B$ at locus 0 at time $t$. In this way, we finally obtain the ancestral marked partition $\pi(0)$.

Since the rates of the various recombination and coalescence events depend on $(X_t)_{0\le t\le T}$, to obtain the distribution of the ancestral partition, we will first describe the conditional law of $\pi(0)$ given $(X_t)_{0\le t\le T}$, and then average over all the paths $X = (X_t)_{0\le t\le T}$. We denote this final distribution by $\Gamma$.

## 3. Main results

### 3.1. Approximate model

From now on, for any tree, a branch denotes an edge connecting two consecutive nodes of this tree.

Consider the following procedure.

1. Start with the partition $\pi = \{\{-mn+1, -(m-1)n+1, \dots, mn+1\}^*, \dots, \{-mn+n, -(m-1)n+n, \dots, mn+n\}^*\}$.

2. Draw $(U_{a,p})_{-m+1\le a\le m,\, 1\le p\le n}$ independent Bernoulli variables with parameter
$$1-\exp\biggl(-\frac{\gamma_a}{\log\alpha}\sum_{\ell=1}^{\lfloor\alpha\rfloor}\frac{1}{\ell}\biggr).$$
If $U_{a,p} = 1$, realize the event $(B \Rsh B, a)$ for the individual $p$. This produces a new partition $\pi'$, with $|\pi'| = |\pi| + \sum_{a,p} U_{a,p}$.

3. Simulate a Yule tree with rate $\alpha$ until it has $\lfloor\alpha\rfloor$ leaves.

4. Extract a subtree $\mathcal{Y}_{|\pi'|}$ of $|\pi'|$ leaves, which are uniformly chosen among the $\lfloor\alpha\rfloor$ leaves of the simulated Yule tree.

5. On each branch of the subtree, independently put a label $\boldsymbol{r}$ representing the recombination events on the branch according to the following probabilities:

$$\mathrm{P}(\boldsymbol{r} = (r_1, \dots, r_q)) = \prod_{c\in\{r_1,\dots,r_q\}} (1 - p_{\ell_1}^{\ell_2}((r_1, \dots, r_q), c)) \prod_{c\notin\{r_1,\dots,r_q\}} p_{\ell_1}^{\ell_2}((r_1, \dots, r_q), c) \tag{2}$$

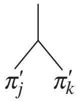
https://doi.org/10.1239/aap/1253281062 Published online by Cambridge University Press

with

$$p_{\ell_1}^{\ell_2}((r_1,\dots,r_q),c) = \begin{cases} \exp\Bigl(-\dfrac{\gamma_c}{\log\alpha}\displaystyle\sum_{\ell=\ell_1+1}^{\ell_2}\frac{1}{\ell}\Bigr) & \text{if } r_{i_0} \le c \le r_{i_0+1}; \\ \exp\Bigl(-\dfrac{\gamma_c}{\log\alpha}\displaystyle\sum_{\ell=1}^{\ell_2}\frac{1}{\ell}\Bigr) & \text{otherwise,} \end{cases}$$

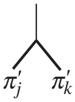
where $i_0 \in \{0,\dots,q\}$ denotes the unique index such that $r_{i_0} \le 0 \le r_{i_0+1}$ (we set $r_0 = -m$ and $r_{q+1} = m+1$).

6. Determine $\pi''$ according to the following equivalence relation. For any $j,k \in \{-mn+1, \dots, (m+1)n\}$, let $\pi'_j$ and $\pi'_k$ be the blocks of $\pi'$ respectively containing $j$ and $k$. Let $-m \le c \le d \le m$ be such that $j$ and $k$ are respectively alleles of gene $R_c$ and gene $R_d$ (even if it means exchanging $j$ and $k$).

Consider the subtree of $\mathcal{Y}_{|\pi'|}$ joining $\pi'_j$ and $\pi'_k$ to the root of $\mathcal{Y}_{|\pi'|}$: 

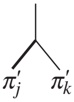
(in the following we will use thick lines to denote branches that we will have to pay attention to).

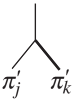
We have $j \sim k$ if and only if one of the following conditions holds true.

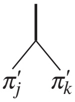
(i) $c = d$ and there exists an $i \in \{0,\dots,q+1\}$ such that $r_i \le 0 \le r_{i+1} - 1$ and $r_i \le d \le r_{i+1} - 1$ for the labels on 

, i.e. if no recombination brings either $j$ or $k$ into background $b$ before the coalescence of their lineage on $\mathcal{Y}_{|\pi'|}$, owing to an event $(\Rsh b, a)$, $0 < |a| \le |d|$.

(ii) $c < d$ and there exist $i_1, i_2, i_3 \in \{0,\dots,q+1\}$ such that

- $r_{i_1} \le 0 \le r_{i_1+1} - 1$ and $r_{i_1} \le c \le r_{i_1+1} - 1$ for the labels on 

;
- $r_{i_2} \le 0 \le r_{i_2+1} - 1$ and $r_{i_2} \le d \le r_{i_2+1} - 1$ for the labels on 

;
- $r_{i_3} \le c < d \le r_{i_3+1} - 1$ for the labels on 

i.e. if no recombination brings either $j$ or $k$ into background $b$ before the coalescence of their lineages on $\mathcal{Y}_{|\pi'|}$ (owing to an event of type $B \Rsh b$), and if, after coalescence, no recombination occurs between the two loci (owing to an event of type $(B \Rsh b, a)$, $c < a \le d$, or, if $cd > 0$, an event $(B \Rsh b, a)$, $|a| \le \min(|c|, |d|)$, followed by an event $(b \Rsh b, a)$, $c < a \le d$).

7. Let $\Gamma_1 = \{\{\pi''_f\}^*, \pi'' \setminus \{\pi''_f\}\}$, where $\pi''_f = \bigcup_{a=-m}^{m}\{j$ allele of $R_a :$ there exists $i \in \{0, \dots, q+1\}$, $r_i \leq 0 \leq r_{i+1} - 1$, and $r_i \leq a \leq r_{i+1} - 1$ for the labels on $\pi'_j\}$ is the nonrecombinant block, that is, the marked block.

We write $\Gamma_1$ for the distribution of the ancestral partition under this model, defined on the set of the partitions of $\{-mn + 1, \dots, mn + n\}$.

Recall that $\Gamma$ is the distribution for the ancestral partition under the Wright–Fisher diffusion model, considered as the *exact* model.

**Theorem 1.** *We have* $d_{\mathrm{TV}}(\Gamma, \Gamma_1) = \mathcal{O}(1/(\log\alpha)^2)$, *where* $d_{\mathrm{TV}}$ *denotes the total variation distance.*

Section 4 will be devoted to the proof of Theorem 1.

### 3.2. Approximate distribution for the size of the hitchhiking set

We consider a sample of $n$ individuals taken at the end of the sweep. For each individual, we define the hitchhiking set.

**Definition 1.** The *hitchhiking set* of an individual is the set of loci (the selected site included) between the last recombination $B \Rsh b$ on the left of locus 0 and the first recombination $B \Rsh b$ on the right of locus 0.

We are interested in the size, $H$, of the hitchhiking set.

We use symbols to specify whether an allele comes from the ancestral individual where $B$ appeared: an allele is symbolized by a circle if it comes from this individual and by a diamond otherwise. The size of the hitchhiking set can then be seen as the length of the block of circles containing $R_0$. For example, in Figure 4 we have $H = 6$.

Note that the size of the hitchhiking set may be smaller than the total number of 'circle' symbols. An example is given in Figure 5, where $H = 3$, but four alleles are inherited from the initial carrier of $B$.

For simplicity, we assume that $\gamma_a$ does not depend on $a$: $\gamma_a = \gamma$ for all $-m + 1 \leq a \leq m$. We draw a sample composed of $n$ individuals at the end of the selective sweep. For $1 \leq p \leq n$, let $H_p$ be the size of the hitchhiking set for individual $p$. We study the joint distribution of $H_1, \dots, H_n$. To write our approximate joint distribution, we have to define $\mathbb{N}$-valued random variables. For $1 \leq p \leq n$, let $h_p \in \{1, \dots, m + 1\}$. Let $U_{\cdot,p} = \tilde{U}_{\cdot,p} + 1$, where $\tilde{U}_{\cdot,p}$, $1 \leq p \leq n$, are independent random variables of binomial distribution with parameters $(h_p - 1, 1 - \exp((-\gamma/\log\alpha)\sum_{\ell=1}^{\lfloor\alpha\rfloor}(1/\ell)))$.

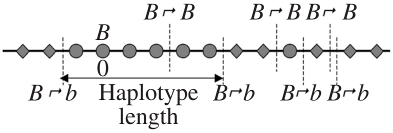


FIGURE 4: An example of the size of the hitchhiking set. The circles and diamonds respectively mark alleles inherited from an ancestor carrying $B$ and an ancestor carrying $b$ at locus 0. Here, $H = 6$.

https://doi.org/10.1239/aap/1253281062 Published online by Cambridge University Press

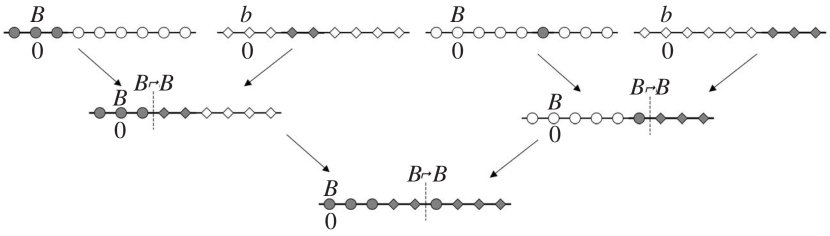


FIGURE 5: An example where the size of the hitchhiking set is smaller than the number of alleles inherited from the original carrier of mutation $B$. The gene under selection is at locus 0. The studied individual, located at the bottom of the figure, carries four alleles symbolized by circles, whereas the size of the hitchhiking set is equal to 3. The filled symbols correspond to the genetic material carried by the studied individual.

Given $U_{\cdot,p} = u_p$, let $Y_{2,p}, \dots, Y_{u_p,p}$ be $\mathbb{N}$-valued random variables such that

$$\mathrm{P}((Y_{2,p} = y_{2,p}, \dots, Y_{u_p,p} = y_{u_p,p})) = \begin{cases} \dbinom{h_p - 1}{u_p - 1}^{-1} & \text{if } 1 \le y_{2,p} < \cdots < y_{u_p,p} \le h_p - 1; \\ 0 & \text{otherwise.} \end{cases}$$

We set $y_{1,p} = 0$, $y_{u_p+1,p} = h_p$, and $U = \sum_{p=1}^{n} U_{\cdot,p}$.

Let $S$ and $F$ be two $\mathbb{N}$-valued random variables, conditionally independent given $U = u$, defined respectively on $\{0, \dots, u\}$ and $\{2, \dots, \lfloor \alpha \rfloor\}$ as follows. For all $0 \le s \le u$,

$$\mathrm{P}(S = s \mid U = u) = \begin{cases} 1 - \dfrac{\gamma u}{\log(\alpha)} \displaystyle\sum_{k=1}^{u-1} \frac{1}{k} & \text{if } s = 0; \\ \dfrac{\gamma u}{\log(\alpha)} \displaystyle\sum_{k=2}^{u-1} \frac{1}{k} & \text{if } s = 1; \\ \dfrac{\gamma u}{\log(\alpha)} \dfrac{1}{s(s-1)} & \text{if } 2 \le s \le u - 1; \\ \dfrac{\gamma u}{\log(\alpha)} \dfrac{1}{u-1} & \text{if } s = u. \end{cases} \tag{3}$$

For all $2 \le f \le \lfloor \alpha \rfloor$,

$$\mathrm{P}(F \le f \mid U = u) = \frac{(f - (u-1)) \cdots (f-1)}{(f + (u-1)) \cdots (f+1)}. \tag{4}$$

Given $u_p, y_{a,p}$, $1 \le a \le u_p$, $1 \le p \le n$, and $1 \le k \le m$, let $\Theta_k$ be the cardinality of $\{(a, p);\, y_{a+1,p} \le k\}$.

**Theorem 2.** *We have*

$$\mathrm{P}(H_1 = h_1, \dots, H_n = h_n) = \sum_{k_1=1}^{h_1} \cdots \sum_{k_n=1}^{h_n} \mathrm{P}(H_1^{\ell} = k_1, \dots, H_n^{\ell} = k_n) \\ \times \mathrm{P}(H_1^{r} = h_1 - k_1 + 1, \dots, H_n^{r} = h_n - k_n + 1),$$

*where $(H_1^\ell, \dots, H_n^\ell)$ and $(H_1^r, \dots, H_n^r)$ are two independent, identically distributed vectors. For any individual $p$, $H_p^r$ and $H_p^\ell$ stand for the restriction of its hitchhiking set to $\{0, \dots, m\}$ and $\{-m, \dots, 0\}$, respectively.*

*The distribution of these random vectors is easily deduced from*

$$\mathrm{P}(H_1^r \geq h_1, \dots, H_n^r \geq h_n) = \mathrm{E}\Bigg[\Bigg(\frac{q(U)}{\binom{U}{S}}\Bigg[\sum_{k=1}^{m}\binom{\Theta_k}{S}\Bigg] + 1 - (\max(h_1, \dots, h_n) - 1)q(U)\Bigg) \times p_{F-1}{}^{\sum_{p=1}^{n}\sum_{a=2}^{U_{\cdot,p}+1}(Y_{a,p}-1)}\Bigg] + \mathcal{O}\bigg(\frac{1}{(\log\alpha)^2}\bigg),$$

*where $q(U) = (U\gamma/\log\alpha)\sum_{\ell=1}^{U-1}(1/\ell)$ and $p_{F-1} = \exp(-(\gamma/\log\alpha)\sum_{\ell=F}^{\lfloor\alpha\rfloor}(1/\ell))$.*

**Remark 2.** Note that the selected locus is present in both of the restrictions. This is why we have to consider $\mathrm{P}(H_p^r = h_p - k_p + 1)$ instead of $\mathrm{P}(H_p^r = h_p - k_p)$ in the convolution formula.

Section 5 will be devoted to the proof of Theorem 2.

### 3.3. Numerical results

In this subsection we present some numerical results to confirm that our approximations perform well.

We consider two examples. In both examples, we chose $N = 10\,000$ and $s = 0.1$, that is, $\alpha = 2000$. In the first example, we set $\rho = 42$, whereas in the second example we set $\rho = 100$. Note that $1/(\log\alpha)^2 \simeq 0.0173$.

We implemented a program which simulates recombination and coalescence events during the selective sweep as explained in Section 2, without any approximation and using a discretization of the diffusion (1). We obtained the corresponding size of the hitchhiking set. We simulated 30 000 genealogies for each value of the set of parameters. The following tables and figures compare the empirical distribution with our approximation.

In Tables 1 and 2 we compare our approximation of the size of the hitchhiking set restricted to $\{0, \dots, m\}$ to the simulations with $\rho = 100$ and $\rho = 42$, respectively, when $n = 1$. For each length we give the value of $\mathrm{P}(H_1^r = h_1)$ according to our approximation and the estimation obtained from the 30 000 genealogies. The value for the last locus, $h_1 = 10$ for $\rho = 100$ and

TABLE 1: $\rho = 100$, $\alpha = 2000$, and $n = 1$

| $h_1$ | Approximation | Simulations | $\Delta \times (\log\alpha)^2$ |
|---|---|---|---|
| 1 | 0.336 | 0.321 | 0.826 |
| 2 | 0.240 | 0.220 | 1.086 |
| 3 | 0.161 | 0.150 | 0.635 |
| 4 | 0.104 | 0.103 | 0.029 |
| 5 | 0.065 | 0.067 | 0.087 |
| 6 | 0.041 | 0.045 | 0.277 |
| 7 | 0.025 | 0.032 | 0.393 |
| 8 | 0.016 | 0.022 | 0.306 |
| 9 | 0.011 | 0.012 | 0.075 |
| $\geq 10$ | 0.002 | 0.027 | 1.467 |

https://doi.org/10.1239/aap/1253281062 Published online by Cambridge University Press

TABLE 2: $\rho = 42$, $\alpha = 2000$, and $n = 1$

| $h_1$ | Approximation | Simulations | $\Delta \times (\log \alpha)^2$ |
|---|---|---|---|
| 1 | 0.149 | 0.149 | 0.006 |
| 2 | 0.132 | 0.130 | 0.081 |
| 3 | 0.115 | 0.109 | 0.335 |
| 4 | 0.099 | 0.096 | 0.179 |
| 5 | 0.084 | 0.082 | 0.150 |
| 6 | 0.071 | 0.066 | 0.295 |
| 7 | 0.060 | 0.059 | 0.075 |
| 8 | 0.050 | 0.048 | 0.150 |
| 9 | 0.042 | 0.041 | 0.040 |
| 10 | 0.035 | 0.034 | 0.052 |
| 11 | 0.029 | 0.031 | 0.110 |
| 12 | 0.024 | 0.025 | 0.046 |
| 13 | 0.020 | 0.020 | 0.000 |
| 14 | 0.016 | 0.018 | 0.104 |
| $\geq 15$ | 0.075 | 0.093 | 1.046 |

$h_1 = 15$ for $\rho = 42$, is $\mathrm{P}(H_1^r \geq h_1)$. We can see that our approximation is very close to the exact distribution.

In Tables 1 and 2 we also give the ratio of the absolute difference $\Delta$ between the two results over $1/(\log \alpha)^2$. With this rescaling, we can see that the constants hidden behind the $\mathcal{O}$s remain small and that our approximation performs well.

Note that the half-length of the 95% confidence interval due to the Monte Carlo method is bounded by $1.96\sqrt{0.5 \times 0.5/30\,000} = 0.0057 < 1/(\log \alpha)^2$, so the error due to the Monte Carlo method is not significant.

We have also checked that the order of the approximation is $\mathcal{O}(1/(\log \alpha)^2)$ when $n = 1$. To do this, we carried out the same calculations as in Tables 1 and 2 with different values for $\alpha$ while $\gamma = \rho \log \alpha / \alpha$ remained constant. We fixed $\gamma = 100 \times \log(2000)/2000 \simeq 0.38$, and we studied the cases $\alpha = 10\,000$, $\alpha = 2000$, $\alpha = 500$, and $\alpha = 200$. For all the sizes between 1 and 9, we obtained $\Delta \times (\log \alpha)^2 \leq 1.5$ (results not shown). So the approximation error of $\mathcal{O}(1/(\log \alpha)^2)$ cannot be improved.

Moreover, we also analysed the quality of our approximation when $n = 3$ and $n = 5$. Because the probabilities of the various $n$-tuples are generally smaller than $1/(\log \alpha)^2$, we calculated the probabilities of the events $\{H_1^r \geq h_1, \dots, H_n^r \geq h_n\}$ when $1 \leq h_1 \leq \cdots \leq h_n \leq 9$ for $\rho \in \{42, 100\}$, $n \in \{3, 5\}$, and $m = 8$. The results are presented in Tables 3 and 4.

Table 3 shows the repartition of the values of $\Delta \times (\log \alpha)^2$ for the 50 most probable events, the others being negligible, when $n = 3$ and $\rho = 100$, for the 78 most probable events when $n = 5$ and $\rho = 100$, and for the 162 most probable events when $n = 3$ and $\rho = 42$. Again, we can see that the approximation performs well.

Unfortunately, for the approximate model, as already noted in [14], the approximation does not perform well as soon as the size of the sample is greater than 5. This remark is also true for our approximation when $n = 5$ and $\rho = 42$, as shown by several examples in Table 4.

Moreover, a star-like approximation is sometimes used for the genealogy at the selected site [9], [16]. This approximation, whose order of accuracy is only $\mathcal{O}(1/(\log \alpha))$, leads to a simpler formula for the size of the hitchhiking set because recombinations independently

https://doi.org/10.1239/aap/1253281062 Published online by Cambridge University Press

Table 3: Repartition of the values of $\Delta \times (\log \alpha)^2$ for the most probable events, when $n = 3$ and $\rho = 100$, $n = 5$ and $\rho = 100$, and $n = 3$ and $\rho = 42$.

| $n$ | $\rho$ | Minimum | Maximum | Mean | Variance |
|---|---|---|---|---|---|
| 3 | 100 | 0 | 3.03 | 0.65 | 1.32 |
| 5 | 100 | 0 | 4.65 | 1.13 | 1.70 |
| 3 | 42 | 0 | 6.07 | 1.54 | 2.76 |

Table 4: Comparisons between our approximation and simulations when $\rho = 42$ and $n = 5$. We present values of $\mathrm{P}(H_1^r \geq h_1, \dots, H_5^r \geq h_5)$ for various values of $(h_1, \dots, h_5)$ and the absolute difference scaled by $(\log \alpha)^2$.

| $(h_1, \dots, h_5)$ | Theorem 2 | Simulations | $\Delta \times (\log \alpha)^2$ |
|---|---|---|---|
| (1,4,4,4,6) | 0.2331 | 0.1669 | 3.8225 |
| (2,2,3,4,5) | 0.3454 | 0.2437 | 5.8743 |
| (2,3,3,6,6) | 0.2098 | 0.1458 | 3.6981 |
| (3,4,4,6,8) | 0.0940 | 0.0765 | 1.0132 |
| (4,4,4,5,5) | 0.1778 | 0.1298 | 2.7772 |
| (4,5,6,7,8) | 0.0234 | 0.0446 | 1.2244 |
| (5,5,5,5,5) | 0.1188 | 0.0941 | 1.4269 |

impact the individuals of the sample: under the model presented in [9], the distribution of the size for each individual is geometric with parameter $1 - \exp(-\gamma)$, i.e.

$$\mathrm{P}(H_1 \geq h_1, \dots, H_n \geq h_n) = \exp\biggl(-\gamma \sum_{p=1}^{n} (h_p - 1)\biggr).$$

We compared this formula to our approximation and the simulations for various values of $n$ and $(h_1, \dots, h_n)$ (30 000 genealogies have been simulated in each case). We chose $\alpha = 2000$ and $\rho = 100$ ($\gamma = 0.38$). The results are presented in Table 5. We have to keep in mind that we do not know the constants in $\mathcal{O}(1/\log \alpha)$ and $\mathcal{O}(1/(\log \alpha)^2)$, so we are not sure that the error for our approximate model is smaller. However, the numerical results show that our

Table 5: Comparisons between our approximation, simulations, and the star-like model when $\alpha = 2000$ and $\rho = 100$. We present the value of $\mathrm{P}(H_1^r \geq h_1, \dots, H_n^r \geq h_n)$ for various values of $n$ and $(h_1, \dots, h_n)$ in the three cases.

| $n$ | $(h_1, \dots, h_n)$ | Theorem 2 | Simulations | Star-like model |
|---|---|---|---|---|
| 3 | (2,2,2) | 0.4414 | 0.3963 | 0.3198 |
| 3 | (3,3,3) | 0.1609 | 0.1558 | 0.1023 |
| 3 | (3,3,4) | 0.0964 | 0.1052 | 0.0699 |
| 3 | (2,4,4) | 0.0941 | 0.1011 | 0.0699 |
| 5 | (2,2,2,2,2) | 0.2735 | 0.2626 | 0.1496 |
| 5 | (2,2,2,2,4) | 0.0795 | 0.1227 | 0.0699 |
| 5 | (2,2,2,3,3) | 0.0956 | 0.1349 | 0.0699 |
| 5 | (2,2,2,3,4) | 0.0222 | 0.0922 | 0.0478 |

https://doi.org/10.1239/aap/1253281062 Published online by Cambridge University Press

approximation is in general better than the star-like tree approximation. This is due to the fact that a recombination may impact on an ancestor of several individuals of the sample. This dependence between individuals is not taken into account with the star-like model.

## 4. Construction of the approximate model and proof of Theorem 1

### 4.1. Evolution at the neutral sites

In this section we assume that $X = (X_t)_{0\le t\le T}$ is given. We explain how to model the transition events with independent Poisson processes.

The rates given in the following paragraph are those for a single lineage. The first idea would be to consider a single Poisson process per lineage. However, the number of lineages under consideration at any specific moment depends on the past events. So we will define independent Poisson processes such that at any given Poisson jump time, we will choose uniformly at random among the living lineages, the lineage undergoing transition at this moment. This choice is modeled by independent random variables $W$, given below. Since the events of type $(B \Rsh b, a)$ and $(b \Rsh b, a)$ have the same rate but occur under two mutually excluding conditions, we use the same Poisson process $\overleftarrow{\tau}_{\Rsh b,a}$ to account for both types. Note that we could also combine $(B \Rsh B, a)$ and $(b \Rsh B, a)$ in a similar Poisson process. However, as we will see in Section 4.2.1, the probability of events of type $b \Rsh B$ is negligible, so these events will not be taken into account in the approximate models.

Let $\overleftarrow{\tau}_{\mathrm{coal}B}$, $\overleftarrow{\tau}_{\mathrm{coal}b}$, $\overleftarrow{\tau}_{\Rsh b,a}$, $\overleftarrow{\tau}_{b\Rsh B,a}$, and $\overleftarrow{\tau}_{B\Rsh B,a}$, $-m+1 \le a \le m$, denote Poisson processes, which are conditionally independent given the random frequency path $(X_t)_{0\le t\le T}$. Since recombination and coalescence events are considered back in time, these processes are defined for the reversed time $\beta = T - t$ from $\beta = 0$ to $\beta = T$:

$$(\overleftarrow{\tau}_{\mathrm{coal}B}(\beta))_{0\le\beta\le T} \text{ has rate } \binom{(2m+1)n}{2}\frac{2}{X_{T-\beta}} \text{ at time } \beta,$$

$$(\overleftarrow{\tau}_{\mathrm{coal}b}(\beta))_{0\le\beta\le T} \text{ has rate } \binom{(2m+1)n}{2}\frac{2}{1-X_{T-\beta}} \text{ at time } \beta,$$

$$(\overleftarrow{\tau}_{\Rsh b,a}(\beta))_{0\le\beta\le T} \text{ has rate } n\rho_a(1-X_{T-\beta}) \text{ at time } \beta,\ -m+1 \le a \le m,$$

$$(\overleftarrow{\tau}_{b\Rsh B,a}(\beta))_{0\le\beta\le T},\ (\overleftarrow{\tau}_{B\Rsh B,a}(\beta))_{0\le\beta\le T} \text{ have rate } n\rho_a X_{T-\beta} \text{ at time } \beta,\ -m+1 \le a \le m.$$

We then denote by $\tau_{\mathrm{coal}B}$, $\tau_{\mathrm{coal}b}$, $\tau_{\Rsh b,a}$, $\tau_{b\Rsh B,a}$, and $\tau_{B\Rsh B,a}$ the Poisson processes defined for $t \in [0, T]$ by $\tau_{\mathrm{coal}B}(t) = \overleftarrow{\tau}_{\mathrm{coal}B}(T-t)$, $\tau_{\mathrm{coal}b}(t) = \overleftarrow{\tau}_{\mathrm{coal}b}(T-t)$, $\tau_{\Rsh b}(t) = \overleftarrow{\tau}_{\Rsh b}(T-t)$, $\tau_{b\Rsh B}(t) = \overleftarrow{\tau}_{b\Rsh B}(T-t)$, and $\tau_{B\Rsh B}(t) = \overleftarrow{\tau}_{B\Rsh B}(T-t)$.

Let

$$W_{\mathrm{coal}B} = (W^{(s)}_{\mathrm{coal}B})_{s\in\mathbb{N}^*}, \qquad W_{\mathrm{coal}b} = (W^{(s)}_{\mathrm{coal}b})_{s\in\mathbb{N}^*}, \qquad W_{\Rsh b,a} = (W^{(s)}_{\Rsh b,a})_{s\in\mathbb{N}^*},$$
$$W_{b\Rsh B,a} = (W^{(s)}_{b\Rsh B,a})_{s\in\mathbb{N}^*}, \quad \text{and} \quad W_{B\Rsh B,a} = (W^{(s)}_{B\Rsh B,a})_{s\in\mathbb{N}^*},$$

for $-m+1 \le a \le m$ be sequences of independent discrete variables, specified as follows: $W^{(s)}_{\mathrm{coal}B}$ and $W^{(s)}_{\mathrm{coal}b}$ are uniform on the set of pairs of elements of $\{-mn+1, \dots, (m+1)n\}$, with cardinality

$$\binom{(2m+1)n}{2},$$

while $W^{(s)}_{\Rsh b,a}$, $W^{(s)}_{B\Rsh b,a}$, and $W^{(s)}_{B\Rsh B,a}$ are uniform on $\{an+1, \dots, an+n\}$, $-m+1 \le a \le m$.

We assume that $W^{(s)}_{\mathrm{coal}B}$, $W^{(s)}_{\mathrm{coal}b}$, $W^{(s)}_{\Rsh b,a}$, $W^{(s)}_{b\Rsh B,a}$, and $W^{(s)}_{B\Rsh B,a}$, $-m+1 \le a \le m$, $s \in \mathbb{N}^*$, are also independent of $(X, \tau_{\mathrm{coal}B}, \tau_{\mathrm{coal}b}, \tau_{\Rsh b,a}, \tau_{b\Rsh B,a}, \tau_{B\Rsh B,a}, -m+1 \le a \le m)$.

To construct the ancestral genealogy, we consider the successive jumps of the Poisson processes. More precisely, suppose that a jump occurs at $t = t_0$ (i.e. $\beta = T - t_0$ for the reversed time). Let $\pi(t_0^+)$ be the genealogy just before the jump. Denote by $\pi_j$ the block of $\pi(t_0^+)$ containing allele $j$, which will be identified with the ancestor of $j$ at $t = t_0^+$.

If the jump corresponds to the $s$th jump of $\overleftarrow{\tau}_{\mathrm{coal}B}$ or $\overleftarrow{\tau}_{\mathrm{coal}b}$, we choose a pair $(j, k)$ of alleles according to the realization of $W^{(s)}_{\mathrm{coal}B}$ or, respectively, $W^{(s)}_{\mathrm{coal}b}$. If $j$ and $k$ are respectively the smallest elements of $\pi_j$ and $\pi_k$, and if both $\pi_j$ and $\pi_k$ carry $B$ or $b$, i.e. both or, respectively, none of them are marked, then the two blocks coalesce and the new block is marked or, respectively, not marked. Otherwise, nothing happens.

Note that we only consider the case where $j$ and $k$ are the smallest elements in order to avoid size biasing.

If the jump corresponds to the $s$th jump of $\overleftarrow{\tau}_{\Rsh b,a}$, we choose an allele $j$ from the realization of $W^{(s)}_{\Rsh b,a}$. If $j$ is the smallest element of $\pi_j \cap \{an+1, \dots, an+n\}$ then $\pi_j$ splits into $\pi_j \cap \{-mn+1, \dots, an\}$ and $\pi_j \cap \{an+1, \dots, (m+1)n\}$. When $\pi_j$ carries $B$ or $b$ and $a > 0$, only the first block or, respectively, none of the blocks will be marked. When $\pi_j$ carries $B$ or $b$ and $a \le 0$, only the second block or, respectively, none of the blocks will be marked. This event is of type $(B \Rsh b, a)$ or, respectively, $(b \Rsh b, a)$.

Finally, if the jump corresponds to the $s$th jump of $\overleftarrow{\tau}_{b\Rsh B,a}$ or $\overleftarrow{\tau}_{B\Rsh B,a}$, we choose an allele $j$ from the realization of $W^{(s)}_{b\Rsh B,a}$ or, respectively, $W^{(s)}_{B\Rsh B,a}$. If $j$ is the smallest element of $\pi_j \cap \{an+1, \dots, an+n\}$ and if $\pi_j$ carries $b$ or, respectively, $B$, then $\pi_j$ splits into $\pi_j \cap \{-mn+1, \dots, an\}$ and $\pi_j \cap \{an+1, \dots, (m+1)n\}$. When $a > 0$, only the second block will be marked when $\pi_j$ carries $b$ and both of the blocks will be marked when $\pi_j$ carries $B$. When $a \le 0$, only the first block will be marked when $\pi_j$ carries $b$ and both of the blocks will be marked when $\pi_j$ carries $B$.

In this way, we obtain $\pi(t_0)$ from $\pi(t_0^+)$ and $(W_{\mathrm{coal}B}, W_{\mathrm{coal}b}, W_{\Rsh b,a}, W_{b\Rsh B,a}, W_{B\Rsh B,a})$. An example is given in Figure 6.

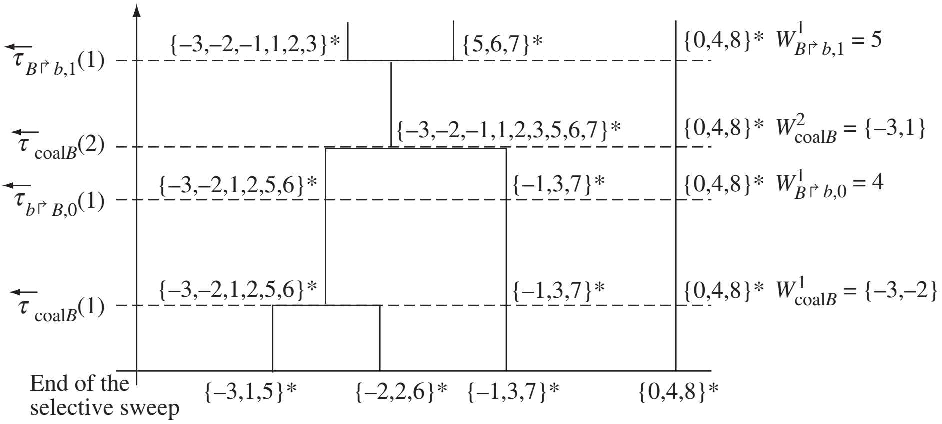

$\overleftarrow{\tau}_{B\Rsh b,1}(1)$
$\overleftarrow{\tau}_{\mathrm{coal}B}(2)$
$\overleftarrow{\tau}_{b\Rsh B,0}(1)$
$\overleftarrow{\tau}_{\mathrm{coal}B}(1)$
End of the selective sweep
{−3,−2,−1,1,2,3}*
{5,6,7}*
{0,4,8}*
$W^1_{B\Rsh b,1} = 5$
{−3,−2,−1,1,2,3,5,6,7}*
{0,4,8}*
$W^2_{\mathrm{coal}B} = \{-3,1\}$
{−3,−2,1,2,5,6}*
{−1,3,7}*
{0,4,8}*
$W^1_{B\Rsh b,0} = 4$
{−3,−2,1,2,5,6}*
{−1,3,7}*
{0,4,8}*
$W^1_{\mathrm{coal}B} = \{-3,-2\}$
{−3,1,5}*
{−2,2,6}*
{−1,3,7}*
{0,4,8}*

FIGURE 6: An example of evolution. Here $m = 1$ and $n = 4$, so that the alleles at loci $-1$, 0, and 1 of four individuals are involved. Note that the second event (looking back in time) does not modify the partition since the impacted block is marked.

Let $E$ be the set of Poisson processes defined on $[0, T]$, and let $\mathcal{P}$ be the set of partitions of $\{-mn+1, \dots, (m+1)n\}$. Given $(X_t)_{0 \le t \le T}$, this procedure specifies a map

$$f \colon E \times E \times (E \times E \times E)^{2m} \times (\mathbb{N}^2)^{\mathbb{N}^*} \times (\mathbb{N}^2)^{\mathbb{N}^*} \times (\mathbb{N}^{\mathbb{N}^*} \times \mathbb{N}^{\mathbb{N}^*} \times \mathbb{N}^{\mathbb{N}^*})^{2m} \mapsto \mathcal{P}$$

such that

$$f(\tau_{\mathrm{coal}B}, \tau_{\mathrm{coal}b}, \tau_{\Rsh b,a}, \tau_{b\Rsh B,a}, \tau_{B\Rsh B,a}, W_{\mathrm{coal}B}, W_{\mathrm{coal}b}, W_{\Rsh b,a}, W_{b\Rsh B,a}, W_{B\Rsh B,a},\\ -m+1 \le a \le m) = \pi(0).$$

Now we successively construct several simplified models, where in the final model we no longer need to compute $(X_t)_{0 \le t \le T}$ to obtain a sample of ancestral partitions from the distribution of $\pi(0)$.

From now on, by *background* $B$ and *background* $b$ we mean the subpopulation carrying allele $B$ and allele $b$, respectively.

### 4.2. Suppression of rare events

In this section, a path $X$ is given and $\mathrm{P}^X$ denotes the conditional probability given $X$, while P denotes the unconditional probability.

4.2.1. *Events* coal$b$ *and* $b \Rsh B$ *are negligible.* First, we see why it is possible to ignore the occurrence of events coal$b$ and $b \Rsh B$ in the approximate models.

**Proposition 1.** *The probability of events of type* coal$b$ *and* $b \Rsh B$ *is* $\mathcal{O}(1/(\log\alpha)^2)$.

*Proof.* The following two statements are proved in [2, Proposition 3.4.].

1. The probability that two lineages coalesce in background $b$ is $\mathcal{O}(1/(\log\alpha)^2)$.
2. The probability that, looking in reversed time, a lineage goes from background $B$ to background $b$ and then goes back to background $B$ is also $\mathcal{O}(1/(\log\alpha)^2)$.

The events of type $b \Rsh B$ bring lineages from background $b$ to background $B$. However, since the sample was taken at the end of the selective sweep, all the lineages began in background $B$, and considering an individual in background $b$ implicitly implies that it has already moved at least once from background $B$ to background $b$. Consequently, owing to statement 2, we find that the probability of an event $b \Rsh B$ is $\mathcal{O}(1/(\log\alpha)^2)$.

**Remark 3.** Our simulations (not shown) indicate that the error induced by the approximation in Proposition 1 is truly of order $\mathcal{O}(1/(\log\alpha)^2)$, and not of a smaller order; so the accuracy of our model cannot actually be improved.

**Corollary 1.** *Let* $\Theta \in \mathcal{P}$. *Let* $\tilde{f} \colon E \times (E \times E)^{2m} \times (\mathbb{N}^2)^{\mathbb{N}^*} \times (\mathbb{N}^{\mathbb{N}^*} \times \mathbb{N}^{\mathbb{N}^*})^{2m} \mapsto \mathcal{P}$ *be the map defined by*

$$\begin{aligned}\tilde{f}(&\tau_{\mathrm{coal}B}, \tau_{\Rsh b,a}, \tau_{B\Rsh B,a}, W_{\mathrm{coal}B}, W_{\Rsh b,a}, W_{B\Rsh B,a}, -m+1 \le a \le m)\\ &= f(\tau_{\mathrm{coal}B}, \tau_{\varnothing}, \tau_{\Rsh b,a}, \tau_{\varnothing}, \tau_{B\Rsh B,a}, W_{\mathrm{coal}B}, \omega_1, W_{\Rsh b,a}, \omega_2, W_{B\Rsh B,a}, -m+1 \le a \le m),\end{aligned} \tag{5}$$

*where* $\tau_{\varnothing}$ *denotes a Poisson process without jumps,* $\omega_1 \in \mathbb{N}^{\mathbb{N}^*}$ *is such that, for all* $s \in \mathbb{N}^*$, $\omega_1(s) = 0$, *and* $\omega_2 \in (\mathbb{N} \times \mathbb{N})^{\mathbb{N}^*}$ *is such that, for all* $s \in \mathbb{N}^*$, $\omega_2(s) = (0, 1)$. *Then*

$$\mathrm{P}^X(\tilde{f}^{-1}(\Theta)) = \mathrm{P}^X(f^{-1}(\Theta)) + \mathcal{O}\biggl(\frac{1}{(\log\alpha)^2}\biggr).$$

The choices of $\omega_1$ and $\omega_2$ are arbitrary. Indeed, the corresponding Poisson processes have no jump, so the values of $\omega_1$ and $\omega_2$ do not influence the realization constructed by (5).

4.2.2. *Order of the events.* Let us first focus on events of type $B \Rsh B$, and show that we can assume that they all occur before any of the other events (still looking back in time). To see this, we define $\varepsilon = (\log\alpha)^2/\alpha$ and $T_\varepsilon = \inf\{t \geq 0 \colon X_t = \varepsilon\}$. Since we consider large values of $\alpha$, note that $\varepsilon$ is small.

In the same spirit as Equations (5.10), (5.11), and (5.12) of [13], we have the following proposition.

**Proposition 2.** *The Poisson processes described in Section 4.1 have the following properties:*

$$\mathrm{P}(\tau_{\mathrm{coal}B} \cap [T_\varepsilon, T] \neq \varnothing) = \mathcal{O}\biggl(\frac{1}{(\log\alpha)^2}\biggr), \tag{6}$$

$$\mathrm{P}\biggl(\bigcup_{a=-m+1}^{m} \{\tau_{B\Rsh B,a} \cap [0, T_\varepsilon] \neq \varnothing\}\biggr) = \mathcal{O}\biggl(\frac{(\log\alpha)^2}{\alpha}\biggr) \ll \mathcal{O}\biggl(\frac{1}{(\log\alpha)^2}\biggr), \tag{7}$$

$$\mathrm{P}\Bigl(\min_{-m+1\leq a\leq m}(\min \tau_{B\Rsh B,a}) < \max_{-m+1\leq a\leq m} \max(\tau_{\Rsh b,a})\Bigr) = \mathcal{O}\biggl(\frac{1}{(\log\alpha)^2}\biggr). \tag{8}$$

We use the notation $H = n\sum_{a=-m+1}^{m} \gamma_a$, which is a constant that depends only on $n$ and $m$.

*Proof of Proposition 2.* Equation (6) is proved in [2, Lemma 4.3]. To prove (7), we use Jensen's inequality to compute

$$\begin{aligned}
\mathrm{P}\biggl(\bigcup_{a=-m+1}^{m} \tau_{B\Rsh B,a} \cap [0, T_\varepsilon] = \varnothing\biggr) &= \mathrm{E}\biggl[\exp\biggl(-n\sum_{a=-m+1}^{m} \rho_a \int_0^{T_\varepsilon} X_s \,\mathrm{d}s\biggr)\biggr] \\
&\geq \exp\biggl(-n\sum_{a=-m+1}^{m} \rho_a \varepsilon\, \mathrm{E}[T_\varepsilon]\biggr) \\
&= \exp\biggl(-H\frac{(\log\alpha)^2}{\alpha} + \mathcal{O}\biggl(\frac{\log\alpha}{\alpha}\biggr)\biggr) \\
&\geq 1 - \mathcal{O}\biggl(\frac{(\log\alpha)^2}{\alpha}\biggr).
\end{aligned}$$

For (8), we write

$$\begin{aligned}
&\mathrm{P}\Bigl(\min_{-m+1\leq a\leq m}(\min \tau_{B\Rsh B,a}) < \max_{-m+1\leq a\leq m}(\max \tau_{\Rsh b,a})\Bigr) \\
&\quad = \mathrm{P}\biggl(\min \bigcup_{a=-m+1}^{m} \tau_{B\Rsh B,a} < \max \bigcup_{a=-m+1}^{m} \tau_{\Rsh b,a}\biggr).
\end{aligned}$$

Given $X$, all the Poisson processes are independent, so $\bigcup_{a=-m+1}^{m} \tau_{B\Rsh B,a}$ and $\bigcup_{a=-m+1}^{m} \tau_{\Rsh b,a}$ are independent Poisson processes. From now on, $|\cdot|$ denotes the cardinality. We have

$$\begin{aligned}
&\mathrm{P}\Bigl(\min_{-m+1\leq a\leq m}(\min \tau_{B\Rsh B,a}) < \max_{-m+1\leq a\leq m}(\max \tau_{\Rsh b,a})\Bigr) \\
&\quad = \mathrm{E}\biggl[\int_0^T \mathrm{P}^X\biggl(\biggl|\bigcup_{a=-m+1}^{m} \tau_{B\Rsh B,a} \cap [0, t]\biggr| > 0\biggr) \mathrm{P}^X\biggl(\max \bigcup_{a=-m+1}^{m} \tau_{\Rsh b,a} \in dt\biggr)\biggr]
\end{aligned}$$

https://doi.org/10.1239/aap/1253281062 Published online by Cambridge University Press

$$
\begin{aligned}
&= \mathrm{E}\bigg[\int_0^T \bigg(1-\exp\bigg(-\int_0^t H\frac{\alpha}{\log\alpha} X\,\mathrm{d}s\bigg)\bigg)\\
&\qquad\qquad \times \exp\bigg(-\int_t^T H\frac{\alpha}{\log\alpha}(1-X_s)\,\mathrm{d}s\bigg) H\frac{\alpha}{\log\alpha}(1-X_t)\,\mathrm{d}t\bigg] && (9)\\
&\le H^2\frac{\alpha^2}{(\log\alpha)^2}\,\mathrm{E}\bigg[\int_0^T \bigg(\int_0^t X_s\,\mathrm{d}s\bigg)(1-X_t)\,\mathrm{d}t\bigg] && (10)\\
&= \mathcal{O}\bigg(\frac{1}{(\log\alpha)^2}\bigg).
\end{aligned}
$$

Equation (9) follows from the fact that, under P, $|\bigcup_{a=-m+1}^{m} \tau_{B\Rsh B,a} \cap [0,t]|$ is Poisson with parameter

$$\int_0^t n \sum_{a=-m+1}^{m} \rho_a X_s\,\mathrm{d}s = \int_0^t H\frac{\alpha}{\log\alpha} X_s\,\mathrm{d}s$$

and $|\bigcup_{a=-m+1}^{m} \tau_{\Rsh b,a} \cap [t,T]|$ is Poisson with parameter $\int_t^T H(\alpha/\log\alpha)(1-X_s)\,\mathrm{d}s$.

The expectation in (10) is estimated in [2] (see Equation (4.5) therein and the following estimates) and the result follows.

Proposition 2 immediately implies the following result.

**Corollary 2.** *With probability* $1-\mathcal{O}(1/(\log\alpha)^2)$, *all the events of type $B \Rsh B$ occur before all the other events.*

**Remark 4.** The presence of $H^2$ in (10) shows that the sample size and the number of loci enter, at least quadratically, the global error term $\mathcal{O}(1/(\log\alpha)^2)$.

4.2.3. *Interchangeability of the events of type $\Rsh b$ or $B \Rsh B$.* The following proposition is similar to Equation (5.9) of [13], but extended to events of type $B \Rsh B$.

**Proposition 3.** *Let $\tau = \tau_{\mathrm{coal}B} \bigcup_{a=-m+1}^{m} (\tau_{\Rsh b,a} \cup \tau_{B\Rsh B,a})$. This union is called the superposition of Poisson processes. Let $s$ and $t$ be two consecutive points of $\tau$ such that $s \in \tau_{\Rsh b,c}$ and $t \in \tau_{\Rsh b,d}$, with $c < d$ (we do not impose $s < t$). Let $\tau'_{\Rsh b,c} = (\tau_{\Rsh b,c} \setminus \{s\}) \cup \{t\}$, $\tau'_{\Rsh b,d} = (\tau_{\Rsh b,d} \setminus \{t\}) \cup \{s\}$, and $\tau'_{\Rsh b,a} = \tau_{\Rsh b,a}$ for all $a \notin \{c,d\}$. Then*

$$
\begin{aligned}
&\tilde f(\tau_{\mathrm{coal}B}, \tau_{\Rsh b,a}, \tau_{B\Rsh B,a}, W_{\mathrm{coal}B}, W_{\Rsh b,a}, W_{B\Rsh B,a},\ -m+1 \le a \le m)\\
&\quad = \tilde f(\tau_{\mathrm{coal}B}, \tau'_{\Rsh b,a}, \tau_{B\Rsh B,a}, W_{\mathrm{coal}B}, W_{\Rsh b,a}, W_{B\Rsh B,a},\ -m+1 \le a \le m).
\end{aligned}
$$

*The same result can also be shown for two consecutive events of type $B \Rsh B$.*

*Proof.* Assume that $s$ is the $x$th point of $\tau_{\Rsh b,c}$ and that $t$ is the $y$th point of $\tau_{\Rsh b,d}$. Write $j = W^{(x)}_{\Rsh b,c}$ and $k = W^{(y)}_{\Rsh b,d}$. Let $\pi_j$ and $\pi_k$ be the blocks respectively containing $j$ and $k$ at time $(s \vee t)^+$.

Consider $\pi_j \neq \pi_k$. Even if the two recombinations are realized, the first one splits one of the blocks—for example, $\pi_j$—into two blocks that are still different from $\pi_k$. Consequently, it has no effect on the realization of the second recombination.

Consider $\pi_j = \pi_k$. If $j > \min(\pi_j \cap \{cn+1, \dots, cn+n\})$ and if $k > \min(\pi_k \cap \{dn+1, \dots, dn+n\})$, no recombination is realized.

Now suppose that $j = \min(\pi_j \cap \{cn+1, \dots, cn+n\})$ and $k > \min(\pi_k \cap \{dn+1, \dots, dn+n\})$. If we begin with the recombination $\tau_{\Rsh b,c}$, $\pi_j$ splits into $\pi_j \cap \{-mn+1, \dots, cn\}$ and

$\pi_j \cap \{cn+1, \dots, (m+1)n\}$. We have $c < d$, so $k \in \pi_j \cap \{cn+1, \dots, (m+1)n\}$ and we still have

$$k > \min(\pi_j \cap \{cn+1, \dots, (m+1)n\} \cap \{dn+1, \dots, dn+n\}) = \min(\pi_j \cap \{dn+1, \dots, dn+n\}).$$

The recombination at locus $d$ is not realized. When we change the order or the two recombinations, the first one at locus $d$ is not realized since $k > \min(\pi_k \cap \{dn+1, \dots, dn+n\})$ and the one at locus $c$ is realized. So, the result is the same.

The arguments are the same when $j > \min(\pi_j \cap \{cn+1, \dots, cn+n\})$ and $k = \min(\pi_k \cap \{dn+1, \dots, dn+n\})$, and when $j = \min(\pi_j \cap \{cn+1, \dots, cn+n\})$ and $k = \min(\pi_k \cap \{dn+1, \dots, dn+n\})$.

By (6)–(8), with probability $1 - \mathcal{O}(1/(\log\alpha)^2)$, events of types $\Rsh b$ and $B \Rsh B$ restricted to $[T_\varepsilon, T]$ are consecutive. Applying Proposition 3, we find that the order of events $\Rsh b$ and $B \Rsh B$ is not important during $[T_\varepsilon; T]$, that is, all the information is given by $|\tau_{\Rsh b,a} \cap [T_\varepsilon; T]|$ and $|\tau_{B \Rsh B,a} \cap [T_\varepsilon; T]|$, $-m+1 \le a \le m$.

Recall that $E$ is the set of Poisson processes defined on $[0, T]$. We denote by $E_\varepsilon$ the set of Poisson processes restricted to $[0, T_\varepsilon]$.

Let $\tau_1 \in E_\varepsilon$. Let $\tau^0_{\mathrm{coal}B} \in E$ such that $\tau^0_{\mathrm{coal}B} \cap [0; T_\varepsilon] = \tau_1$ and $|\tau^0_{\mathrm{coal}B} \cap [T_\varepsilon, T]| = 0$.

For $-m+1 \le a \le m$, let $\tau_{2,a} \in E_\varepsilon$, $c_a \in \mathbb{N}$, and $d_a \in \mathbb{N}$. Let $\tau^0_{\Rsh b,a} \in E$ such that $\tau^0_{\Rsh b,a} \cap [0; T_\varepsilon] = \tau_{2,a}$ and $|\tau^0_{\Rsh b,a} \cap [T_\varepsilon, T]| = c_a$. Let $\tau^0_{B\Rsh B,a} \in E$ such that $|\tau^0_{B\Rsh B,a} \cap [0; T_\varepsilon]| = 0$ and $|\tau^0_{B\Rsh B} \cap [T_\varepsilon, T]| = d_a$.

We define the map $g \colon E_\varepsilon \times (E_\varepsilon \times \mathbb{N} \times \mathbb{N})^{2m} \times (\mathbb{N}^2)^{\mathbb{N}^*} \times (\mathbb{N}^{\mathbb{N}^*} \times \mathbb{N}^{\mathbb{N}^*})^{2m} \mapsto \mathcal{P}$ such that

$$\begin{aligned} &g(\tau_1, \tau_{2,a}, c_a, d_a, W_{\mathrm{coal}B}, W_{\Rsh b,a}, W_{B\Rsh B,a},\ -m+1 \le a \le m) \\ &\qquad = \tilde{f}(\tau^0_{\mathrm{coal}B}, \tau^0_{\Rsh b,a}, \tau^0_{B\Rsh B,a}, W_{\mathrm{coal}B}, W_{\Rsh b,a}, W_{B\Rsh B,a},\ -m+1 \le a \le m). \end{aligned}$$

**Corollary 3.** *For all $\Theta \in \mathcal{P}$,*

$$\mathrm{P}^X(g^{-1}(\Theta)) = \mathrm{P}^X(\tilde{f}^{-1}(\Theta)) + \mathcal{O}\biggl(\frac{1}{(\log\alpha)^2}\biggr).$$

4.2.4. *Approximate independence.* In this section we explain why the events of type $(B \Rsh B, a)$ can be approximated by the realizations (before any other event) of independent Bernoulli random variables with parameter $1 - \exp(\mathrm{E}[-\rho_a \int_0^T X_s \,\mathrm{d}s])$, where here E denotes the integration over all the paths $(X_t)_{0 \le t \le T}$.

We first show that the number of $B \Rsh B$ events is approximately independent of the number of events of type $\Rsh b$ (events of type $\Rsh b$ are the other possible events occurring during the time interval $[T_\varepsilon; T]$). We do not condition upon $X$ anymore. More precisely, we have the following result.

**Proposition 4.** *For any $u_a, v_a \in \mathbb{N}$, $-m+1 \le a \le m$, we have*

$$\begin{aligned} &\mathrm{P}(|\tau_{\Rsh b,a} \cap [T_\varepsilon; T]| = u_a,\ |\tau_{B\Rsh B,a} \cap [T_\varepsilon; T]| = v_a;\ -m+1 \le a \le m) \\ &\quad = \prod_{a=-m+1}^{m} (\mathrm{P}(|\tau_{\Rsh b,a} \cap [T_\varepsilon; T]| = u_a)\, \mathrm{P}(|\tau_{B\Rsh B,a} \cap [T_\varepsilon; T]| = v_a)) \\ &\qquad + \mu(u_a, v_a;\ -m+1 \le a \le m)\,\mathcal{O}\biggl(\frac{1}{(\log\alpha)^2}\biggr), \end{aligned}$$

*where* $\sum_{u_{-m+1},\dots,u_m,v_{-m+1},\dots,v_m=0}^{\infty} \mu(u_a, v_a,\ -m+1 \le a \le m) < \infty$.

The proof of this proposition, which is quite technical, is given in Section 6. We already know that recombination and coalescence events are independent. Thus, the events $B \Rsh B$ can be realized independently of all the other events.

Combined with the strong Markov property, Proposition 4 leads to the following result.

**Corollary 4.** *For any* $\Theta \in \mathcal{P}$,

$$\begin{aligned}
&\mathrm{P}((\tau_{\mathrm{coal}B} \cap [0; T_\varepsilon], \tau_{\Rsh b,a} \cap [0; T_\varepsilon], |\tau_{\Rsh b,a} \cap [T_\varepsilon; T]|, |\tau_{B\Rsh B,a} \cap [T_\varepsilon; T]|,\\
&\quad W_{\mathrm{coal}B}, W_{\Rsh b,a}, W_{B\Rsh B,a}, -m+1 \le a \le m) \in g^{-1}(\Theta))\\
&\quad = \sum \biggl( \mathrm{P}(\tau_{\mathrm{coal}B} \cap [0; T_\varepsilon] = \eta_{\mathrm{coal}B},\ \tau_{\Rsh b,a} \cap [0; T_\varepsilon] = \eta_{\Rsh b,a},\ -m+1 \le a \le m)\\
&\qquad\qquad \times \prod_{a=-m+1}^{m} \mathrm{P}(|\tau_{\Rsh b,a} \cap [T_\varepsilon; T]| = u_a)\,\mathrm{P}(|\tau_{B\Rsh B,a} \cap [T_\varepsilon; T]| = v_a)\\
&\qquad\qquad \times \mathrm{P}(W_{\mathrm{coal}B} = w_{\mathrm{coal}B},\ W_{\Rsh b,a} = w_{\Rsh b},\ W_{B\Rsh B,a} = w_{B\Rsh B},\ -m+1 \le a \le m) \biggr)\\
&\qquad + \mathcal{O}\biggl(\frac{1}{(\log(\alpha))^2}\biggr),
\end{aligned}$$

*where the sum is over* $\{(\eta_{\mathrm{coal}B}, \eta_{\Rsh b,a}, u_a, v_a, w_{\mathrm{coal}B}, w_{\Rsh b}, w_{B\Rsh B}, -m+1 \le a \le m) \in g^{-1}(\Theta)\}$.

We now prove that, for fixed $a$, the cardinality of $\tau_{B\Rsh B,a}$ approximately follows a Poisson distribution with parameter $\mathrm{E}[n\rho_a \int_0^T X_s\,\mathrm{d}s]$, denoted by $P_{\mathrm{E}[n\rho_a \int_0^T X_s\,\mathrm{d}s]}$. More precisely, we have the following result.

**Proposition 5.** *For all* $-m+1 \le a \le m$, *the total variation distance between the laws of* $|\tau_{B\Rsh B,a}|$ *and* $P_{\mathrm{E}[n\rho_a \int_0^T X_s\,\mathrm{d}s]}$ *is* $\mathcal{O}(1/(\log\alpha)^2)$*:*

$$\frac{1}{2}\sum_{k\in\mathbb{N}} \biggl| \mathrm{P}(|\tau_{B\Rsh B,a}| = k) - \frac{\mathrm{e}^{-\mathrm{E}[p]}\,\mathrm{E}[(p)^k]}{k!} \biggr| = \mathcal{O}\biggl(\frac{1}{(\log\alpha)^2}\biggr),$$

*where* $p = n\rho_a \int_0^T X_s\,\mathrm{d}s$.

*Proof.* Fix $k \in \mathbb{N}$. From the Taylor expansion of order 2 between $p$ and $\mathrm{E}[p]$, there exists a $\tilde{p}$ such that $|\tilde{p} - \mathrm{E}[p]| \le |p - \mathrm{E}[p]|$ and

$$\begin{aligned}
\mathrm{P}(|\tau_{B\Rsh B,a}| = k) &= \mathrm{E}\biggl[\frac{\mathrm{e}^{-p}p^k}{k!}\biggr]\\
&= \mathrm{E}\biggl[\frac{\mathrm{e}^{-\mathrm{E}[p]}\,\mathrm{E}(p)^k}{k!} + (p - \mathrm{E}[p])\frac{\partial}{\partial p}\frac{\mathrm{e}^{-p}p^k}{k!}\bigg|_{p=\mathrm{E}[p]}\\
&\qquad + \frac{1}{2}(p - \mathrm{E}[p])^2 \frac{\partial^2}{\partial p^2}\frac{\mathrm{e}^{-p}p^k}{k!}\bigg|_{p=\tilde{p}}\biggr]\\
&= \frac{\mathrm{e}^{-\mathrm{E}[p]}\,\mathrm{E}[(p)^k]}{k!}\\
&\quad + \frac{1}{2}\mathrm{E}\biggl[(p - \mathrm{E}[p])^2 \mathrm{e}^{-\tilde{p}}\biggl(\frac{\tilde{p}^{k-2}}{(k-2)!}\mathbf{1}_{\{k\ge2\}} - 2\frac{\tilde{p}^{k-1}}{(k-1)!}\mathbf{1}_{\{k\ge1\}} + \frac{\tilde{p}^k}{k!}\biggr)\biggr],
\end{aligned}$$

$$\begin{aligned}
\sum_{k\in\mathbb{N}} &\left| \mathrm{P}(|\tau_{B\Rsh B,a}| = k) - \frac{\mathrm{e}^{-\mathrm{E}[p]}\,\mathrm{E}[(p)^k]}{k!} \right| \\
&\le \frac{1}{2}\,\mathrm{E}\biggl[(p-\mathrm{E}[p])^2\mathrm{e}^{-\tilde{p}}\biggl(\sum_{k\ge 2}\frac{\tilde{p}^{k-2}}{(k-2)!} + 2\sum_{k\ge 1}\frac{\tilde{p}^{k-1}}{(k-1)!} + \sum_{k\ge 0}\frac{\tilde{p}^{k}}{k!}\biggr)\biggr] \\
&\le 2\,\mathrm{E}[(p-\mathrm{E}[p])^2] \\
&= 2\,\mathrm{var}(p).
\end{aligned}$$

Moreover, using a proof similar to that of [13, Equation (5.22)], we obtain

$$\mathrm{var}(p) = \mathrm{var}\biggl(n\rho_a\int_0^T X_s\,\mathrm{d}s\biggr) \le 2n^2\rho_a^2\,\mathrm{var}(T) = \mathcal{O}\biggl(\frac{1}{(\log\alpha)^2}\biggr).$$

Furthermore, since the events of type $B \Rsh B$ occur before all the other events, they only take place for the individuals of the sample. For each event, the implied individual is chosen uniformly in the sample and the corresponding split is done. Consequently, $P_{\mathrm{E}[n\rho_a\int_0^T X_s\,\mathrm{d}s]}$ can be seen as the superposition of $n$ independent Poisson variables whose rate is

$$\mathrm{E}\biggl[\rho_a\int_0^T X_s\,\mathrm{d}s\biggr],$$

one for each individual of the sample.

**Corollary 5.** *Events of type $B \Rsh B$ can be approximated by realizations of $U_{a,p}$, $-m+1 \le a \le m$, $1 \le p \le n$, independent Bernoulli variables with parameter* $1-\exp(\mathrm{E}[-\rho_a\int_0^T X_s\,\mathrm{d}s])$.

*Proof.* The impact of the Poisson process $\overleftarrow{\tau}_{B\Rsh B,a}$ on the evolution of the genealogy is the same as long as at least one of its event points happens, because once the split has occurred, another event will not change anything: one event or several events of type $B \Rsh B$ for the same $(a, p)$ produce the same result. This is why we can use $(U_{a,p},\ -m+1 \le a \le m,\ 1 \le p \le n)$ independent Bernoulli variables, $U_{a,1}, \dots, U_{a,n}$, with parameter

$$1 - \mathrm{P}(P_{\mathrm{E}[\rho_a\int_0^T X_s\,\mathrm{d}s]} = 0) = 1 - \exp\biggl(-\mathrm{E}\biggl[\rho_a\int_0^T X_s\,\mathrm{d}s\biggr]\biggr).$$

If $U_{a,p} = 1$, the event described in event $(B \Rsh B, a)$ is realized for individual $p$. If $U_{a,p} = 0$, nothing happens. For a given individual, the order of the realizations is not important because the partition becomes finer and finer.

Finally, we obtain the following model.

1. Start with the partition $\pi = \{\{-mn+1, -(m-1)n+1, \dots, mn+1\}^*, \dots, \{-mn+n, -(m-1)n+n, \dots, mn+n\}^*\}$.

2. Draw a path $(X_t)_{0\le t\le T}$ as in (1).

3. Draw $(U_{a,p})_{-m+1\le a\le m,\,1\le p\le n}$ independent Bernoulli variables with parameter

$$1 - \mathrm{E}\biggl[\rho_a\int_0^T X_s\,\mathrm{d}s\biggr].$$

   If $U_{a,p} = 1$, realize the event $(B \Rsh B, a)$ for individual $p$.
   A new partition $\tilde{\pi}'$ is obtained. Note that $|\tilde{\pi}'| = |\pi| + \sum_{a,p} U_{a,p}$.

4. Use this new partition from $t = T$ to construct the genealogy until $t = 0$, via Poisson processes $\overleftarrow{\tau}_{\mathrm{coal}B,a}$ and $\overleftarrow{\tau}_{\Rsh b,a}$, $-m+1 \le a \le m$.

5. This gives us a random approximate ancestral partition $\tilde{\pi}(0)$, which depends on the path $(X_t)_{0 \le t \le T}$.

6. The distribution of the approximate ancestral partition, obtained by integrating the law of $\tilde{\pi}(0)$ over all the paths $X = (X_t)_{0 \le t \le T}$, is denoted by $\tilde{\Gamma}$: for any set $A$ of partitions of $\{-mn+1, \dots, (m+1)n\}$, $\mathrm{P}(\tilde{\Gamma} \in A) = \mathrm{E}[\mathrm{P}^X(\tilde{\pi}(0) \in A)]$.

**Theorem 3.** *We have* $d_{\mathrm{TV}}(\Gamma, \tilde{\Gamma}) = \mathcal{O}(1/(\log\alpha)^2)$.

*Proof.* Since the number of partitions of $\{-mn+1, \dots, (m+1)n\}$ is finite, it is enough to prove that, for a fixed partition $\Theta$, we have $|\,\mathrm{P}(\Gamma = \Theta) - \mathrm{P}(\tilde{\Gamma} = \Theta)| = \mathcal{O}(1/(\log\alpha)^2)$.

In the following equation, E corresponds to the integration over the random paths $X$. Recall that $\pi(0)$ and $f$ are defined conditionally to a given path $X$. The sum is still over $\{(\eta_{\mathrm{coal}B}, \eta_{\Rsh b,a}, u_a, v_a, w_{\mathrm{coal}B}, w_{\Rsh b}, w_{B\Rsh B}, -m+1 \le a \le m) \in g^{-1}(\Theta)\}$. We have

$$\begin{aligned}
&\mathrm{P}(\Gamma = \Theta)\\
&\quad = \mathrm{E}[\mathrm{P}^X(\pi(0) = \Theta)]\\
&\quad = \mathrm{E}[\mathrm{P}^X((\tau_{\mathrm{coal}B}, \tau_{\mathrm{coal}b}, \tau_{\Rsh b,a}, \tau_{b\Rsh B,a}, \tau_{B\Rsh B,a},\\
&\qquad\qquad W_{\mathrm{coal}B}, W_{\mathrm{coal}b}, W_{\Rsh b,a}, W_{b\Rsh B,a}, W_{B\Rsh B,a}) \in f^{-1}(\Theta))]\\
&\quad = \sum\Big[\mathrm{P}(\tau_{\mathrm{coal}B} \cap [0; T_\varepsilon] = \eta_{\mathrm{coal}B}, \tau_{\Rsh b,a} \cap [0; T_\varepsilon] = \eta_{\Rsh b,a}, |\tau_{\Rsh b,a} \cap [T_\varepsilon; T]| = u_a,\\
&\qquad\qquad W_{\mathrm{coal}B} = w_{\mathrm{coal}B}, W_{\Rsh b,a} = w_{\Rsh b,a}, -m+1 \le a \le m)\\
&\qquad\quad \times \prod_{a,p} \mathrm{P}\Big(U_{a,p} = \mathbf{1}_{p \in \{w^{(s)}_{B\Rsh B,a}, 1 \le s \le v_a\}}\Big)\Big] + \mathcal{O}\bigg(\frac{1}{(\log\alpha)^2}\bigg),
\end{aligned}$$

by application of Corollaries 1, 3, 4, and 5 and Proposition 5.

### 4.3. Modeling evolution at the selected locus with a Yule tree

4.3.1. *Time rescaling and the Yule tree.* We argued in Section 4.2.1 that it is sufficient to consider events of types coal$B$, $\Rsh b$, and $B \Rsh B$. The events $B \Rsh B$ happen before all the others and give a new partition $\tilde{\pi}'$ at $t = T$. So, the remaining events are those of types coal$B$ (at rate $2/X_t$) and $(\Rsh b, a)$ (at rate $\rho_a(1 - X_t)$).

First, note that, by rescaling time with $\mathrm{d}\tau = (1 - X_t)\,\mathrm{d}t$, the proportion of $B$ in the population is now the solution on $[0, \inf\{\tau > 0 \colon Z_\tau = 1\}]$ of the supercritical branching process:

$$\mathrm{d}Z_\tau = \alpha Z_\tau \coth\bigg(\frac{\alpha}{2} Z_\tau\bigg)\,\mathrm{d}\tau + \sqrt{2 Z_\tau}\,\mathrm{d}W_\tau, \qquad Z_0 = 0,$$

where $W$ is still a standard Brownian motion (see [3, Chapter 6]).

With this time rescaling, the rates of recombination of events $(\Rsh b, a)$ become constant, equal to $\rho_a$. The coalescence rate becomes $2/Z_\tau(1 - Z_\tau)$. However, when the coalescence rates are $2/Z_\tau(1 - Z_\tau)$ and $2/Z_\tau$, the genealogies coincide with probability $1 - \mathcal{O}(\log\alpha/\alpha)$ (see [2, Proposition 4.2]). Moreover, the pair coalescence rate of the individuals with an infinite line of descent is $2/Z_\tau$ (see [2, Lemma 4.5]), so lines taken at random from the supercritical branching process coalesce as infinite lines of descent with probability $1 - \mathcal{O}(\log\alpha/\alpha)$.

https://doi.org/10.1239/aap/1253281062 Published online by Cambridge University Press

O'Connell [12] showed that the genealogy of individuals with an infinite line of descent is a Yule tree with birth rate $\alpha$, and that the number of lineages, $D$, at the end of the selective sweep is Poisson with parameter $\alpha$ (from [4] and the Markov property). Consequently, we can extract a subtree from a Yule tree $\mathcal{Y}$ with birth rate $\alpha$ stopped at Poisson$(\alpha)$ leaves to simulate the genealogy of the sample at the selected locus. Of course, in order for this procedure to make sense, the number of leaves of the stopped Yule tree must be bigger than $|\tilde{\pi}'|$. Let us explain why this is the case.

In fact, with probability $1-\mathcal{O}(1/(\log\alpha)^2)$, the number of lineages, $D$, of the Yule tree $\mathcal{Y}$ at the end of the selective sweep is $\lfloor\alpha\rfloor$. Indeed, by the Chebyshev inequality,

$$\mathrm{P}(|D-\alpha|>\alpha^{3/4})\leq\frac{\operatorname{var}(D)}{(\alpha^{3/4})^2}=\frac{1}{\sqrt{\alpha}}\leq\mathcal{O}\biggl(\frac{1}{(\log\alpha)^2}\biggr),$$

so we can make the approximation $\lceil\alpha-\alpha^{3/4}\rceil\leq D\leq\lfloor\alpha+\alpha^{3/4}\rfloor$. Moreover, as will be explained in Section 4.3.2, the probability that a recombination event occurs (for example, at locus $a$) between the instants when the Yule tree has $\lfloor\alpha-\alpha^{3/4}\rfloor$ and $\lceil\alpha+\alpha^{3/4}\rceil$ lines is bounded by

$$\begin{aligned}
1-\exp\biggl(-\frac{\gamma_a}{\log\alpha}\sum_{\ell=\lceil\alpha-\alpha^{3/4}\rceil+1}^{\lfloor\alpha+\alpha^{3/4}\rfloor}\frac{1}{\ell}\biggr)&+\mathcal{O}\biggl(\frac{1}{(\log\alpha)^2}\biggr)\\
&\leq\frac{\gamma_a}{\log\alpha}\sum_{\ell=\lceil\alpha-\alpha^{3/4}\rceil+1}^{\lfloor\alpha+\alpha^{3/4}\rfloor}\frac{1}{\ell}+\mathcal{O}\biggl(\frac{1}{(\log\alpha)^2}\biggr)\\
&\leq\frac{\gamma_a}{\log\alpha}\frac{2\alpha^{3/4}}{\alpha-\alpha^{3/4}}+\mathcal{O}\biggl(\frac{1}{(\log\alpha)^2}\biggr)\\
&=\mathcal{O}\biggl(\frac{1}{(\log\alpha)^2}\biggr).
\end{aligned}$$

The probability that such a recombination event happens is negligible. Consequently, since $\lceil\alpha-\alpha^{3/4}\rceil\leq\lfloor\alpha\rfloor\leq\lfloor\alpha+\alpha^{3/4}\rfloor$ for large $\alpha$, we can approximate $D$ by $\lfloor\alpha\rfloor$ and use a Yule tree with $\lfloor\alpha\rfloor$ leaves.

By hypothesis, $1\ll\alpha$, so $|\tilde{\pi}'|<\lfloor\alpha\rfloor$ and the above sampling procedure is possible.

Now, randomly choose $|\tilde{\pi}'|$ of the $\lfloor\alpha\rfloor$ leaves of the tree. The corresponding subtree, denoted by $\mathcal{Y}_{|\tilde{\pi}'|}$, is the genealogy at locus 0 for the sample after events $B\Rsh B$. Hence, the Yule subtree can be used to replace the draw of successive events coal$B$ with a probability error of order $\mathcal{O}(1/(\log\alpha)^2)$.

4.3.2. *Use of labels for recombination events.* The leaves of the Yule tree are the blocks of $\tilde{\pi}'$ (all these blocks are marked). Since we are interested in the genealogy of the neutral genes, we have to place events of type $\Rsh b$ along this tree.

Recall that events $\Rsh b$ correspond to recombinations $B\Rsh b$ and $b\Rsh b$. These recombinations split the blocks. Each time a nonmarked block is produced, it escapes the tree (because the tree describes the descendants of the individual where $B$ appeared at $t=0$) and then splits in background $b$ according to recombinations $b\Rsh b$ until $t=0$. Consequently, at each node of the tree, only one marked block is present.

To place events $(\Rsh b,a)$, we use Poisson processes with rate $\rho_a$ on the branches of $\mathcal{Y}_{|\tilde{\pi}'|}$ (a branch denotes an edge of the tree connecting two consecutive nodes). The branches where

https://doi.org/10.1239/aap/1253281062 Published online by Cambridge University Press

jumps happen are labeled. Each label accounts for all the events on the corresponding branch, so that there is at most one label per branch.

A label is written as $\boldsymbol{r} = (r_1, \dots, r_q)$, $1 \leq q \leq 2m$, where $-m < r_1 < r_2 < \cdots < r_q \leq m$ are the locations of the recombination splits. Its meaning is that $q$ recombinations occurred and, for $1 \leq i \leq q$, the $i$th recombination occurred between loci $r_i - 1$ and $r_i$. It must have happened either on that branch, and it is then a $B \Rsh b$ recombination, or after the allele at locus $r_i$ has escaped $\mathcal{Y}_{|\tilde{\pi}'|}$ (owing to a prior $\Rsh b$ recombination), and it is then a $b \Rsh b$ recombination.

We write $r_0 = -m$ and $r_{q+1} = m + 1$. Let $i_0 \in \{0, \dots, q\}$ denote the unique index such that $r_{i_0} \leq 0 < r_{i_0+1}$. Let $\zeta$ be the block present at the node determining the beginning (looking back in time) of the branch. This block is marked.

The above label $\boldsymbol{r}$ implies that, at the time corresponding to the end of the branch, $\zeta$ is replaced by $\zeta \cap (R_{-m}, \dots, R_{r_1-1})$, $\zeta \cap (R_{r_1}, \dots, R_{r_2-1}), \dots, \zeta \cap (R_{r_q}, \dots, R_m)$ in the new partition, and among these new blocks, only $\zeta \cap (R_{r_{i_0}}, \dots, R_{r_{i_0+1}-1})$ is marked and is present at the node determining the end of the branch.

Consider a branch of the extracted tree. At the beginning of this branch, the full Yule tree has $\ell_2$ lineages, and at the end of this branch, it has only $\ell_1 < \ell_2$ lineages (there is at least the coalescence of the studied branch).

When the Yule tree has $\ell$ lineages, the time between two consecutive coalescence events is exponential with rate $\ell\alpha$. Consider a Poisson process with rate $\rho$ on a lineage simulating instants of recombination. Coalescence and recombination events on this lineage are independent. Thus, the probability that there is no recombination on the chunk of the branch corresponding to having $\ell$ lineages on the full Yule tree is the probability that a coalescence event occurs before a recombination event, that is $\ell\alpha/(\ell\alpha + \rho)$, and the probability of having no label on the whole branch is therefore $\prod_{\ell=\ell_1+1}^{\ell_2} \ell\alpha/(\ell\alpha + \rho)$.

**Proposition 6.** *With*

$$q_x^y(\gamma_c) = \prod_{\ell=x+1}^{y} \frac{\ell\alpha}{\ell\alpha + \rho_c},$$

*the probability of the label* $\boldsymbol{r} = (r_1, \dots, r_q)$ *is*

$$\Bigg[\prod_{c=r_{i_0}+1}^{r_{i_0+1}-1} q_{\ell_1}^{\ell_2}(\gamma_c)\Bigg](1 - q_{\ell_1}^{\ell_2}(\gamma_{r_{i_0}}))(1 - q_{\ell_1}^{\ell_2}(\gamma_{r_{i_0+1}}))$$
$$\times \prod_{i \neq i_0,\, i_0+1} (1 - q_0^{\ell_2}(\gamma_{r_i}))\Bigg[\prod_{(c>r_{i_0+1} \text{ or } c<r_{i_0}),\, c \notin \boldsymbol{r}} q_0^{\ell_2}(\gamma_c)\Bigg].$$

*Proof.* The label $\boldsymbol{r} = (r_1, \dots, r_q)$ is the intersection of the following events.

$e_1$: No recombination between loci $r_{i_0}$ and $r_{i_0+1} - 1$ along the branch (that is, when the full Yule tree goes from $\ell_2$ to $\ell_1$ lineages).

$e_2$: At least one recombination between loci $r_{i_0} - 1$ and $r_{i_0}$ along the branch.

$e_3$: At least one recombination between loci $r_{i_0+1} - 1$ and $r_{i_0+1}$ along the branch.

$e_4$: Recombinations at loci $r_i$, $i \notin \{i_0, i_0 + 1\}$.

$e_5$: No recombination elsewhere (even after the lineage has escaped $\mathcal{Y}_{|\tilde{\pi}'|}$).

https://doi.org/10.1239/aap/1253281062 Published online by Cambridge University Press

We have

$$\mathrm{P}(e_1 \cap e_2 \cap e_3 \cap e_4 \cap e_5) = \mathrm{P}(e_1)\,\mathrm{P}(e_2 \mid e_1)\,\mathrm{P}(e_3 \mid e_1, e_2)\,\mathrm{P}(e_4 \mid e_1, e_2, e_3)\,\mathrm{P}(e_5 \mid e_1, e_2, e_3, e_4),$$

which explains the five terms of the product.

It is important to observe that in the last two terms, we have $1 - q_0^{\ell_2}$ and not $1 - q_{\ell_1}^{\ell_2}$ because involved recombination events can happen either before the recombinations, between loci $r_{i_0} - 1$ and $r_{i_0}$ and between loci $r_{i_0+1} - 1$ and $r_{i_0+1}$ (still looking in reversed time)—which explains the $\ell_2$ superscript—or later, but the lineage has then escaped $\mathcal{Y}_{|\tilde{\pi}'|}$ and the event can happen until $t = 0$.

**Proposition 7.** *With an error of $\mathcal{O}(1/(\log\alpha)^2)$, the probability of the label $\boldsymbol{r} = (r_1, \dots, r_q)$ is given by (2).*

*Proof.* By Taylor's expansion,

$$\begin{aligned}\prod_{\ell=\ell_1+1}^{\ell_2} \frac{\ell\alpha}{\ell\alpha + \rho_c} &= \exp\biggl(\sum_{\ell=\ell_1+1}^{\ell_2} \log\biggl(1 - \frac{\rho_c}{\ell\alpha + \rho_c}\biggr)\biggr) \\ &= \exp\biggl(-\frac{\gamma_c}{\log\alpha}\sum_{\ell=\ell_1+1}^{\ell_2}\frac{1}{\ell}\biggr)\exp\biggl(\mathcal{O}\biggl(\frac{1}{(\log\alpha)^2}\biggr)\biggr) \\ &= p_{\ell_1}^{\ell_2}(\boldsymbol{r}, c) + \mathcal{O}\biggl(\frac{1}{(\log\alpha)^2}\biggr).\end{aligned}$$

It is possible to construct the successive partitions after each event of coalescence or recombination, until $t = 0$. Nevertheless, we can also define an equivalence relation that will allow us to decide (directly) if, given the labels on the subtree, two alleles are in the same block of $\pi(0)$. This equivalence relation and the attainment of the marked partition are presented in Section 3.1.

4.3.3. *The final approximation of the process.* We want to establish a final model where we no longer need to simulate a path $X$ in order to construct the ancestral genealogy. For the moment, the parameters of the Bernoulli variables still depend on $X$.

**Proposition 8.** *With an error of $\mathcal{O}(1/(\log\alpha)^2)$, $U_{a,p}$ follows a Bernoulli distribution with parameter*

$$1 - p_0^{\lfloor\alpha\rfloor}(\boldsymbol{r}, a) = 1 - \exp\biggl(-\frac{\gamma_a}{\log\alpha}\sum_{\ell=1}^{\lfloor\alpha\rfloor}\frac{1}{\ell}\biggr).$$

**Remark 5.** Note that $p_0^{\lfloor\alpha\rfloor}(\boldsymbol{r}, a)$ depends only on $a$ and $\alpha$.

*Proof of Proposition 8.* By the time reversibility of $(X_t)_{0\le t\le T}$, the parameter of $U_{a,p}$ is

$$1 - \exp\biggl(-\mathrm{E}\biggl[\rho_a\int_0^T X_s\,\mathrm{d}s\biggr]\biggr) = 1 - \exp\biggl(-\mathrm{E}\biggl[\rho_a\int_0^T (1 - X_s)\,\mathrm{d}s\biggr]\biggr).$$

Let $(N_t)_{t\le 0}$ be a Poisson process with intensity $\mathrm{E}[\rho_a(1 - X_t)\,\mathbf{1}_{\{t\le T\}}]$ for $t \ge 0$. Then

$$\exp\biggl(-\mathrm{E}\biggl[\rho_a\int_0^T (1 - X_s)\,\mathrm{d}s\biggr]\biggr) = \mathrm{P}(N_T = 0).$$

https://doi.org/10.1239/aap/1253281062 Published online by Cambridge University Press

With $\mathrm{d}\tau = (1 - X_t)\,\mathrm{d}t$, $\mathrm{P}(N_T = 0) = \mathrm{P}(N'_{\lfloor\alpha\rfloor} = 0)$, where $N'$ is a Poisson process with intensity $\rho_a$ along a whole line of the Yule tree, that is, while the tree goes from $\lfloor\alpha\rfloor$ to 0 leaves. The probability of a whole line without any label is the product of the probabilities for each branch constituting the line. Therefore, $\mathrm{P}(N'_{\lfloor\alpha\rfloor} = 0) = p_0^{\lfloor\alpha\rfloor}(\boldsymbol{r}, a) + \mathcal{O}(1/(\log\alpha)^2)$.

Finally, we obtain the approximate model presented in Section 3.1.

**Theorem 4.** *We have* $d_{\mathrm{TV}}(\Gamma_1, \tilde{\Gamma}) = \mathcal{O}(1/(\log\alpha)^2)$.

*Proof.* Let $\Theta \in \mathcal{P}$.

$$\begin{aligned}\mathrm{P}(\Gamma_1 = \Theta) - \mathrm{P}(\tilde{\Gamma} = \Theta) = {}& \sum_{\mu} (\mathrm{P}(\Gamma_1 = \Theta \mid \pi' = \mu) - \mathrm{P}(\tilde{\Gamma} = \Theta \mid \tilde{\pi}' = \mu))\,\mathrm{P}(\pi' = \mu) \\ & + \sum_{\mu} \mathrm{P}(\tilde{\Gamma} = \Theta \mid \tilde{\pi}' = \mu)(\mathrm{P}(\pi' = \mu) - \mathrm{P}(\tilde{\pi}' = \mu)),\end{aligned}$$

where we sum over all the partitions $\mu \in \mathcal{P}$. Their number is finite, so it is enough to estimate each term of the sum. We have $\mathrm{P}(\pi' = \mu) - \mathrm{P}(\tilde{\pi}' = \mu) = \mathcal{O}(1/(\log\alpha)^2)$ and $\mathrm{P}(\Gamma_1 = \Theta \mid \pi' = \mu) - \mathrm{P}(\tilde{\Gamma} = \Theta \mid \tilde{\pi}' = \mu) = \mathcal{O}(1/(\log\alpha)^2)$. These two estimates follow from a combination of the following facts.

- Proposition 3.6 of [2].
- The equivalence relation with the labels gives the same partition as the events of type $\Rsh b$ in $\Gamma_1$.
- There is a finite number of labels (independent of $\alpha$) and the probability of each event differs by $\mathcal{O}(1/(\log\alpha)^2)$ according to Proposition 7.
- Proposition 8.

Finally, combining Theorems 3 and 4 yields Theorem 1.

## 5. Approximate distribution of the size of the hitchhiking set

In this section we assume that the recombination rate is the same for all the loci: $\gamma_a = \gamma$ for all $-m + 1 \le a \le m$.

We are interested in the joint distribution of the sizes of the hitchhiking sets of $n$ individuals taken at the end of the selective sweep ($n \ge 1$). Let $H_p$ be the size of the hitchhiking set of individual $p$.

In this Yule-approximation, the evolution to the left of the selected locus is independent of the evolution to the right because events happening to the right do not influence the genealogy to the left, and vice versa. We note that this point does not hold for the 'exact' model, where, for example, recombined lines from both sides of the selected locus may coalesce in the wild-type background, which would make their genealogy dependent.

If we denote by $H_p^{\ell}$ and $H_p^{r}$ the sizes of the hitchhiking set restricted to $\{-m, \dots, 0\}$ and $\{0, \dots, m\}$, respectively, then $H_p^{r}$ and $H_p^{\ell}$ are independent and identically distributed, so that

$$\begin{aligned}&\mathrm{P}(H_1 = h_1, \dots, H_n = h_n) \\ &\quad = \sum_{k_1=1}^{h_1} \cdots \sum_{k_n=1}^{h_n} \mathrm{P}(H_1^{\ell} = k_1, \dots, H_n^{\ell} = k_n)\,\mathrm{P}(H_1^{r} = h_1 - k_1 + 1, \dots, H_n^{r} = h_n - k_n + 1).\end{aligned}$$

https://doi.org/10.1239/aap/1253281062 Published online by Cambridge University Press

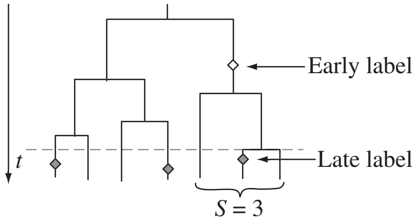


Figure 7: An example of early and late labels. Late labels (*filled diamonds*) occur before the first coalescence event (*dotted line*), and early labels (*open diamonds*) occur after the first coalescence event. Here $S$ is the number of leaves of the subtree under the early label. It is the number of individuals in the sample that is impacted by this early label.

We consider the evolution to the right of locus 0. In particular, the labels are $\boldsymbol{r} = (r_1, \dots, r_q)$ with $0 < r_1 < \cdots < r_q \le m$. For simplicity, we write

$$p_{\ell_1} \equiv p_{\ell_1}^{\lfloor\alpha\rfloor}(\boldsymbol{r}, c) = \exp\biggl(-\frac{\gamma}{\log(\alpha)} \sum_{\ell=\ell_1+1}^{\lfloor\alpha\rfloor} \frac{1}{\ell}\biggr),$$

since it does not depend on $\boldsymbol{r}$ and $c$ anymore. We wish to determine an approximate formula for $\mathrm{P}(H_1^r \ge h_1, \dots, H_n^r \ge h_n)$.

## 5.1. Concepts established in [2]

In this subsection we recall the definitions and main results obtained by Etheridge *et al.* [2] that will be necessary in the next subsection. Note that these results were established for a single neutral gene, so we generalize them.

**Definition 2.** A *late label* is a label attached to recombinations occurring between the end of the selective sweep and the first coalescence event in the sample.

An *early label* is a label attached to recombinations occurring between the first coalescence event in the sample and the beginning of the selective sweep.

**Remark 6.** The term ‘first coalescence event’ corresponds to reversed time, as for transition events in most of this paper. On the other hand, the terms ‘early’ and ‘late’ introduced in [2] should be understood in nonreversed time. However, to use the same terminology as [2], we keep these two adjectives (see Figure 7).

**Theorem 5.** *We consider m neutral genes located to the right of the site under selection.*

1. *With probability* $1 - \mathcal{O}(1/(\log\alpha)^2)$, *there is at most one early label.*
2. *Considering the subtree with* $U = |\pi'|$ *leaves associated to the sample, let S be the number of leaves under the early label (*$S = 0$ *if no early label). Up to a total variation distance of* $\mathcal{O}(1/(\log\alpha)^2)$, *given that* $U = u$, *the distribution of S is given by (3).*
3. *Let F be the number of living lineages in the full Yule tree* $\mathcal{Y}$ *at the time of the first coalescence event in the sample. Up to a total variation distance of* $\mathcal{O}(1/(\log\alpha)^2)$, *given* $U = u$, *the cumulative distribution of F is given by (4).*

https://doi.org/10.1239/aap/1253281062 Published online by Cambridge University Press

4. *Up to a total variation distance of $\mathcal{O}(1/(\log\alpha)^2)$, given that $F = f$, the probability that a late recombination occurs between $R_0$ and $R_a$ on a fixed branch of $\mathcal{Y}_{|\pi'|}$ is*

$$1 - (p_{f-1})^a = 1 - \exp\biggl(-\frac{a\gamma}{\log(\alpha)}\sum_{\ell=f}^{\lfloor\alpha\rfloor}\frac{1}{\ell}\biggr).$$

An illustration is given in Figure 7.

### 5.2. Approximate distribution of the size of the hitchhiking set

Let $h_p \in \{1, \dots, m+1\}$, $1 \le p \le n$.

We wish to obtain an approximate formula for $\mathrm{P}(H_1^r \ge h_1, \dots, H_n^r \ge h_n)$. We always have $H_p^r \ge 1$ because of the site under selection. Looking at individual $p$, the history at loci between $h_p - 1$ and $m$ have no interest for this computation. Consequently, we focus on alleles $0, 1, \dots, h_p - 1$ for individual $p$, $1 \le p \le n$, and set $m = \max(h_1, \dots, h_n) - 1$.

To simplify the notation, we write $U_{\cdot,p} = 1 + \sum_{a=1}^{h_p-1} U_{a,p}$ for all $1 \le p \le n$ and then $U = \sum_{p=1}^{n} U_{\cdot,p} = |\pi'|$. Consider individual $p$ $(1 \le p \le n)$. Given that $U_{\cdot,p} = u_p$, let $Y_{2,p}, \dots, Y_{u_p,p}$ be the ordered locations of the $(u_p - 1)$ recombinations of type $B \Rsh B$. We set $y_{1,p} = 0$ and $y_{u_p+1,p} = h_p$.

After all the $B \Rsh B$ recombinations, the genetic material of individual $p$ is carried by $U_p$ ancestors, with the $i$th ancestor, called $A_{i,p}$, carrying the alleles from locus $y_{i,p}$ to locus $y_{i+1,p} - 1$. Using Proposition 8, we have, for all $u_p \ge 1$ and $1 \le y_{2,p} < \cdots < y_{u_p,p} \le h_p - 1$,

$$\begin{aligned}&\mathrm{P}(U_{\cdot,p} = u_p, Y_{2,p} = y_{2,p}, \dots, Y_{u_p,p} = y_{u_p,p})\\ &\quad= (1-p_0)^{u_p-1} p_0^{(h_p-1)-(u_p-1)} + \mathcal{O}\biggl(\frac{1}{(\log\alpha)^2}\biggr).\end{aligned}$$

Moreover, $B \Rsh B$ recombinations independently affect the $n$ individuals of the sample, so

$$\begin{aligned}&\mathrm{P}(U_{\cdot,p} = u_p,\ Y_{a,p} = y_{a,p},\ 2 \le a \le u_p,\ 1 \le p \le n)\\ &\quad= \prod_{p=1}^{n} \mathrm{P}(U_{\cdot,p} = u_p,\ Y_{a,p} = y_{a,p},\ 2 \le a \le u_p).\end{aligned}$$

Let $\mathcal{L}_p(h_p)$ be the event that none of the neutral alleles at loci $1, \dots, h_p - 1$ for individual $p$ escaped the selective sweep because of a late label. Let $\mathcal{E}_p(h_p)$ be the event that none of the neutral alleles at loci $1, \dots, h_p - 1$ for individual $p$ escaped the selective sweep because of an early label. Then

$$\begin{aligned}&\mathrm{P}(H_1^r \ge h_1, \dots, H_n^r \ge h_n)\\ &\quad= \mathrm{E}[\mathrm{P}(H_1^r \ge h_1, \dots, H_n^r \ge h_n \mid U_{\cdot,p}, Y_{1,p}, \dots, Y_{U_{\cdot,p},p},\ 1 \le p \le n)]\\ &\quad= \mathrm{E}\biggl[\mathrm{P}\biggl(\bigcap_{p=1}^{n} \mathcal{E}_p(h_p) \cap \mathcal{L}_p(h_p) \biggm| U_{\cdot,p}, Y_{1,p}, \dots, Y_{U_{\cdot,p},p},\ 1 \le p \le n\biggr)\biggr]\\ &\quad= \mathrm{E}\biggl[\mathrm{P}\biggl(\bigcap_{p=1}^{n} \mathcal{E}_p(h_p) \biggm| \bigcap_{p=1}^{n} \mathcal{L}_p(h_p), U_{\cdot,p}, Y_{1,p}, \dots, Y_{U_{\cdot,p},p},\ 1 \le p \le n\biggr)\\ &\qquad\quad \times \mathrm{P}\biggl(\bigcap_{p=1}^{n} \mathcal{L}_p(h_p) \biggm| U_{\cdot,p}, Y_{1,p}, \dots, Y_{U_{\cdot,p},p},\ 1 \le p \le n\biggr)\biggr]. \qquad (11)\end{aligned}$$

For $1 \le i \le U_{\cdot,p}$ and $1 \le p \le n$, the alleles between loci $Y_{i,p}$ and $Y_{i+1,p} - 1$ do not escape the selective sweep because of a late label if and only if no late recombination occurs between $R_0$ and $R_{Y_{i+1,p}-1}$ for the ancestor $A_{i,p}$. Since the late recombinations occur independently on the $U$ ancestors, we obtain

$$\mathrm{P}\biggl(\biggl(\bigcap_{p=1}^{n} \mathcal{L}_p(h_p) \biggm| U_{\cdot,p}, Y_{1,p}, \dots, Y_{U_{\cdot,p},p},\ 1 \le p \le n\biggr)\biggr)$$

$$= \mathrm{E}\Bigl[p_{F-1}{}^{\sum_{p=1}^{n} \sum_{a=2}^{U_{\cdot,p}+1} (Y_{a,p}-1)}\Bigr] + \mathcal{O}\biggl(\frac{1}{(\log \alpha)^2}\biggr). \qquad (12)$$

If $\boldsymbol{r} = (r_1, \dots, r_q)$ is the early label (if there is no early label, $\boldsymbol{r} = \varnothing$), define $E(a) = |\boldsymbol{r} \cap \{1, \dots, a\}|$ for all $1 \le a \le m$. Then $\mathrm{P}(E(a) \ge 1)$ is the probability that at least one early recombination took place between $R_0$ and $R_a$. We set $E(0) = 0$. Since, by Theorem 5,

$$\mathrm{P}(S > 0 \mid U = u) = \frac{u\gamma}{\log(\alpha)} \sum_{k=1}^{u-1} \frac{1}{k} + \mathcal{O}\biggl(\frac{1}{(\log \alpha)^2}\biggr)$$

is the probability that there is at least one early recombination at a given locus, we have

$$\mathrm{P}(E(a) \ge 1 \mid U = u) = \frac{ua\gamma}{\log(\alpha)} \sum_{k=1}^{u-1} \frac{1}{k} + \mathcal{O}\biggl(\frac{1}{(\log \alpha)^2}\biggr), \qquad (13)$$

and, in particular, for any $k \in \{1, \dots, m\}$,

$$\mathrm{P}(E(k) \ge 1 \mid E(k-1) = 0,\ U = u) = \frac{u\gamma}{\log(\alpha)} \sum_{\ell=1}^{u-1} \frac{1}{\ell} + \mathcal{O}\biggl(\frac{1}{(\log \alpha)^2}\biggr). \qquad (14)$$

The total probability formula leads to

$$\mathrm{P}\biggl(\bigcap_{p=1}^{n} \mathcal{E}_p(h_p) \biggm| \bigcap_{p=1}^{n} \mathcal{L}_p(h_p), U_{\cdot,p}, Y_{1,p}, \dots, Y_{U_{\cdot,p},p},\ 1 \le p \le n\biggr)$$

$$= \sum_{k=1}^{m} \mathrm{P}\biggl(\bigcap_{p=1}^{n} \mathcal{E}_p(h_p) \biggm| \bigcap_{p=1}^{n} \mathcal{L}_p(h_p), U_{\cdot,p}, Y_{1,p}, \dots, Y_{U_{\cdot,p},p},\ 1 \le p \le n,$$

$$E(k-1) = 0,\ E(k) \ge 1\biggr)$$

$$\times \mathrm{P}\biggl(E(k-1) = 0,\ E(k) \ge 1 \biggm| \bigcap_{p=1}^{n} \mathcal{L}_p(h_p), U_{\cdot,p}, Y_{1,p}, \dots, Y_{U_{\cdot,p},p},\ 1 \le p \le n\biggr)$$

$$+ \mathrm{P}\biggl(\bigcap_{p=1}^{n} \mathcal{E}_p(h_p) \biggm| \bigcap_{p=1}^{n} \mathcal{L}_p(h_p), U_{\cdot,p}, Y_{1,p}, \dots, Y_{U_{\cdot,p},p},\ 1 \le p \le n,\ E(m) = 0\biggr)$$

$$\times \mathrm{P}\biggl(E(m) = 0 \biggm| \bigcap_{p=1}^{n} \mathcal{L}_p(h_p), U_{\cdot,p}, Y_{1,p}, \dots, Y_{U_{\cdot,p},p},\ 1 \le p \le n\biggr)$$

https://doi.org/10.1239/aap/1253281062 Published online by Cambridge University Press

$$
\begin{aligned}
&= \sum_{k=1}^{m} \mathrm{P}(\text{individuals } A_{i,p} \text{ such that } y_{i+1,p}-1 \geq k \text{ are not impacted by the early label}) \\
&\qquad \times \mathrm{P}(E(k) \geq 1 \mid E(k-1)=0, U_{\cdot,p}, Y_{1,p}, \ldots, Y_{U_{\cdot,p},p},\ 1 \leq p \leq n) \\
&\qquad \times \mathrm{P}(E(k-1)=0 \mid U_{\cdot,p}, Y_{1,p}, \ldots, Y_{U_{\cdot,p},p},\ 1 \leq p \leq n) + \mathrm{P}(E(m)=0).
\end{aligned}
$$

We remark that the above individuals are not impacted by the early label if and only if the $S$ picked individuals are among the ancestors $A_{i,p}$ which are such that $Y_{i+1,p}-1<k$. Let $\Theta_k$ be the number of such individuals, and let $u=\sum_{p=1}^{n} u_p$. Let $q(u)=(u\gamma/\log\alpha)\sum_{\ell=1}^{u-1}(1/\ell)$. Using (13), (14), and the convention $\binom{n}{p}=0$ if $n<p$, we obtain

$$
\begin{aligned}
&\mathrm{P}\biggl(\bigcap_{p=1}^{n} \mathcal{E}_p(h_p) \biggm| \bigcap_{p=1}^{n} \mathcal{L}_p(h_p),\ U_{\cdot,p}=u_p,\ Y_{1,p}=y_{1,p}, \ldots, Y_{U_{\cdot,p},p}=y_{u_p,p},\ 1 \leq p \leq n\biggr) \\
&\quad = \sum_{k=1}^{m} \mathrm{E}\biggl[\binom{\Theta_k}{S}\binom{u}{S}^{-1}\biggr] \frac{u\gamma}{\log\alpha}\biggl(\sum_{\ell=1}^{u-1}\frac{1}{\ell}\biggr)\biggl(1-\frac{u(k-1)\gamma}{\log\alpha}\sum_{\ell=1}^{u-1}\frac{1}{\ell}\biggr) \\
&\qquad + 1 - mq(u) + \mathcal{O}\biggl(\frac{1}{(\log\alpha)^2}\biggr) \\
&\quad = q(u)\,\mathrm{E}\biggl[\binom{u}{S}^{-1}\sum_{k=1}^{m}\binom{\Theta_k}{S}\biggr] + 1 - mq(u) + \mathcal{O}\biggl(\frac{1}{(\log\alpha)^2}\biggr). \qquad (15)
\end{aligned}
$$

Theorem 2 follows from (11), (12), and (15).

## 6. Proof of Proposition 4

We are going to use the following notation. To any random variable $\xi$ we associate the random variable $P_\xi$, whose conditional law given that $\xi=\lambda$ is Poisson with parameter $\lambda$. We have

$$
\begin{aligned}
\mathrm{P}(|\tau_{\Rsh b,a} \cap [T_\varepsilon; T]| = u_a) &= \mathrm{E}[\mathrm{P}^X(P_{\psi_a(\delta)} = u_a)], \\
\mathrm{P}(|\tau_{B\Rsh B,a} \cap [T_\varepsilon; T]| = v_a) &= \mathrm{E}[\mathrm{P}^X(P_{\theta_a(\delta)} = v_a)],
\end{aligned}
$$

where $\psi_a(\delta)=n\rho_a\int_{T_\varepsilon}^{T}(1-X_s)\,\mathrm{d}s$, $\theta_a(\delta)=n\rho_a\int_{T_\varepsilon}^{T}X_s\,\mathrm{d}s$, and $\delta=1/(\log\alpha)^2$.

We highlight the dependence on $\delta$ due to the presence of $\rho_a$. It has already been shown in the proof of Proposition 5 that $\operatorname{var}(\theta_a(\delta))=\mathcal{O}((\log\alpha)^{-2})=\mathcal{O}(\delta)$. Similarly, we can check, by the time reversibility of $(X_t)_{0\leq t\leq T}$, that

$$
\operatorname{var}(\psi_a(\delta)) \leq \operatorname{var}\biggl(n\rho_a\int_0^T(1-X_s)\,\mathrm{d}s\biggr) = \operatorname{var}\biggl(n\rho_a\int_0^T X_s\,\mathrm{d}s\biggr) = \mathcal{O}\biggl(\frac{1}{(\log\alpha)^2}\biggr).
$$

If the variance was null, we would have independence of the Poisson processes. In our case, we have in fact a mixture of Poisson processes for which the variance of the parameters is small.

The key observation is that, for all $a$,

$$
\psi_a(\delta) = \frac{\rho_a}{\rho_1}\psi(\delta) = \frac{\gamma_a}{\gamma_1}\psi(\delta) = c_a\psi(\delta),
$$

where $c_a$ is independent of $\delta$ and $\psi=\psi_1$. Similarly, $\theta_a(\delta)=c_a\theta(\delta)$ for all $a$.

https://doi.org/10.1239/aap/1253281062 Published online by Cambridge University Press

Consequently, $\mathrm{P}(|\tau_{\Rsh b,a}\cap[T_\varepsilon, T]| = u_a) = \mathrm{E}[\mathrm{P}^X(P_{c_a\psi(\delta)} = u_a)]$ and $\mathrm{P}(|\tau_{B\Rsh B,a}\cap[T_\varepsilon, T]| = v_a) = \mathrm{E}[\mathrm{P}^X(P_{c_a\theta(\delta)} = v_a)]$. Define $Z_a = \mathrm{P}^X(P_{c_a\psi(\delta)} = u_a)$ and $Y_d = \mathrm{P}^X(P_{c_d\theta(\delta)} = v_d)$ for $-m+1 \le a \le m$ and $-m+1 \le d \le m$. Since the Poisson processes are conditionally independent given $X$, we have to prove that

$$\begin{aligned}\mathrm{E}\Bigg[\prod_{a=-m+1}^{m} Z_a \prod_{d=-m+1}^{m} Y_d\Bigg] - \prod_{a=-m+1}^{m} \mathrm{E}[Z_a] \prod_{d=-m+1}^{m} \mathrm{E}[Y_d]&\\ = \mu(u_{-m+1}, \dots, u_m, v_{-m+1}, \dots, v_m)\mathcal{O}(\delta).&\end{aligned}$$

*First step.* The idea in the following calculations is to use the properties of the variances of $Z_a$ and $Y_d$. To do this, we rewrite the above difference as a sum of differences, adding terms just after removing them. Covariance terms then appear:

$$\begin{aligned}
&\mathrm{E}\Bigg[\prod_{a=-m+1}^{m} Z_a \prod_{d=-m+1}^{m} Y_d\Bigg] - \prod_{a=-m+1}^{m} \mathrm{E}[Z_a] \prod_{d=-m+1}^{m} \mathrm{E}[Y_d]\\
&\quad= \mathrm{E}\Bigg[\prod_{a=-m+1}^{m} Z_a \prod_{d=-m+1}^{m} Y_d\Bigg] - \mathrm{E}\Bigg[\prod_{a=-m+1}^{m} Z_a\Bigg]\mathrm{E}\Bigg[\prod_{d=-m+1}^{m} Y_d\Bigg]\\
&\qquad+ \mathrm{E}\Bigg[\prod_{a=-m+1}^{m} Z_a\Bigg]\mathrm{E}\Bigg[\prod_{d=-m+1}^{m} Y_d\Bigg] - \mathrm{E}\Bigg[\prod_{a=-m+1}^{m} Z_a\Bigg]\mathrm{E}[Y_{-m+1}]\,\mathrm{E}\Bigg[\prod_{d=-m+2}^{m} Y_d\Bigg]\\
&\qquad+ \sum_{K=-m+1}^{m-2}\Bigg(\mathrm{E}\Bigg[\prod_{a=-m+1}^{m} Z_a\Bigg]\prod_{d=-m+1}^{K}\mathrm{E}[Y_d]\,\mathrm{E}\Bigg[\prod_{d=K+1}^{m} Y_d\Bigg]\\
&\qquad\qquad\qquad - \mathrm{E}\Bigg[\prod_{a=-m+1}^{m} Z_a\Bigg]\prod_{d=-m+1}^{K+1}\mathrm{E}[Y_d]\,\mathrm{E}\Bigg[\prod_{d=K+2}^{m} Y_d\Bigg]\Bigg)\\
&\qquad+ \mathrm{E}\Bigg[\prod_{a=-m+1}^{m} Z_a\Bigg]\prod_{d=-m+1}^{m}\mathrm{E}[Y_d] - \mathrm{E}[Z_{-m+1}]\,\mathrm{E}\Bigg[\prod_{a=-m+2}^{m} Z_a\Bigg]\prod_{d=-m+1}^{m}\mathrm{E}[Y_d]\\
&\qquad+ \sum_{K=-m+1}^{m-2}\Bigg(\prod_{a=-m+1}^{K}\mathrm{E}[Z_a]\,\mathrm{E}\Bigg[\prod_{a=K+1}^{m} Z_a\Bigg]\prod_{d=-m+1}^{m}\mathrm{E}[Y_d]\\
&\qquad\qquad\qquad - \prod_{a=-m+1}^{K+1}\mathrm{E}[Z_a]\,\mathrm{E}\Bigg[\prod_{a=K+2}^{m} Z_a\Bigg]\prod_{d=-m+1}^{m}\mathrm{E}[Y_d]\Bigg)\\
&\quad= \operatorname{cov}\Bigg(\prod_{a=-m+1}^{m} Z_a, \prod_{d=-m+1}^{m} Y_d\Bigg) + \mathrm{E}\Bigg[\prod_{a=-m+1}^{m} Z_a\Bigg]\operatorname{cov}\Bigg(Y_{-m+1}, \prod_{d=-m+2}^{m} Y_d\Bigg)\\
&\qquad+ \sum_{K=-m+1}^{m-2}\Bigg(\mathrm{E}\Bigg[\prod_{a=-m+1}^{m} Z_a\Bigg]\prod_{d=-m+1}^{K}\mathrm{E}[Y_d]\operatorname{cov}\Bigg(Y_{K+1}, \prod_{d=K+2}^{m} Y_d\Bigg)\Bigg)\\
&\qquad+ \operatorname{cov}\Bigg(Z_{-m+1}, \prod_{a=-m+2}^{m} Z_a\Bigg)\prod_{d=-m+1}^{m}\mathrm{E}[Y_d]\\
&\qquad+ \sum_{K=-m+1}^{m-2}\Bigg(\operatorname{cov}\Bigg(Z_{K+1}, \prod_{a=K+2}^{m} Z_a\Bigg)\prod_{a=-m+1}^{K}\mathrm{E}[Z_a]\prod_{d=-m+1}^{m}\mathrm{E}[Y_d]\Bigg).
\end{aligned}$$

https://doi.org/10.1239/aap/1253281062 Published online by Cambridge University Press

Since, for all $-m+1 \le a \le m$ and $-m+1 \le d \le m$, we have $0 \le Z_a \le 1$ and $0 \le Y_d \le 1$,

$$\begin{aligned}\bigg|\mathrm{E}\bigg[\prod_{a=-m+1}^{m} Z_a \prod_{d=-m+1}^{m} Y_d\bigg] - \prod_{a=-m+1}^{m} \mathrm{E}[Z_a] \prod_{d=-m+1}^{m} \mathrm{E}[Y_d]\bigg| &\\ \le \sqrt{\operatorname{var}\bigg(\prod_{a=-m+1}^{m} Z_a\bigg) \operatorname{var}\bigg(\prod_{d=-m+1}^{m} Y_d\bigg)} + \sqrt{\operatorname{var}(Y_{-m+1}) \operatorname{var}\bigg(\prod_{d=-m+2}^{m} Y_d\bigg)} &\\ + \sum_{K=-m+1}^{m-2} \sqrt{\operatorname{var}(Y_{K+1}) \operatorname{var}\bigg(\prod_{j=K+2}^{m} Y_j\bigg)} + \sqrt{\operatorname{var}(Z_{-m+1}) \operatorname{var}\bigg(\prod_{i=-m+2}^{m} Z_i\bigg)} &\\ + \sum_{K=-m+1}^{m-2} \sqrt{\operatorname{var}(Z_{K+1}) \operatorname{var}\bigg(\prod_{i=K+2}^{m} Z_i\bigg)}. &\end{aligned}$$

Since, for all $a, b \ge 0$, $\sqrt{ab} \le \max(a, b)$, we have to prove that, for all $-m+1 \le K_0 \le K \le m$, $\operatorname{var}(\prod_{a=K_0}^{K} Z_a) = \mu(u_{K_0}, \dots, u_K)\mathcal{O}(\delta)$ and $\operatorname{var}(\prod_{d=K_0}^{K} Y_d) = \mu(v_{K_0}, \dots, v_K)\mathcal{O}(\delta)$ with $\sum_{u_{K_0},\dots,u_K=0}^{\infty} \mu(u_{K_0}, \dots, u_K) < \infty$ and $\sum_{v_{K_0},\dots,v_K=0}^{\infty} \mu(v_{K_0}, \dots, v_K) < \infty$.

(To simplify, the function is still denoted by $\mu$ even though the number of parameters is different.)

*Second step.* Let $-m+1 \le K_0 \le K \le m$. We are going to show that $\operatorname{var}(\prod_{a=K_0}^{K} Z_a) = \mu(u_{K_0}, \dots, u_K)\mathcal{O}(\delta)$ with $\sum_{u_{K_0},\dots,u_K=0}^{\infty} \mu(u_{K_0}, \dots, u_K) < \infty$. The proof is exactly the same for $\prod_{d=K_0}^{K} Y_d$.

With $Z_a = \exp(-c_a\psi)(c_a\psi)^{u_a}/u_a!$ for all $a$, we obtain

$$\begin{aligned}\operatorname{var}\bigg(\prod_{a=K_0}^{K} Z_a\bigg) &= \mathrm{E}\bigg[\prod_{a=K_0}^{K} Z_a^2\bigg] - \bigg(\mathrm{E}\bigg[\prod_{a=K_0}^{K} Z_a\bigg]\bigg)^2 \\ &= \mathrm{E}\bigg[\frac{\exp(-2(\sum_{a=K_0}^{K} c_a)\psi)\psi^{2\sum_{a=K_0}^{K} u_a} \prod_{a=K_0}^{K} c_a^{2u_a}}{\prod_{a=K_0}^{K} (u_a!)^2}\bigg] \\ &\quad - \bigg(\mathrm{E}\bigg[\frac{\exp(-(\sum_{a=K_0}^{K} c_a)\psi)\psi^{\sum_{a=K_0}^{K} u_a} \prod_{a=K_0}^{K} c_a^{u_a}}{\prod_{a=K_0}^{K} u_a!}\bigg]\bigg)^2 \\ &=: A - B.\end{aligned}$$

*Third step.* In this step we show that

$$B = \frac{\exp(-2(\sum_{a=K_0}^{K} c_a)\,\mathrm{E}[\psi])\,\mathrm{E}(\psi)^{2\sum_{a=K_0}^{K} u_a} \prod_{a=K_0}^{K} c_a^{2u_a}}{\prod_{a=K_0}^{K} (u_a!)^2} + \mu_B(u_{K_0}, \dots, u_K)\mathcal{O}(\delta).$$

Using an order 1 Taylor expansion between the values $\psi$ and $\mathrm{E}[\psi]$ of the function

$$f\colon \psi \mapsto f(\psi) = \frac{\exp(-(\sum_{a=K_0}^{K} c_a)\psi)\psi^{\sum u_a} \prod_{a=K_0}^{K} c_a^{u_a}}{\prod_{a=K_0}^{K} u_a!},$$

and taking the expectation and the square of the result, there exists a $\tilde{\psi}$ such that $|\tilde{\psi} - \mathrm{E}[\psi]| \leq |\psi - \mathrm{E}[\psi]|$ and

$$B = \biggl(f(\mathrm{E}[\psi]) + \mathrm{E}\biggl[(\psi - \mathrm{E}[\psi])\frac{\exp(-(\sum_{a=K_0}^{K} c_a)\tilde{\psi})(\prod_{a=K_0}^{K} c_a^{u_a})}{\prod_{a=K_0}^{K} u_a!} \times \biggl(\biggl(-\sum_{a=K_0}^{K} c_a\biggr)\tilde{\psi}^{\sum_{a=K_0}^{K} u_a} + \biggl(\sum_{a=K_0}^{K} u_a\biggr)\tilde{\psi}^{\sum_{a=K_0}^{K} u_a - 1}\biggr)\biggr]\biggr)^2.$$

Expanding this equation and applying Schwarz's inequality, we obtain

$$\begin{aligned} &|B - f(\mathrm{E}[\psi])^2| \\ &\leq 2f(\mathrm{E}[\psi])\biggl|\mathrm{E}\biggl[(\psi - \mathrm{E}[\psi])\frac{\exp(-(\sum_{a=K_0}^{K} c_a)\tilde{\psi})(\prod_{a=K_0}^{K} c_a^{u_a})}{\prod_{a=K_0}^{K} u_a!} \\ &\qquad \times \biggl(\biggl(-\sum_{a=K_0}^{K} c_a\biggr)\tilde{\psi}^{\sum_{a=K_0}^{K} u_a} + \biggl(\sum_{a=K_0}^{K} u_a\biggr)\tilde{\psi}^{\sum_{a=K_0}^{K} u_a - 1}\biggr)\biggr]\biggr| \\ &\quad + \mathrm{E}[|\psi - \mathrm{E}[\psi]|^2]\,\mathrm{E}\biggl[\biggl(\frac{\exp(-(\sum_{a=K_0}^{K} c_a)\tilde{\psi})(\prod_{a=K_0}^{K} c_a^{u_a})}{\prod_{a=K_0}^{K} u_a!} \\ &\qquad \times \biggl(\biggl(-\sum_{a=K_0}^{K} c_a\biggr)\tilde{\psi}^{\sum_{a=K_0}^{K} u_a} + \biggl(\sum_{a=K_0}^{K} u_a\biggr)\tilde{\psi}^{\sum_{a=K_0}^{K} u_a - 1}\biggr)\biggr)^2\biggr]. \end{aligned} \tag{16}$$

We can rewrite the term whose expectation of the square is the last factor above as

$$\prod_{a=K_0}^{K}\biggl(\frac{\mathrm{e}^{-c_a\tilde{\psi}}(c_a\tilde{\psi})^{u_a}}{u_a!}\biggr)\biggl(-\sum_{a=K_0}^{K} c_a\biggr) + \sum_{\substack{r=K_0 \\ u_r>0}}^{K}\prod_{a\neq r}\biggl(\frac{\mathrm{e}^{-c_a\tilde{\psi}}(c_a\tilde{\psi})^{u_a}}{u_a!}\biggr)\frac{\mathrm{e}^{-c_r\tilde{\psi}}(c_r\tilde{\psi})^{u_r-1}}{(u_r-1)!}c_r.$$

So,

$$\begin{aligned} &\biggl[\frac{\exp(-(\sum_{a=K_0}^{K} c_a)\tilde{\psi})(\prod_{a=K_0}^{K} c_a^{u_a})}{\prod_{a=K_0}^{K} u_a!}\biggl(\biggl(-\sum_{a=K_0}^{K} c_a\biggr)\tilde{\psi}^{\sum_{a=K_0}^{K} u_a} \\ &\qquad + \biggl(\sum_{a=K_0}^{K} u_a\biggr)\tilde{\psi}^{\sum_{a=K_0}^{K} u_a - 1}\biggr)\biggr]^2 \\ &\leq \biggl(\sum_{a=K_0}^{K} c_a\biggr)^2\biggl(\prod_{a=K_0}^{K}\beta_a(u_a) + \sum_{\substack{r=K_0 \\ u_r>0}}^{K}\biggl(\prod_{a\neq r}\beta_a(u_a)\biggr)\beta_r(u_r-1)\biggr)^2 \\ &= \mu_{B,1}(u_{K_0}, \dots, u_K), \end{aligned}$$

where $\beta_a(u) = \mathrm{e}^{-c_a\tilde{\psi}}(c_a\tilde{\psi})^u/u!$ and $\sum_{u_{K_0},\dots,u_K=0}^{\infty}\mu_{B,1}(u_{K_0}, \dots, u_K) < \infty$ owing to

$$\sum_{u=0}^{\infty}\beta_a(u) = 1.$$

https://doi.org/10.1239/aap/1253281062 Published online by Cambridge University Press

For the other term on the right-hand side of (16), we again use a Taylor expansion of order 1, between $\tilde{\psi}$ and $\mathrm{E}[\psi]$ to find $\hat{\psi}$ such that $|\hat{\psi} - \mathrm{E}[\psi]| \le |\tilde{\psi} - \mathrm{E}[\psi]| \le |\psi - \mathrm{E}[\psi]|$ and

$$
\begin{aligned}
&\exp\biggl(-\biggl(\sum_{a=K_0}^{K} c_a\biggr)\tilde{\psi}\biggr)\biggl(\prod_{a=K_0}^{K} c_a^{u_a}\biggr)\frac{((-\sum_{a=K_0}^{K} c_a)\tilde{\psi}^{\sum_{a=K_0}^{K} u_a} + (\sum_{a=K_0}^{K} u_a)\tilde{\psi}^{\sum_{a=K_0}^{K} u_a - 1})}{\prod_{a=K_0}^{K} u_a!}\\
&\quad = \exp\biggl(-\biggl(\sum_{a=K_0}^{K} c_a\biggr)\mathrm{E}[\psi]\biggr)\biggl(\prod_{a=K_0}^{K} c_a^{u_a}\biggr)\\
&\qquad \times \frac{((-\sum_{a=K_0}^{K} c_a)\,\mathrm{E}[(\psi)^{\sum_{a=K_0}^{K} u_a}] + (\sum_{a=K_0}^{K} u_a)\,\mathrm{E}[(\psi)^{\sum_{a=K_0}^{K} u_a - 1}])}{\prod_{a=K_0}^{K} u_a!}\\
&\qquad + (\tilde{\psi} - \mathrm{E}[\psi])\frac{\exp(-\sum_{a=K_0}^{K} c_a\hat{\psi})}{\prod_{a=K_0}^{K} u_a!}\prod_{a=K_0}^{K} c_a^{u_a}\\
&\qquad \times \Biggl(\biggl(\sum_{a=K_0}^{K} c_a\biggr)^2 \hat{\psi}^{\sum_{a=K_0}^{K} u_a} - 2\biggl(\sum_{a=K_0}^{K} c_a\biggr)\biggl(\sum_{a=K_0}^{K} u_a\biggr)\hat{\psi}^{\sum_{a=K_0}^{K} u_a - 1}\\
&\qquad\qquad + \biggl(\sum_{a=K_0}^{K} u_a\biggr)\biggl(\sum_{a=K_0}^{K} u_a - 1\biggr)\hat{\psi}^{\sum_{a=K_0}^{K} u_a - 2}\Biggr)\\
&\quad = B_1 + (\tilde{\psi} - \mathrm{E}[\psi]) \times B_2.
\end{aligned}
$$

The first term, $B_1$, is deterministic, so its contribution to the right-hand side of (16) is zero. Moreover, some easy but technical calculations (not shown), based on the inequality $\exp(-x)x^u/u! \le 1$ for all $x \ge 0$, $u \in \mathbb{N}$, show that $B_2$ is bounded by $(\sum_{a=K_0}^{K} c_a)^2 + 2(\sum_{a=K_0}^{K} c_a)^2 + (\sum_{a=K_0}^{K} c_a)^2 = C$, with $C$ independent of $u_a$. Therefore, using Schwarz's inequality,

$$
\begin{aligned}
&|B - f(\mathrm{E}[\psi])^2|\\
&\quad \le 2f(\mathrm{E}[\psi]) \times C \times \mathrm{E}[|\psi - \mathrm{E}[\psi]||\tilde{\psi} - \mathrm{E}[\psi]|] + \mu_{B,1}(u_{K_0}, \dots, u_K) \times \mathcal{O}(\delta)\\
&\quad \le 2f(\mathrm{E}[\psi]) \times C \times \mathrm{E}[|\psi - \mathrm{E}[\psi]|^2] + \mu_{B,1}(u_{K_0}, \dots, u_K)\mathcal{O}(\delta)\\
&\quad \le \mu_B(u_{K_0}, \dots, u_K)\mathcal{O}(\delta)
\end{aligned}
$$

with $\sum_{u_{K_0},\dots,u_K=0}^{\infty} \mu_B(u_{K_0}, \dots, u_K) < \infty$, because

$$
f(\mathrm{E}[\psi]) = \prod_{a=K_0}^{K} \bar{\beta}_a(u_a),
$$

where $\bar{\beta}_a(u_a) = \exp(-c_i\,\mathrm{E}[\psi])(c_a\,\mathrm{E}[\psi])^{u_a}/u_a!$.

*Fourth step.* In this step we show that

$$
A = \frac{\exp(-2(\sum_{a=K_0}^{K} c_a)\,\mathrm{E}[\psi])\,\mathrm{E}(\psi)^{2\sum_{a=K_0}^{K} u_a}\prod_{a=K_0}^{K} c_a^{2u_a}}{\prod_{a=K_0}^{K}(u_a!)^2} + \mu_A(u_{K_0}, \dots, u_K) \times \mathcal{O}(\delta),
$$

where $\sum_{u_{K_0},\dots,u_K=0}^{\infty} \mu_A(u_{K_0}, \dots, u_K) < \infty$.

As above, we use a Taylor expansion, but now at order 2, and we compute the expectation: there exists a $\tilde{\psi}$ such that $|\tilde{\psi} - \mathrm{E}[\psi]| \le |\psi - \mathrm{E}[\psi]|$ and

$$A = \frac{\exp(-2(\sum_{a=K_0}^{K} c_a)\,\mathrm{E}[\psi])\,\mathrm{E}(\psi)^{2\sum_{a=K_0}^{K} u_a} \prod_{a=K_0}^{K} c_a^{2u_a}}{\prod_{a=K_0}^{K}(u_a!)^2} + \frac{1}{2}\,\mathrm{E}\biggl[|\psi - \mathrm{E}[\psi]|^2 \frac{\partial^2}{\partial\psi^2}\biggl(\frac{\exp(-2(\sum_{a=K_0}^{K} c_a)\psi)\psi^{2\sum_{a=K_0}^{K} u_a} \prod_{a=K_0}^{K} c_a^{2u_a}}{\prod_{a=K_0}^{K}(u_a!)^2}\biggr)\bigg|_{\psi=\tilde{\psi}}\biggr].$$

We expand the last term:

$$\begin{aligned}
&\frac{\partial^2}{\partial\psi^2}\biggl(\frac{\exp(-2(\sum_{a=K_0}^{K} c_a)\psi)\psi^{2\sum_{a=K_0}^{K} u_a} \prod_{a=K_0}^{K} c_a^{2u_a}}{\prod_{a=K_0}^{K}(u_a!)^2}\biggr)\bigg|_{\psi=\tilde{\psi}}\\
&\quad = \exp\biggl(-2\biggl(\sum_{a=K_0}^{K} c_a\biggr)\tilde{\psi}\biggr)\\
&\qquad \times \biggl(\frac{4(\sum_{a=K_0}^{K} c_a)^2\tilde{\psi}^{2\sum_{a=K_0}^{K} u_a} \prod_{a=K_0}^{K} c_a^{2u_a}}{\prod_{a=K_0}^{K}(u_a!)^2}\\
&\qquad\qquad - \frac{8(\sum_{a=K_0}^{K} c_a)(\sum_{a=K_0}^{K} u_a)\tilde{\psi}^{2\sum_{a=K_0}^{K} u_a - 1} \prod_{a=K_0}^{K} c_a^{2u_a}}{\prod_{a=K_0}^{K}(u_a!)^2}\\
&\qquad\qquad + \frac{(2\sum_{a=K_0}^{K} u_a)(2(\sum_{a=K_0}^{K} u_a) - 1)\tilde{\psi}^{2\sum_{a=K_0}^{K} u_a - 2} \prod_{a=K_0}^{K} c_a^{2u_a}}{\prod_{a=K_0}^{K}(u_a!)^2}\biggr)\\
&\quad =: A_1 + A_2 + A_3.
\end{aligned}$$

If, for all $a$, $u_a = 0$, then

$$\frac{\partial^2}{\partial\psi^2}\biggl(\exp\biggl(-2\biggl(\sum_{a=K_0}^{K} c_a\biggr)\psi\biggr)\biggr)\bigg|_{\psi=\tilde{\psi}} = 4\biggl(\sum_{a=K_0}^{K} c_a\biggr)^2 \exp\biggl(-2\biggl(\sum_{a=K_0}^{K} c_a\biggr)\tilde{\psi}\biggr).$$

Again, after easy but technical calculations, we find that if there exists an $a_0$ such that $u_{a_0} \ge 1$ then, writing $\beta_a(u) = \exp(-c_a\tilde{\psi})(c_a\tilde{\psi})^u/u!$,

- $A_1$ is bounded by $4(\sum_{a=K_0}^{K} c_a)^2 \prod_{a=K_0}^{K} \beta_a(u_a)$,
- $A_2$ is bounded by $8(\sum_{a=K_0}^{K} c_a)^2 \prod_{a=K_0}^{K} \beta_a(u_a)$,
- $A_3$ is bounded by

$$4\biggl(\sum_{a=K_0}^{K} c_a\biggr)\biggl[\sum_{\substack{b=K_0\\u_b>0}}^{K}\biggl(\prod_{a\ne b}\beta_a(u_a)\biggr)\beta_b(u_b-1)c_b\biggr] + 2\sum_{\substack{r=K_0\\u_r>0}}^{K}\biggl(\prod_{a\ne r}\beta_a(u_a)\biggr)\beta_r(u_r-1)c_r c_{a_0}.$$

Finally,

$$\frac{1}{2}\frac{\partial^2}{\partial\psi^2}\biggl(\frac{\exp(-2(\sum_{a=K_0}^{K} c_a)\psi)\psi^{2\sum_{a=K_0}^{K} u_a} \prod_{a=K_0}^{K} c_a^{2u_a}}{\prod_{a=K_0}^{K}(u_a!)^2}\biggr)\bigg|_{\psi=\tilde{\psi}} \le \mu_A(u_{K_0}, \dots, u_K)$$

https://doi.org/10.1239/aap/1253281062 Published online by Cambridge University Press

with $\sum_{u_{K_0},\dots,u_K=0}^{\infty} \mu_A(u_{K_0},\dots,u_K) < \infty$ and

$$\begin{aligned}\left| A - \frac{\exp(-2(\sum_{a=K_0}^{K} c_a)\,\mathrm{E}[\psi])\,\mathrm{E}(\psi)^{2\sum_{a=K_0}^{K} u_a} \prod_{a=K_0}^{K} c_a^{2u_a}}{\prod_{a=K_0}^{K} (u_a!)^2} \right| & \\ \le \mu_A(u_{K_0},\dots,u_K)\,\mathrm{E}[|\psi - \mathrm{E}[\psi]|^2] & \\ = \mu_A(u_{K_0},\dots,u_K) \times \mathcal{O}(\delta). &\end{aligned}$$

Combining the third and fourth steps gives the result announced in the second step, which completes the proof.

## Acknowledgements

I thank my thesis advisor, Etienne Pardoux, and Vlada Limic for many discussions on this work.

## References

[1] Depaulis, F. and Veuille, M. (1998). Neutrality tests based on the number of haplotypes under an infinite site model. *Molec. Biol. Evol.* **15,** 1788–1790.

[2] Etheridge, A., Pfaffelhuber, P. and Wakolbinger, A. (2006). An approximate sampling formula under genetic hitchhiking. *Ann. Appl. Prob.* **16,** 685–729.

[3] Ethier, S. N. and Kurtz, T. G. (1986). *Markov Processes*. John Wiley, New York.

[4] Evans, S. N. and O'Connell, N. (1994). Weighted occupation time for branching particle systems and a representation for the supercritical superprocess. *Canad. Math. Bull.* **37,** 187–196.

[5] Griffiths, R. C. (2003). The frequency spectrum of a mutation and its age, in a general diffusion model. *Theoret. Pop. Biol.* **64,** 241–251.

[6] Hanchard, N. *et al.* (2006). Screening for recently selected alleles by analysis of human haplotype similarity. *Amer. J. Human Genet.* **78,** 153–159.

[7] Hudson, R. *et al.* (1994). Evidence for positive selection in the superoxide dismutase *(Sod)* region of *Drosophila melanogaster*. *Genetics* **136,** 1329–1340.

[8] Kaj, I. and Krone, S. M. (2003). The coalescent process in a population with stochastically varying size. *J. Appl. Prob.* **40,** 33–48.

[9] Lehnert, A., Pfaffelhuber, P. and Stephan, W. (2008). Linkage disequilibrium under genetic hitchhiking in finite populations. *Genetics* **179,** 527–537.

[10] Maynard-Smith, J. and Haigh, J. (1974). The hitch-hiking effect of a favorable gene. *Genet. Res.* **23,** 23–35.

[11] McVean, G. (2007). The structure of linkage disequilibrium around a selective sweep. *Genetics* **175,** 1395–1406.

[12] O'Connell, N. (1993). Yule process approximation for the skeleton of a branching process. *J. Appl. Prob.* **30,** 725–729.

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

[14] Pfaffelhuber, P., Haubold, B. and Wakolbinger, A. (2006). Approximate genealogies under genetic hitchhiking. *Genetics* **174,** 1995–2008.

[15] Sabeti, P. C. *et al.* (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419,** 832–837.

[16] Schweinsberg, J. and Durrett, R. (2005). Random partitions approximating the coalescence of lineages during a selective sweep. *Ann. Appl. Prob.* **15,** 1591–1651.

[17] Stephan, W., Song, Y. S. and Langley, C. H. (2006). The hitchhiking effect on linkage disequilibrium between linked loci. *Genetics* **172,** 2647–2663.

[18] Wang, E., Komada, G., Baldi, P. and Moyzis, R. (2006). Global landscape of recent inferred Darwinian selection for *Homo sapiens*. *Proc. Nat. Acad. Sci. USA* **103,** 135–140.

https://doi.org/10.1239/aap/1253281062 Published online by Cambridge University Press