

Approximating genealogies for partially linked neutral loci under a selective sweep

P. Pfaffelhuber · A. Studeny

Received: 31 October 2006 / Published online: 30 March 2007
© Springer-Verlag 2007

Abstract Consider a genetic locus carrying a strongly beneficial allele which has recently fixed in a large population. As strongly beneficial alleles fix quickly, sequence diversity at partially linked neutral loci is reduced. This phenomenon is known as a selective sweep. The fixation of the beneficial allele not only affects sequence diversity at single neutral loci but also the joint allele distribution of several partially linked neutral loci. This distribution can be studied using the ancestral recombination graph for samples of partially linked neutral loci during the selective sweep. To approximate this graph, we extend recent work by Etheridge et al. (Ann Appl Probab 16:685–729, 2006) and Schweinsberg and Durrett (Ann Appl Probab 15:1591–1651, 2005) using a marked Yule tree for the genealogy at a single neutral locus linked to a strongly beneficial one. We focus on joint genealogies at two partially linked neutral loci in the case of large selection coefficients and recombination rates $= \mathcal{O}(\frac{1}{\log s})$ between loci. Our approach leads to a full description of the genealogy with accuracy of $\mathcal{O}((\log s)^{-2})$ in probability. As an application, we derive the expectation of Lewontin's D as a measure for non-random association of alleles.

Keywords Selective sweep Genetic hitchhiking Diffusion approximation Yule process Ancestral recombination graph Random background

Mathematics Subject Classification (2000) 92D15 (Primary) · 60J80 · 60J85 · 60K37 · 92D10 (Secondary)

P. Pfaffelhuber (✉) · A. Studeny
Ludwig-Maximilian University Munich, Munich, Germany
e-mail: p.p@lmu.de

1 Introduction

The model of selective sweeps, also known as genetic hitchhiking introduced by Maynard-Smith and Haigh in [7], is the starting point for a large body of both empirical and theoretical population genetic studies. It predicts that sequence diversity is reduced close to a strongly selected locus on a recombining genome near the time of fixation of the beneficial allele. Theoretical studies aim at describing these patterns of genetic diversity in detail while empirical work uses this prediction to identify genes under selection.

If a species or a population adapts to its environment, several genes might be under strong selection. Moreover, if the function of genes were known, we would have predictions as to which genes are responsible for the adaptive process. Unfortunately, functional information is scarce. Without functional knowledge and in the presence of recombination, the model of selective sweeps helps to identify candidate genes affected by recent selective pressures. Genome scans are carried out for a sample of individuals, which show patterns of sequence diversity at lots of marker loci in the whole genome [8]. If a marker shows low diversity, statistical tests help to decide if a gene under selection is located nearby [6].

Most theoretical studies of selective sweeps have focused on a model with one selective and one partially linked neutral locus [3, 11, 17, 24, 26]. This simple model already describes the reduction in sequence diversity. However genetic data are frequently available for many partially linked loci. This raises the question of whether selective sweeps also generate distinct patterns of multi-locus allele frequencies. We will follow [25] and study a three locus model with one selective and two partially linked neutral loci. Using this model, it is possible to study the non-random association of allelic types at the two neutral loci, which is usually called linkage disequilibrium.

An influential idea in the analysis of selective sweeps was to study approximate genealogies describing relationships between the individuals in a sample from the population. Studying genealogies at the selected site started with [4] and was carried further to linked neutral loci in [11].

The genealogy at a single neutral locus can be constructed as structured coalescent. Here, the beneficial and wild-type allele at the selected locus form two subpopulations. Their sizes are determined by the frequency path of the beneficial allele during the selective sweep. Assume a new gamete is built (forward in time) by recombination of a beneficial allele at the selected locus and a neutral variant linked to a wild-type. Following the neutral variant backward in time leads to a migration event from the beneficial to the wild-type background. Therefore, recombination acts as migration between the beneficial and the wild-type backgrounds.

Genealogies of two or more loci can be constructed using the ancestral recombination graph [7, 9]. Therefore, we will construct ancestries of two partially linked neutral loci under a selective sweep by structured ancestral recombination graphs. In the case of only one locus, the two subpopulations are distinguished by the beneficial and wild-type allele at the selected locus, respectively. This ancestral recombination graph will serve as the exact model for genealogies at partially linked loci under a selective sweep. However, an exact analysis is hard to obtain, because the graph must be conditioned on the random frequency path of the beneficial allele.

An alternative approach uses a two-step procedure for genealogies at the selective and the neutral locus. First, the (approximate) genealogy at the selective locus is generated and second, the genealogy at the neutral locus is added, which might differ due to recombination. Two approximate genealogies at the selected site have been proposed. First, a star-like genealogy, which means that the most recent common ancestor of all pairs in the population is the individual which carried the beneficial allele first [18,24]. Second, a Yule process, i.e., a pure birth process, which allows for coalescences also during the selective sweep [24]. It was shown in [24, Theorems 1.1, 1.2] that the Yule process approximation is more exact than the star-like approximation. Therefore, we will use this Yule process approximation for the genealogy at the selected site to study the three locus model [25] for selective sweeps. We will show that the analysis carried out in [8] in the two locus case can be extended to the three locus case (Theorem 1). Moreover, the approximation by a Yule process can be used to calculate characteristics of linkage disequilibrium explicitly (Theorem 2).

2 The model

Consider a beneficial allele which enters a population of (haploid) size N at time $t = 0$ and has a selective advantage s with respect to the wild-type allele. Set $\gamma = sN$, which is called the scaled selection coefficient. As selection can only be detected if the beneficial allele exists in the population, we condition on fixation of the beneficial allele and let T be the (random) time of fixation.

Assume reproduction in the population follows a Wright–Fisher model, or, more generally, a Cannings model with individual offspring variance 1. In the limit of infinite N and a time rescaling in units of generations, the frequency path of the beneficial allele is the solution of the SDE

$$dX = X(1 - X)\coth(\gamma X)dt + \sqrt{X(1 - X)}dW, \tag{2.1}$$

with a standard Brownian motion W and $X_0 = 0$. This diffusion arises as a transform of the process describing the unconditional frequency path with the fixation probability of the beneficial allele as a harmonic function and has 0 as an entrance boundary. (See e.g. [8], p. 245 and [3], (2.1)).

Two neutral loci are partially linked to the selected locus. For simplicity, we refer to the two neutral loci as the left and right neutral locus, denoted by L and R . As illustrated in Fig. 1, the selected locus lies either (i) outside or (ii) in between the neutral loci. All other possible geometries are equivalent to either (i) or (ii) because of the symmetry in the model.

Recombination can break up the association of these three loci. (We only consider recombination as simple crossing over. Gene conversion is not considered in our model.) As we take a limiting infinite population and rescale time by a factor of N , we have to consider scaled recombination rates. These are different for the two geometries. For geometry (i) we denote the recombination rates between the selective and neutral loci by s_L, s_R and for geometry (ii) by s_{LS}, s_{SR} respectively.

The two linked neutral loci do not affect the frequency path of the beneficial allele. In contrast, neutral variants which are linked to the beneficial allele at the beginning

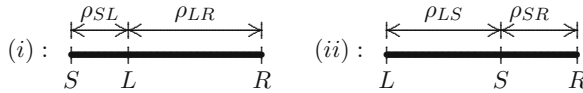


Fig. 1 The two possible geometries of the selected locus (S) and the two neutral loci (L and R). The scaled recombination rates between loci are given by ρ_{SL} , ρ_{LR} , ρ_{LS} and ρ_{SR}

of the selective sweep rise in frequency. Looking backward in time from the time of fixation, we can trace back the history of a finite sample at all three loci. As the neutral loci are linked to the selected one, the genealogies at all three loci are correlated.

For the construction of the ancestral recombination graph relating all loci, time is running backward, so we set $t = T - t$. Conditioned on a frequency path $(X_t)_{0 \leq t \leq T}$, given by (2.1), we will describe the ancestral recombination graph as a partition-valued process $X = (X_t)_{0 \leq t \leq T}$.

Assume we take a sample from the population at time T . Every individual in the sample carries one L- and one R-locus. Of all L- and R-loci present in the sample we want to trace back a number ℓ of L- and r of R-loci. These loci are represented by sets ℓ for the L- and r for the R-loci. So, $|\ell| = \ell$, $|r| = r$. To define the state space of the structured ancestral recombination graph denote by $\mathcal{P}_{\ell \cup r}$ the set of partitions of A for a finite set A and define

$$\mathcal{P}'_{\ell \cup r} := \{ (\mathcal{B}, \mathcal{b}), \mathcal{B} \cup \mathcal{b} \in \mathcal{P}_{\ell \cup r}, \mathcal{B} \cap \mathcal{b} = \emptyset \}.$$

The coordinates \mathcal{B} and \mathcal{b} contain partition elements located in the beneficial and the wild-type background, respectively. For $(\mathcal{B}, \mathcal{b}) \in \mathcal{P}'_{\ell \cup r}$ we write (j) for the partition element containing $j \in \ell \cup r$.

The ancestral process is started at the time $t = 0$ of fixation of the beneficial allele. So, the sample of L- and R-loci is linked to the beneficial allele. Therefore, we start the process in $X_0 = (\emptyset, \emptyset)$ for some $(\mathcal{B}, \mathcal{b}) \in \mathcal{P}'_{\ell \cup r}$. Assume the state at time t is $X_t = (\mathcal{B}, \mathcal{b}) \in \mathcal{P}'_{\ell \cup r}$. For $j \in \ell \cup r$ the partition element which contains j , i.e., $(X_t)_{(j)}$, encodes the set of L- and R-loci, taken from the population at time $T - t$, which have the same ancestor at time $T - t$. Usually we will study the genealogy of n pairs of L- and R-loci. In this case set $\ell := \{1, \dots, n\}$ and $r := \{n + 1, \dots, 2n\}$ and start the process with $\mathcal{B} = \{\{1, n + 1\}, \dots, \{n, 2n\}\}$.

The dynamics of the process is given as follows: Coalescence events occur for lines in the beneficial and the wild-type background with pair coalescence rates λ and $\lambda(1 - X_{T-t})$ at time t , respectively. So, given $X_t = (\mathcal{B}, \mathcal{b})$ with $\mathcal{B} = \{\mathcal{B}_1, \dots, \mathcal{B}_{|\mathcal{B}|}\}$ and $\mathcal{b} = \{\mathcal{b}_1, \dots, \mathcal{b}_{|\mathcal{b}|}\}$ transitions occur for $1 \leq j \neq k \leq |\mathcal{B}|$ and $1 \leq j' \neq k' \leq |\mathcal{b}|$ from $(\mathcal{B}, \mathcal{b})$ to

$$\begin{aligned} & ((\mathcal{B} \setminus \{\mathcal{B}_j, \mathcal{B}_k\}) \cup \{\mathcal{B}_j \cup \mathcal{B}_k\}, \mathcal{b}) \quad \text{with rate } \frac{\lambda}{X_{T-t}}, \quad (1) \\ & (\mathcal{B}, (\mathcal{b} \setminus \{\mathcal{b}_{j'}, \mathcal{b}_{k'}\}) \cup \{\mathcal{b}_{j'} \cup \mathcal{b}_{k'}\}) \quad \text{with rate } \frac{\lambda(1 - X_{T-t})}{1 - X_{T-t}}, \quad (2) \end{aligned} \tag{2.2}$$

respectively. For transitions in the process \mathcal{X} due to recombination we focus on geometry (i) first. A recombination event hits one line between the S and the L locus with rate s_L and between the L and the R locus with rate s_R . If a recombination event occurs between the S and the L locus, it may be that both recombining chromosomes carry the same allele at the S locus. This gives a recombination event which cannot be seen effectively and we ignore it in the process \mathcal{X} . All other recombination events must be modeled. If $\mathcal{X} = (\mathcal{B}, \mathcal{b})$ with $\mathcal{B} = \{1, \dots, |\mathcal{B}|\}$ and $\mathcal{b} = \{1, \dots, |\mathcal{b}|\}$, transitions occur for $1 \leq j \leq |\mathcal{B}|$ and $1 \leq k \leq |\mathcal{b}|$ from $(\mathcal{B}, \mathcal{b})$ to

$$(\mathcal{B} \setminus \{j\}, \mathcal{b} \cup \{j\}) \quad \text{with rate } s_L(1 - X_{T-}) \quad (3_i)$$

$$((\mathcal{B} \setminus \{j\}) \cup \{j \cap \ell\}, \mathcal{b} \cup \{j \cap r\}) \quad \text{with rate } s_R(1 - X_{T-}) \quad (4_i)$$

$$((\mathcal{B} \setminus \{j\}) \cup \{j \cap \ell, j \cap r\}, \mathcal{b}) \quad \text{with rate } s_L X_{T-} \quad (5_i)$$

$$(\mathcal{B}, (\mathcal{b} \setminus \{k\}) \cup \{k \cap \ell, k \cap r\}) \quad \text{with rate } s_R(1 - X_{T-}) \quad (6_i)$$

$$(\mathcal{B} \cup \{k\}, \mathcal{b} \setminus \{k\}) \quad \text{with rate } s_L X_{T-} \quad (7_i)$$

$$(\mathcal{B} \cup \{k \cap r\}, (\mathcal{b} \setminus \{k\}) \cup \{k \cap \ell\}) \quad \text{with rate } s_R X_{T-} \quad (8_i)$$

(2.3)

Here, (3_i) encodes a recombination event which takes a pair of linked L - and R -loci from the beneficial to the wild-type background; an event (4_i) separates the R -locus of a line and takes it to the wild-type background; (5_i) the L and R loci of a line in the beneficial background are split but remain both in the same background; (6_i) describes the same transition for a line in the wild-type background. The transitions (7_i) and (8_i) describe the back-recombination of loci into the beneficial background.

Example 2.1 An example displaying the dynamics of the process \mathcal{X} for geometry (i) is shown in Fig. 2. The sets of L - and R -loci are $\ell = \{1, 2, 3\}$ and $r = \{4, 5, 6\}$, respectively. The starting partition is $\mathcal{X}^0 = (\mathcal{B}, \mathcal{b})$ with $\mathcal{B} = \{\{1, 4\}, \{2, 5\}, \{3, 6\}\}$. Several kinds of events can happen; coalescences in the beneficial background, i.e., an event (1), recombinations which leave the two neutral loci together but change the allele at the selected site, i.e., an event (2_i) and recombination events which split the two neutral loci. The last kind of event may either bring one of the two neutral loci in a different background (4_i), or split a line within the beneficial background (5_i), or split a line in the wild-type background (6_i). The final partition is $\mathcal{X}^1 = (\mathcal{B}, \mathcal{b})$ with $\mathcal{B} = \{\{1, 2\}\}$, $\mathcal{b} = \{\{3\}, \{4\}, \{5\}, \{6\}\}$.

For geometry (ii) we have (rescaled) recombination rates s_L and s_R between the left neutral and the selective and the right and the selective locus, respectively. Here, transitions occur from $(\mathcal{B}, \mathcal{b})$ to

$$\left((B \setminus \{j\}) \cup \{j \cap r\}, b \cup \{j \cap \ell\} \right) \text{ with rate } \lambda_S(1 - X_{T-}) \tag{3_i}$$

$$\left((B \setminus \{j\}) \cup \{j \cap \ell\}, b \cup \{j \cap r\} \right) \text{ with rate } \lambda_R(1 - X_{T-}) \tag{4_i}$$

$$\left((B \setminus \{j\}) \cup \{j \cap \ell, j \cap r\}, b \right) \text{ with rate } (\lambda_S + \lambda_R)X_{T-} \tag{5_i}$$

$$\left(B, (b \setminus \{k\}) \cup \{k \cap \ell, k \cap r\} \right) \text{ with rate } (\lambda_S + \lambda_R)(1 - X_{T-}) \tag{6_i}$$

$$\left(B \cup \{k \cap \ell\}, (b \setminus \{k\}) \cup \{k \cap r\} \right) \text{ with rate } \lambda_S X_{T-} \tag{7_{ii}}$$

$$\left(B \cup \{k \cap r\}, (b \setminus \{k\}) \cup \{k \cap \ell\} \right) \text{ with rate } \lambda_R X_{T-} \tag{8_i}$$

(2.4)

These events refer to a change in background from the beneficial to the wild-type background either for the L-locus, (3_i), or the R-locus, (4_i). Splits in the beneficial and wild-type background may happen as in the case of geometry (i); see events (5_i) and (6_i). Back-recombinations to the beneficial background are denoted (7_i) for the L- and (8_i) for the R-locus. Observe that a transition which takes both loci on one line from the beneficial to the wild-type background cannot occur for geometry (ii); cf. event (3_i).

Definition 2.2 Assume ℓ and r are sets of left and right neutral loci, respectively, and $\mathcal{X} = (X_t)_{0 \leq t \leq T}$ is a frequency path of the beneficial allele given by (1).

Conditioned on \mathcal{X} , consider the jump process $\mathcal{X} = (\mathcal{X}^x)_{0 \leq x \leq T}$, which starts in $\mathcal{X}_0^x = (\cdot, \emptyset)$ for $\cdot \in \mathcal{P}_{\ell \cup r}$ and makes transitions by coalescence events (1), (2), given by (2.2) and recombination events (7_i) or (3_i)-(8_i) from (2.3) and (2.4), respectively. This process \mathcal{X} is denoted the structured ancestral recombination graph for the L and R loci conditioned on \mathcal{X} for geometry (i) or (ii), respectively.

The mixture of \mathcal{X}_T^x over the distribution of frequency paths given by (1) defines the random partition $\mathcal{X} = (B, b)$, i.e.,

$$:= \int \mathcal{X}_T^x \mathbb{P}[d\mathcal{X}].$$

3 Main result

We study selective sweeps in the infinite population limit, i.e., the frequency of the beneficial allele follows the SDE given by (1). Moreover, selection is most efficient for large selection coefficients. Our goal is to derive a simpler but approximate expression for $\mathbb{E}[T]$ in the regime of large λ . It was shown in [3] that for the fixation time T of the beneficial allele

$$\mathbb{E}[T] = \frac{2 \log \lambda}{\lambda} + \mathcal{O}\left(\frac{1}{\lambda}\right), \quad \mathbb{V}[T] = \mathcal{O}\left(\frac{1}{\lambda^2}\right) \tag{3.1}$$

for large λ . This suggests that only under the scaling $\lambda \asymp \mathcal{O}(\lambda / \log \lambda)$ for the recombination rate a non-trivial number of recombination events occurs during the sweep for

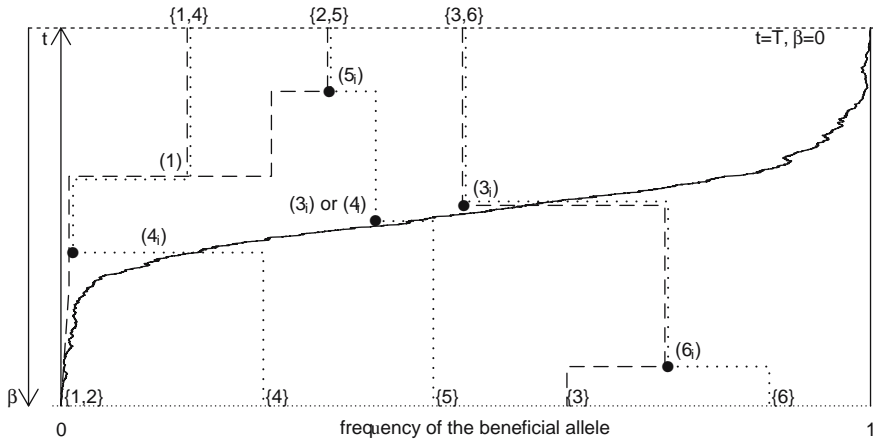


Fig. 2 A structured ancestral recombination graph conditioned on the frequency path of the beneficial allele. Between times $t = 0$ and $t = T$ coalescences may occur at rates β_1 and β_2 . Recombination events happen at rates $\gamma_1 - \gamma_2$. The dashed lines indicate ancestry of the R-locus while the R-locus may be traced along dotted lines

large β . This is true for all possible kinds of recombination events during the sweep, so the recombination rates β_{SL} , β_{LR} and β_{LS} , β_{SR} for geometries (i) and (ii) should be of this order. Henceforth, we assume

$$\text{Geometry (i): } \beta_{SL} = \beta_{SL} \frac{1}{\log}, \quad \beta_{LR} = \beta_{LR} \frac{1}{\log}, \quad 0 < \beta_{SL}, \beta_{LR} < \infty$$

$$\text{Geometry (ii): } \beta_{LS} = \beta_{LS} \frac{1}{\log}, \quad \beta_{SR} = \beta_{SR} \frac{1}{\log}, \quad 0 < \beta_{LS}, \beta_{SR} < \infty.$$

Our approximation of β is based on a Yule tree, which serves as an approximation of the genealogy at the selected locus. A Yule tree is the realization of a Yule process, i.e., a pure birth process which starts with one line and every line splits in two lines after an exponential waiting time.

In our approximation the quantity

$$p_{i_1}^{i_2}(\beta) := \exp\left(-\frac{1}{\log} \sum_{i=i_1+1}^{i_2} \frac{1}{i}\right) \tag{3.2}$$

will play an important role.

Assume ℓ and r are sets of left and right loci and $\mathcal{P} \in \mathcal{P}_{\ell \cup r}$. Three mechanisms determine the Yule approximation of the partition. First, we approximate splits in the beneficial background, i.e., events (s_i) and (5_i) , by the following procedure:

For all partition elements $1, \dots, | \mathcal{P} |$, realize Bernoulli random variables $U_1, \dots, U_{| \mathcal{P} |}$ which are 1 with success probability

$$\text{geometry (i): } 1 - p_0^{i_2} \beta_{LR} \quad \text{geometry (ii): } 1 - p_0^{i_2} \beta_{LS + SR}. \tag{3.3}$$

If $U_i = 1$, split the i th partition element in its left and right locus. Altogether, this defines a partition

$$\nu = \{ i \cap \ell, i \cap r : U_i = 1 \} \cup \{ i : U_i = 0 \}.$$

Next, realize a Yule process with branching rate λ , i.e., each line splits in two lines at rate λ . Stop this process when it has $\lfloor 2 \rfloor$ lines. Call this tree \mathcal{Y} . To obtain the genealogy of a sample of size n from this tree with $\lfloor 2 \rfloor$ extant leaves, we use the following construction:

Start with $\lfloor n \rfloor$ lines from the full Yule tree \mathcal{Y} with $\lfloor 2 \rfloor$ lines. When there are k lines left at the time the full tree has i lines, the probability that a coalescence event occurs among the k lines at the time the full tree goes from i to $i - 1$ lines is

$$\frac{\binom{k}{2}}{\binom{i}{2}}. \tag{3.4}$$

By this construction we build a tree \mathcal{Y}_n with the partition elements of ν as leaves and nodes which record the number of lines in the full Yule tree.

Remark 3.1 To construct the sample tree \mathcal{Y}_n from \mathcal{Y} is a task equivalent to describing an exchangeable sample from a tree which arises by exchangeable binary coalescence dynamics. This has been studied [23] and was recalled in [3, Lemma 4.8]. If $t = i$ is the number of lines in the Yule tree at time t , denote by K_i the number of lines in \mathcal{Y}_n while $t = i$. The process $\{K_i\}_{\lfloor 2 \rfloor \geq i \geq 1}$ is a time-inhomogeneous Markov chain with transition probabilities

$$\mathbb{P}[K_{i-1} = k - 1 | K_i = k] = \frac{\binom{k}{2}}{\binom{i}{2}}, \quad i = 2, \dots, \lfloor 2 \rfloor, k = 2, \dots, \lfloor n \rfloor.$$

Moreover, the sample tree can be described forward in time by noting that

$$\mathbb{P}[K_i = k | K_{i-1} = k - 1] = \frac{\lfloor n \rfloor - k + 1}{\lfloor n \rfloor + i - 1}.$$

□

The sample tree which is pruned out of the full tree in this way represents the genealogy at the selected site. To describe the genealogies at the partially linked neutral sites we mark the sample Yule tree to determine further recombination events. A mark stands for one (or two) recombination events that may occur. This works in the following way:

Table 1 For geometry (i), we mark every branch in the Yule tree by at most one from three different kinds of events

Mark	Probability
SL	$(1 - p_1^2(\text{SL}))p_0^2(\text{LR})$
LR	$p_1^2(\text{SL})(1 - p_1^2(\text{LR}))$
SLR	$(1 - p_1^2(\text{SL}))(1 - p_0^2(\text{LR}))$
no	$p_1^2(\text{SL})p_1^2(\text{LR})$

If a branch starts when the full Yule tree has i_1 lines and ends when it has i_2 lines, the probabilities for all marks are given in the table

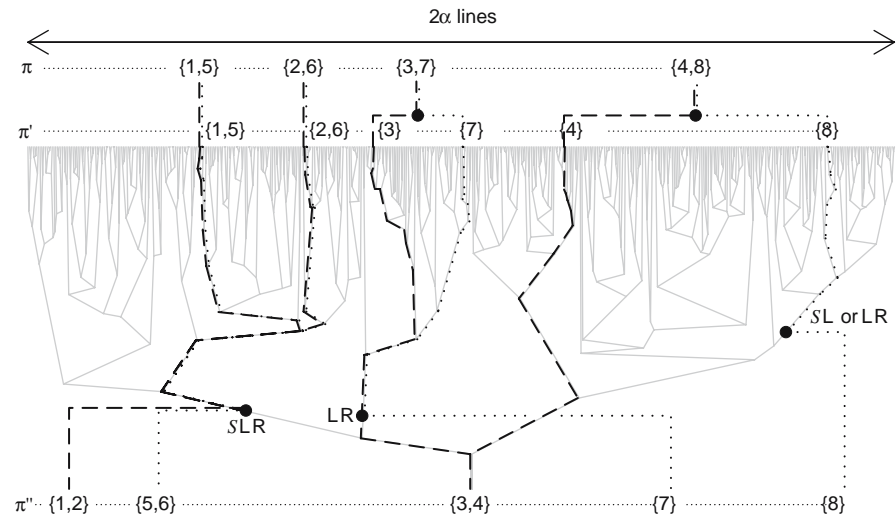


Fig. 3 The Yule process approximation for two linked neutral loci under a selective sweep. Here, we consider geometry (i). The L-locus may be traced back along dashed lines while dotted lines indicate ancestry of the R-locus. See text for explanation

Let a branch in the tree be given which starts when the full tree has i_1 lines and ends when the full genealogy has i_2 lines. For geometry (i), every branch can be hit by at most one of three different kinds of marks indicating recombination events. These are SL-, LR-, and SLR-marks. Their probabilities are given in Table 1. For geometry (ii) the branch is hit independently by LS- and SR-marks with probabilities $(1 - p_1^2(\text{LS}))$ and $(1 - p_1^2(\text{SR}))$. (3.5)

Here, SL-marks separate the S from the L-locus on each branch of the tree, etc. For geometry (ii) LR-marks separate the L from the L- and the L- from the R-locus.

Example 3.2 The above construction is illustrated in Fig. 3. We consider geometry (i) here. A set $\ell = \{1, 2, 3, 4\}$ of L-loci and $r = \{5, 6, 7, 8\}$ of R-loci is given. Starting

with $\mathcal{P} = \{\{1, 5\}, \{2, 6\}, \{3, 7\}, \{4, 8\}\}$, every partition element is split with probability $p_0^{1/2}$ according to (3.3). This results in the finer partition \mathcal{P}' . The partition elements of \mathcal{P}' are used to construct a sample tree from a full Yule tree which has 2n lines. The coalescence probabilities for the sample are given by (3.4). (On the sample tree, branches are marked SL-, LR-, or SLR-marks according to Table 1. The resulting partition \mathcal{P}'' is constructed as given in Definition 3.3.

We are now in a position to define our approximation based on the Yule process.

Definition 3.3 Assume ℓ and r are sets of left and right neutral loci, respectively, and $\mathcal{P} \in \mathcal{P}_{\ell \cup r}$. By (3.3) construct the partition \mathcal{P}' and by (3.4) and (3.5) a Yule tree \mathcal{T}_1 with marks. For geometry (i) define the equivalence relation:

$$j \sim k : \Leftrightarrow \left\{ \begin{array}{ll} \text{no SL-, SLR-mark on } \begin{array}{c} \text{\textit{(j)}} \\ \text{\textit{(k)}} \end{array} & \text{if } j, k \in \ell, \\ \text{no SL-, LR-, SLR-mark on } \begin{array}{c} \text{\textit{(j)}} \\ \text{\textit{(k)}} \end{array} & \text{if } j, k \in r \\ \text{no SL-mark on } \begin{array}{c} \text{\textit{(j)}} \\ \text{\textit{(k)}} \end{array} & \\ \text{no LR-mark on } \begin{array}{c} \text{\textit{(j)}} \\ \text{\textit{(k)}} \end{array} & \text{if } j \in \ell, k \in r \\ \text{no SLR-mark on } \begin{array}{c} \text{\textit{(j)}} \\ \text{\textit{(k)}} \end{array} & \end{array} \right. \quad (3.6)$$

where the bold lines indicate for which part of the tree relating two lines with the root of the tree, the constraint on marks applies. For geometry (ii) set

$$j \sim k : \Leftrightarrow \left\{ \begin{array}{ll} \text{no LS-mark on } \begin{array}{c} \text{\textit{(j)}} \\ \text{\textit{(k)}} \end{array} & \text{if } j, k \in \ell, \\ \text{no SR-mark on } \begin{array}{c} \text{\textit{(j)}} \\ \text{\textit{(k)}} \end{array} & \text{if } j, k \in r, \\ \text{no LS-mark on } \begin{array}{c} \text{\textit{(j)}} \\ \text{\textit{(k)}} \end{array} & \\ \text{no SR-mark on } \begin{array}{c} \text{\textit{(j)}} \\ \text{\textit{(k)}} \end{array} & \text{if } j \in \ell, k \in r \end{array} \right. \quad (3.7)$$

(The equations (3.6) and (3.7) indeed define equivalence relations, as can easily be checked.) Each of these equivalence relations defines a partition π . For geometry (i) there is a unique partition element

$$\pi_f = \left\{ j \in \ell : \text{no SL-, SLR-mark on } \mathbf{I}^{(j)} \right\} \cup \left\{ k \in r : \text{no SL-, LR-, SLR-mark on } \mathbf{I}^{(k)} \right\} \tag{3.8}$$

and for geometry (ii) a unique partition element

$$\pi_f = \left\{ j \in \ell : \text{no LS-mark on } \mathbf{I}^{(j)} \right\} \cup \left\{ k \in r : \text{no SR-mark on } \mathbf{I}^{(k)} \right\}. \tag{3.9}$$

Then the random partition

$$:= (\pi_f, \pi \setminus \pi_f)$$

is called the Yule approximation of π .

Example 3.4 For the example in Fig 3 the SL-, LR- and SLR-marks on the sample tree lead to the realization

$$= (\{\{3, 4\}\}, \{\{1, 2\}, \{5, 6\}, \{7\}, \{8\}\}).$$

Theorem 1 Let $\pi \in \mathcal{P}_{\ell \cup r}$ and π_f and $\pi \setminus \pi_f$ be as in Definition 3.2 and 3.3. Then,

$$\sup_{\pi' \in \mathcal{P}'_{\ell \cup r}} |\mathbb{P}[\pi = \pi'] - \mathbb{P}[\pi = \pi'_f]| = \mathcal{O}\left(\frac{1}{(\log)^2}\right).$$

Remark 3.51. The Theorem states that, for large n , the random partitions π and π_f are close in variation distance. Here, variation distance refers to the maximal difference in the probabilities to obtain any partition $\pi' \in \mathcal{P}'_{\ell \cup r}$. The order of accuracy, given by the Landau symbol, still depends on several parameters. These are the cardinalities n_ℓ and n_r and recombination constants $s_{L, LR}$ for geometry (i) and s_{LS} and s_{SR} for geometry (ii). The proof of Theorem 1 will be given in Sect 5.

- At first sight, comparing the Definitions 3.3 and 2.2 the Yule approximation does not look any simpler than the exact model. However, the Yule approximation has advantages both analytically and computationally. The random partitions π relies on constructing a frequency path, while the Yule approximation π_f constructs the ancestral recombination graph for the sample directly. Analytically, as we will see in Sect 4, this means that explicit calculations are possible. Computationally, i.e., for simulations of the ancestral recombination graph, the direct construction of the ancestry of the sample allows for fast algorithms; see for the case of a single neutral locus.

3. The current paper is a generalization of results found in [1] for a two-locus system with only one neutral locus. More precisely, consider the projection π on only one locus, i.e., on either ℓ or r . In Propositions 4.2 and 4.7 of that paper it was shown that the projection of π on ℓ or r is an approximation to a structured coalescent with an error in probability of the order $(\log n)^{-2}$.
4. In [3] an approximate sampling formula was given in the two-locus case. A similar approach would be possible here. However, we refrain from its derivation because it was shown in [2] that the sampling formula in the two-locus case only produces numerically sound results for $n \leq 5$.
5. As indicated numerically in [2], the Yule approximation can be improved. To understand how this works, we need to collect the errors which contribute to the error of order $O(1/(\log n)^2)$. First, the Yule approximation ignores events (6), (7) and (8). Second, as will be clear in the proof of Proposition 5, the coalescent rate in the beneficial background is decreased from $X dt$ to $(1-X)/X dt$ by the Yule process. It is the latter error that dominates, at least in large samples, because the total coalescence rate increases quadratically with the number of lines. However, increasing the coalescence probability α to

$$1 \wedge \frac{\binom{k}{2}}{\binom{i}{2}} \frac{1}{1 - \frac{i-1}{2}}$$

- at the time the Yule tree has i lines corrects for this error.
6. For simulations of genealogies it is most important that the Yule approximation given above is not restricted to the case of two neutral loci. The take-home-message from the construction of the Yule approximation is that splits in the beneficial background are generated first and afterwards marks on a Yule tree determine all recombination events. Both, splits in the beneficial background and recombination events along the Yule tree can be given along a continuous chromosome.

4 Application: *D*

Lewontin's *D* is a measure of linkage disequilibrium (non-random association of alleles) and is frequently used as a simple statistic in a multi-locus setting (see also [5, (2.89)]). Given two loci L and R with alleles 0 or 1 at each locus, it is defined as

$$D = p_{LR} - p_L p_R \tag{4.1}$$

where p_{LR} is the frequency of individuals carrying allele 1 at both loci, p_L is the frequency of 1's at the L locus and p_R is the frequency of 1's at the R locus..

To predict patterns of *D* between pairs of neutral loci at the time of fixation of a beneficial allele we next approximate $E[D(T)]$ using Theorem 1. It is crucial to observe that $E[p_{LR}(T)]$ as well as $E[p_L(T)p_R(T)]$ may be derived by the distribution of genealogies of linked neutral loci under selection and the expected allele frequencies at the beginning of the sweep. To see this, note that $E[p_{LR}(T)]$ equals the probability that the ancestors of the L - and R -locus of one randomly picked individual from the

population at time T carry alleles 1 at both neutral loci. Analogously, $p_L(T)p_R(T)$ is the probability that the ancestors of the L- and R- loci of two different individuals at time T both carry allele 1. Denote q by the probability that both loci L and R from one individual, picked at time T , have a common ancestor at the beginning of the sweep. Analogously, q' is the same probability for the L- and R-loci from two different individuals. Using these definitions we see that

$$\begin{aligned} \mathbb{E}[p_{LR}(T)] &= q \cdot \mathbb{E}[p_{LR}(0)] + (1 - q) \cdot \mathbb{E}[p_L(0)p_R(0)], \\ \mathbb{E}[p_L(T)p_R(T)] &= q' \cdot \mathbb{E}[p_{LR}(0)] + (1 - q') \cdot \mathbb{E}[p_L(0)p_R(0)]. \end{aligned} \tag{4.2}$$

Combining (4.2) with the definition of D from (4.1),

$$\mathbb{E}[D(T)] = (q - q')\mathbb{E}[D(0)]. \tag{4.3}$$

Both q and q' may be approximated by Theorem 1. Formally, setting $g = \{1\}$, $r = \{2\}$,

$$\begin{aligned} q &= \mathbb{P} \left[\binom{B}{\{1,2\}} \cup \binom{b}{\{1,2\}} = \{\{1,2\}\} \right], \\ q' &= \mathbb{P} \left[\binom{B}{\{1\},\{2\}} \cup \binom{b}{\{1\},\{2\}} = \{\{1,2\}\} \right]. \end{aligned} \tag{4.4}$$

As q may be approximated by (4.3) this brings us in a position to predict patterns of D at the end of a selective sweep.

Theorem 2 For geometry (i),

$$\mathbb{E}[D(T)] = p_0^2 (2 - p_{LR}) \left(1 - \sum_{k=2}^{\infty} \frac{2}{k(k+1)} p_k^2 (2 - p_{SL}) \right) \mathbb{E}[D(0)] + \mathcal{O}\left(\frac{1}{(\log T)^2}\right), \tag{4.5}$$

and for geometry (ii),

$$\mathbb{E}[D(T)] = \mathbb{E}[D(0)] \cdot \mathcal{O}\left(\frac{1}{(\log T)^2}\right). \tag{4.6}$$

Remark 4.11. Patterns of Lewontin's D can be studied by deterministic forward calculations instead of our genealogical approach. This was carried out in [26] under the assumption that strong selection leads to a deterministic behavior of allele frequencies. Specially, the frequency of the beneficial allele follows the logistic differential equation

$$dX = X(1 - X)dt, \quad X_0 = \frac{1}{N}$$

instead of the stochastic path given by (4.1). Predictions of D at all times during the selective sweep were given. In particular, their equation (47) approximates values of D at the end of the sweep for geometry (i).

In real populations, random effects due to genetic drift are not negligible. This has been pointed out by [4]. The Yule process approximation captures most random effects. Indeed, comparison with simulations from [4] shows that the results produced by the Yule process approximation are more accurate than those of [25] (see Fig 4).

2. For empirical studies it is most interesting to know which patterns of linkage disequilibrium to look for in real data. The pattern genetic hitchhiking can produce was discussed in [25] and [22]. Surprisingly, hitchhiking reduces levels of linkage disequilibrium compared to the neutral expectation. This is evident from Fig. 4. If the selected locus is far from both neutral loci, linkage disequilibrium between the neutral loci is not affected by hitchhiking. Therefore, values D for large s_L converge to the expectation D under neutrality. This effect was taken up by [22] to argue that genetic hitchhiking produces patterns in the association of alleles similar to recombination hotspots, which are e.g. important in genetic association studies in humans [2]. However, genetic hitchhiking certainly produces patterns different from recombination hotspots in general, e.g., a low neutral diversity or a distinctive site frequency spectrum [6].
3. An accurate approximation of $E[D(T)]$ does not suffice to predict patterns of linkage disequilibrium in general. In addition to genetic drift, random effects which affect $D(T)$ were found in [25] to be the allelic type of the founder of the sweep and its frequency. The resulting variance D can be considerably higher than under neutrality.

Now we come to the proof of Theorem 2.

Proof The key in the proof is to compute the probabilities p and q' . This is achieved by the Yule process approximation of Theorem 1.

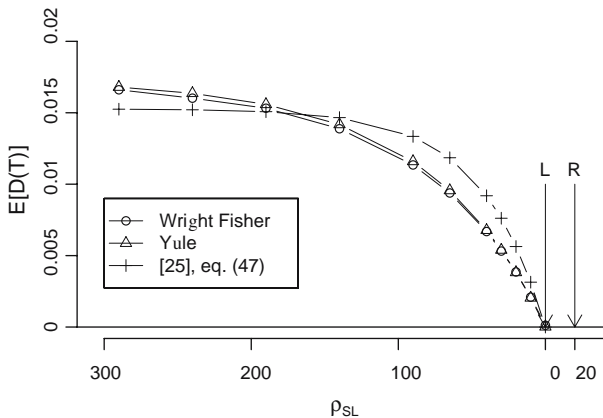


Fig. 4 The effect of Lewontin's D under a selective sweep may be simulated in a Wright–Fisher model. In this process, the frequency path of the beneficial allele is stochastic and the ancestral recombination graph may be built conditioned on this frequency path. The locations of the L and R locus are fixed. The position of the selected site varies along the x -axis. If we compare the result from Fig. 5 to equation (47) of [25] we see that the Yule process approximation is more accurate. The parameters of the Wright–Fisher model are $N = 10^5$, $s = 1000$, $L_R = 20$ and $D(0) = 0.0242$

We start with geometry (ii). Here, we can see from the Yule approximation (that $q = q'$ up to a term of order $1/(\log)^2$ since oneL and oneR locus are identical



by descent iff there is no S mark on  and no SR mark on . This


does not depend on the linkage of theL and theR locus at the end of the sweep. Consequently, (4.6) follows.

For geometry (i), we start with the approximation of For oneL and oneR locus from two different individuals there is a random number of lines in the full tree of the Yule approximation at the time the selected loci which are linked to the neutral ones coalesce. To obtain the distribution of K , we compute


$$\begin{aligned} \mathbb{P}[K = k] &= \prod_{l=k+1}^2 \left(1 - \frac{1}{\binom{l}{2}}\right) \frac{1}{\binom{k}{2}} \\ &= \left(\prod_{l=k+1}^2 \frac{(l+1)(l-2)}{l(l-1)}\right) \frac{2}{k(k-1)} = \frac{2}{k(k+1)} + \mathcal{O}\left(\frac{1}{k}\right), \end{aligned}$$

which is a special case of (4.16). We read from (3.6) that theL and R locus are identical by descent at the beginning of the sweep if and only if (a) no mark on L

mark falls on , (b) no mark hits  and (c) no mark or an R mark

falls on . Hence we compute

$$\begin{aligned} q' &= \sum_{k=2}^2 \frac{2}{k(k+1)} p_0^k(LR) p_k^2(SL) p_k^2(LR) p_k^2(SL) + \mathcal{O}\left(\frac{1}{(\log)^2}\right) \\ &= p_0^2(LR) \sum_{k=2}^2 \frac{2}{k(k+1)} p_k^2(2SL) + \mathcal{O}\left(\frac{1}{(\log)^2}\right). \end{aligned} \tag{4.7}$$

For q we have to distinguish the cases where theL and theR-loci split or not. If they do not split, theL- and R-locus have the same ancestor at the beginning of the sweep if and only if there is neither anR- nor anSLR-mark on . If they split, the probability of a common ancestor is Therefore,

$$q = p_0^2(LR) p_0^2(LR) + (1 - p_0^2(LR)) q' + \mathcal{O}\left(\frac{1}{(\log)^2}\right). \tag{4.8}$$

Hence

$$\mathbb{E}[D(T)] = p_0^2 (\lambda_{LR})(p_0^2 (\lambda_{LR}) - q')\mathbb{E}[D(0)] + \mathcal{O}\left(\frac{1}{(\log T)^2}\right) \tag{4.9}$$

and the result follows.

5 Proof of Theorem 1

The proof deals with geometries (i) and (ii) simultaneously. We will write events at rates (3)–(8) whenever we refer to the rates (3)–(8) for geometry (i) and (3̄)–(8̄) for geometry (ii), respectively.

We will be dealing with several random partitions all of which agree up to an error of order $\mathcal{O}((\log T)^{-2})$. Exactly, we will prove

$$\overset{\text{Prop.5.2}}{\approx} \overset{\text{Prop.5.5}}{\approx} \overset{\text{Prop.5.6}}{\approx}$$

where π , π' , and π'' are given in Definitions 2.2, 5.1, 5.3 and 3.3, respectively and \approx means that the random partitions differ by $\mathcal{O}((\log T)^{-2})$ in variation distance.

While π is the random partition which is defined by the structured ancestral recombination graph, the other random partitions are approximations. First arises by (i) ignoring events which occur according to rates (2), (6_i), (7) and (8) and (ii) realizing all events according to rates (6) first and only afterwards, construct the process using rates (1), (3), (4) and (6_i). Second, π' already deals with the Yule process. It is derived by marking an infinite Yule tree by two constant rate Poisson processes with rates $\lambda_{SL}, \lambda_{LR}$ for geometry (i) and $\lambda_{LS}, \lambda_{SR}$ for geometry (ii). Finally, the Yule approximation π'' arises by considering only the number of lines in an infinite Yule tree at times of coalescence in a sample.

In the whole proof we rely on a probability measure on a probability space on which the solution of 1.1) as well as arbitrarily many independent Poisson processes and other random variables are realized.

Definition 5.1 Define a $\mathcal{P}'_{\ell \cup r}$ -valued random variable π' as follows: starting in $\ell \in \mathcal{P}_{\ell \cup r}$ split all partition elements $\in \ell$ independently into $\cap \ell, \cap r$ with probability

$$1 - \mathbb{E}\left[\exp\left(-\int_0^T X_s ds\right)\right] \tag{5.1}$$

where $\lambda = \lambda_{LR}$ for geometry (i) and $\lambda = \lambda_{LS} + \lambda_{SR}$ for geometry (ii). The resulting partition π' is used for the starting point (π', \emptyset) of a process $X = (X_t)_{0 \leq t \leq T}$, conditioned on a frequency path $\mathbf{h} = (X_t)_{0 \leq t \leq T}$ with transitions according to events (1), (3), (4), (6_i), given by 2.3, for geometry (i) and to events (1), (3̄) and (4_i), given by 2.4, for geometry (ii), respectively. Given X , define

$$:= \int_T^X \mathbb{P}[d\mathcal{X}].$$

Proposition 5.2 Let $\ell \in \mathcal{P}_{\ell \cup r}$ and $r \in \mathcal{P}_r$ be as in Definition 8.2 and 5.1. Then,

$$\sup_{\ell \in \mathcal{P}'_{\ell \cup r}} |\mathbb{P}[X_{T-} = \ell] - \mathbb{P}[X_{T-} = r]| = \mathcal{O}\left(\frac{1}{(\log N)^2}\right).$$

Proof We proceed in several steps. Our arguments in Step 1 show that we may discard events which occur at rates (2), (6), (7) and (8). In Step 2 we use a fixed number of Poisson processes to generate the random partition we want to approximate. Our goal is to separate events (5) from the rest by verifying a certain order of the possible events and establishing an approximate independence of the events (5). Particularly, we show in Step 3 that splits in the beneficial background (i.e., events (5)) take place before all other events with high probability. The approximate independence will be proved in Steps 5 and 6 by an application of a general result on mixed Poisson processes we establish in Step 4.

Step 1 (Small probability of events (2), (6), (7) and (8))

First, note that by Proposition 3.4 of [3] events (2), i.e., coalescences in the wild-type background, have a probability of order $\mathcal{O}((\log N)^{-2})$. Furthermore, events (7) and (8) are back-recombinations into the beneficial background and hence, again referring to [3], have a probability of order $\mathcal{O}((\log N)^{-2})$ as well. Additionally, for geometry (ii), events (6), i.e., splits in the wild-type background, can only occur if a coalescence event (2) has happened before. As a consequence, we can discard events which occur at rates (2), (6), (7) and (8) producing only an error in variation distance of at most $\mathcal{O}((\log N)^{-2})$.

So we are left with a $\mathcal{P}'_{\ell \cup r}$ -valued stochastic process conditioned at $X^x = (X^x)_{0 \leq t \leq T}$, which arises by events (1), (3), (4), (5) and (6_i), and is started in $X^x_0 = (\ell, \emptyset)$.

Step 2 (Construction of X^x by Poisson processes)

Recall that $\ell := |\ell|$ and $r := |r|$ are the number of ℓ and R loci under consideration. Take Poisson processes which are all conditionally independent given the random frequency path X^x of the beneficial allele. For coalescence, take a Poisson process with

$$\text{rate} \left(\frac{\ell + r}{2} \right) \frac{1}{X_{T-}} \quad (\text{coalescence in the beneficial background}) \quad (1), \tag{5.2}$$

at time t ; for recombination events take Poisson processes $\mathcal{T}_{4i}, \mathcal{T}_{5i}$ with

$$\begin{aligned} \text{rate} &= \ell(1 - X_{T-}) \quad (\text{rec. to the wild-type background}) \quad (3_i), \\ \text{rate} &= r(1 - X_{T-}) \quad (\text{rec. to or split in the wild-type background}) \quad (4_i), \\ \text{rate} &= \ell X_{T-} \quad (\text{split in the beneficial background}) \quad (5_i), \end{aligned} \tag{5.3}$$

at time t for geometry (i) and Poisson processes $\mathcal{S}_3, \mathcal{T}_{4ii}, \mathcal{T}_{5ii}$ with

$$\text{rate } \lambda_S(1 - X_{T-}) \quad (\text{rec. to the wild-type background}) \quad (3ii),$$

$$\text{rate } \lambda_{SR}(1 - X_{T-}) \quad (\text{rec. to the wild-type background}) \quad (4ii),$$

$$\text{rate } (\lambda_S + \lambda_{SR})X_{T-} \quad (\text{split in the beneficial background}) \quad (5ii),$$

$$(5.4)$$

at time t for geometry (ii). We have combined recombinations to the wild-type and splits in the wild-type background in case of geometry (ii) since they happen with the same rates.

Additionally, let $W = (W_{i,m})_{i=1,3,4,5,m=1,2,\dots}$ be a random array such that all $W_{i,m}$'s are independent. $W_{1,m}$ is uniformly distributed on all pairs of $\ell \cup r$, $W_{3,m}$ is uniformly distributed on ℓ , and $W_{4,m}$ and $W_{5,m}$ are uniformly distributed on r , $m = 1, 2, \dots$

The set $\ell \cup r$ can be totally ordered, so we may assume that every partition element in $\pi \in \mathcal{P}'_{\ell \cup r}$ has a smallest element. Recall that we write (j) for the partition element containing $j \in \ell \cup r$.

We abbreviate by $\mathcal{T}_3 - \mathcal{T}_5$ the Poisson processes $\mathcal{S}_3 - \mathcal{T}_{5i}$ for geometry (i) and the Poisson processes $\mathcal{S}_{3ii} - \mathcal{T}_{5ii}$ for geometry (ii). Next, we consider that the distribution of π^x is the image measure of the tuple $(\pi, \mathcal{T}_3, \mathcal{T}_4, \mathcal{T}_5, W)$ under a map. Specifically, the distribution of π^x is uniquely determined by the distribution $(\pi, \mathcal{T}_3, \mathcal{T}_4, \mathcal{T}_5, W)$.

To define π^x , consider a discrete set $\mathcal{B}_1 \subseteq [0, T]$ and finite sets $\mathcal{T}_3, \mathcal{T}_4, \mathcal{T}_5 \subseteq [0, T]$ such that $\mathcal{T}_{i_1} \cap \mathcal{T}_{i_2} = \emptyset$ for $i_1 \neq i_2$ and $\text{set } \mathcal{T} = \bigcup_i \mathcal{T}_i$. Furthermore $W = (W_{i,m})_{i=1,3,4,5,m=1,2,\dots}$ such that for all $m = 1, 2, \dots$, $W_{1,m}$ is a pair in $\ell \cup r$, $W_{3,m} \in \ell$ and $W_{4,m}, W_{5,m} \in r$. Given $(\mathcal{T}_3, \mathcal{T}_4, \mathcal{T}_5, W)$ we generate a partition by considering the events in \mathcal{T} in decreasing order. Assume $\pi^x = (\pi, \emptyset)$ and after the $(m - 1)$ st event at time t we obtain a partition $\pi^x = (\mathcal{B}, b) \in \mathcal{P}'_{\ell \cup r}$ and the m th event in \mathcal{T} to be realized happens at time $t \in \mathcal{T}$.

Consider first the case that t is the m th event in \mathcal{T} and the i_1 st event in $\mathcal{T}_1 \in \mathcal{T}_3$. The pair $W_{1,m_1} = (j, k)$ gives a random pair of loci. If $(j), (k) \in \mathcal{B}$ and if both, j and k , are the smallest elements of their partition elements, coalesce these partition elements, i.e., make the transition

$$(\mathcal{B}, b) \rightarrow ((\mathcal{B} \setminus \{(j), (k)\}) \cup \{(j) \cup (k)\}, b).$$

Otherwise do nothing.

The next case to consider is that t is the m_3 rd event in \mathcal{T}_3 and $W_{3,m_3} = j$ for some $j \in \ell$. If $(j) \in \mathcal{B}$ and if j is the smallest element of $(j) \cap \ell$, change the partition element from \mathcal{B} to $\mathcal{B} \cup b$, i.e., make the transition

$$(\mathcal{B}, b) \rightarrow (\mathcal{B} \setminus \{(j)\}, \mathcal{B} \cup \{(j)\}). \quad (5.5)$$

Otherwise do nothing. The case \mathcal{T}_5 is similar and is omitted.

If τ_j is the m_4 th event in T_4 and $w_{4,m_4} = j$ for $j \in r$ the partition again only changes if $j = \min_{(j) \cap r}$. We distinguish two cases $(j) \in B$ and $(j) \in b$. In the former case, split the ℓ - and R -loci in the partition element in two partition elements and bring all R -loci into the wild-type background, i.e., make the transition

$$\left(B, b \right) \longrightarrow \left(\left(B \setminus \left\{ \frac{B}{(j)} \right\} \right) \cup \left\{ \frac{B}{(j)} \cap \ell \right\}, b \cup \left\{ \frac{B}{(j)} \cap r \right\} \right). \tag{5.6}$$

This corresponds to an event (4). In the latter case split ℓ and R -loci of (j) and leave them in the wild-type background, i.e., make the transition

$$\left(B, b \right) \longrightarrow \left(B, \left(b \setminus \left\{ (j) \right\} \right) \cup \left\{ (j) \cap \ell, (j) \cap r \right\} \right), \tag{5.7}$$

which corresponds to an event (6). Recall that for geometry (ii) one ℓ - and one R -locus cannot recombine to the wild-type background together. Hence partition elements in b are either subsets of r or of ℓ such that the last transition must not occur for this geometry.

By generating all events according to this procedure we end with a partition π^X . Therefore we have defined the map $(T_1, T_3, T_4, T_5, w) \mapsto \pi^X$.

The distribution of π^X is the image measure $(\pi^X, T_1, T_3, T_4, T_5, w)$ under the map π . (5.8)

To see this, observe first, that there are only finitely many recombination events (3), (4), (5) and (6). Almost surely, all events in the Poisson processes occur at different times, so π is defined on a set of probability 1. By the above construction, we obtain that two partition elements in B coalesce by event (1). The Poisson processes $\mathcal{T}_3, \mathcal{T}_4, \mathcal{T}_5$ produce exactly the recombination events (3), (4), (5) and (6) and hence (5.8) is proved.

Given w , the random partition (T_1, T_3, T_4, T_5, w) only depends on the order of time points in T_1, T_3, T_4, T_5 . There is another feature we will need:

Let τ', τ'' be consecutive time points in \mathcal{T} with $\tau' \in T_3, \tau'' \in T_4$. Exchanging τ' and τ'' does not alter the random partition (T_1, T_3, T_4, T_5, w) . Formally, if $T \cap (\tau', \tau'') = \emptyset, T'_3 = T_3 \setminus \{\tau'\} \cup \{\tau''\}$ and $T'_4 = T_4 \setminus \{\tau''\} \cup \{\tau'\}$. Then (5.9)

$$(T_1, T'_3, T'_4, T_5, w) = (T_1, T_3, T_4, T_5, w).$$

Assume τ' is the m_3 rd event in $T_3, w_{3,m_3} = j$ and τ'' is the m_4 th event in T_4 and $w_{4,m_4} = k$. If j and k are not in the same partition element for τ', τ'' , the claim is trivial as recombination events only make the partition finer. Similarly, if $j > \min_{(j) \cap \ell}$ or $k > \min_{(k) \cap r}$ only one transition occurs and the claim follows. In the case

$$(j) = (k), \quad j = \min_{(j) \cap \ell}, \quad k = \min_{(j) \cap r}$$

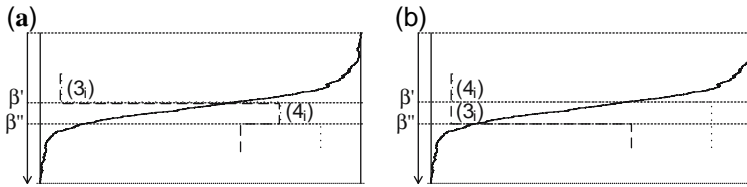


Fig. 5 a A partition element (a line) is hit by an event taking both the L- and the R-locus to the wild-type background at time t' . Afterwards, at time t'' the line is split in the wild-type background. Here, the R-locus is taken to the wild-type background at time t' . Afterwards the L-locus is taken to the same background at time t'' . The outcome is the same. The line moves from the beneficial to the wild-type background and is split there

two transitions occur if and only if $(j) = (k) \in B$. We illustrate this situation in Fig. 5.

Observe that the two-step transitions for the pair $((5.6), (5.7))$ (see Fig 5a) as well as for the pair $((5.6), (5.5))$ (see Fig 5b) are given by

$$(B, b) \rightarrow (B \setminus (j), b \cup \{(j) \cap \ell, (j) \cap r\}),$$

i.e., the partition element both moves from B to b and is split in its L- and R-loci. This proves §.9).

Step 3 (Probable order of events)

Define $\tau := \frac{(\log T)^2}{r}$ and $T := \min\{t \geq 0 : X_t = \emptyset\}$. We will show that (i) no coalescences, i.e., events (5) , occur in $[T, T]$, (ii) no splits in the beneficial background, i.e., events (5) , occur during $[0, T]$ and (iii) splits in the beneficial background, i.e., events (5) do not overlap with other recombination events (4) with high probability. More precisely, we claim

$$\mathbb{P}[\mathcal{T}_1 \cap [T, T] \neq \emptyset] = \mathcal{O}\left(\frac{1}{(\log T)^2}\right), \tag{5.10}$$

$$\mathbb{P}[\mathcal{T}_5 \cap [0, T] \neq \emptyset] = \mathcal{O}\left(\frac{(\log T)^2}{r}\right), \tag{5.11}$$

$$\mathbb{P}[\min \mathcal{T}_5 < \max(\mathcal{T}_3 \cup \mathcal{T}_4)] = \mathcal{O}\left(\frac{1}{(\log T)^2}\right). \tag{5.12}$$

First, (5.10) coincides with the assertion of Lemma 4.3 [1]. [Second, for (5.11), we have $X_t \leq \frac{(\log T)^2}{r}$ for all $t \leq T$. Hence we get

$$\begin{aligned} \mathbb{P}[\mathcal{T}_5 \cap [0, T] = \emptyset] &= \mathbb{E}\left[\exp\left(-r_{LR} \int_0^T X_s ds\right)\right] \\ &\geq \mathbb{E}\left[\exp(-r_{LR} T)\right] \geq \exp(-r_{LR} \mathbb{E}[T]). \end{aligned}$$

By (3.1) we see that $\mathbb{E}[T] = \frac{2 \log}{\log} + \mathcal{O}(1)$. By the choice of ϵ , this finally gives

$$\mathbb{P}[T_5 \cap [0, T] = \emptyset] \geq 1 - \mathcal{O}\left(\frac{(\log)^2}{\log}\right).$$

Third, for (5.12) we write, using $\epsilon = \mathcal{O}\left(\frac{1}{\log}\right)$, which might change from occurrence to occurrence,

$$\begin{aligned} & \mathbb{P}[\min T_5 < \max(T_3 \cup T_4)] \\ &= \mathbb{E} \left[\int_0^T \mathbb{P}[T_5 \cap [0, t] \neq \emptyset \mid \max(T_3 \cup T_4) \in dt, \mathcal{X}] \cdot \mathbb{P}[\max(T_3 \cup T_4) \in dt \mid \mathcal{X}] \right] \\ &\leq \mathbb{E} \left[\int_0^T \left(1 - \exp\left(-\int_0^t X_s ds\right) \right) \cdot (1 - X_t) \exp\left(-\int_t^T (1 - X_s) ds\right) \right] \\ &\leq 2 \cdot \mathbb{E} \left[\int_0^T (1 - X_t) \int_0^t X_s ds dt \right]. \end{aligned} \tag{5.13}$$

The last term can be estimated using the Green function for the diffusion. As the right hand side of (5.13) coincides with the second line of (4.5) in [10], we immediately obtain (5.12).

In the next three steps we will show that realizing the different splits independently from a fixed sample path $\mathcal{X} = (X_t)_{0 \leq t \leq T}$ will cause only a small error. To see this we will establish a general result on mixed Poisson processes in Step 4 and apply it to the Poisson processes introduced in Step 2. The proof of Proposition 5.11 will then be concluded by an application of these two steps.

Step 4 (General approximations of mixed Poisson processes)

Let $\{(\cdot) : \cdot > 0\}, \{(\cdot) : \cdot > 0\}$ be families of random variables taking values in \mathbb{R}^+ . Assume that the expectations $\mathbb{E}[(\cdot)], \mathbb{E}[(\cdot)]$ are bounded in and

$$\mathbb{V}[(\cdot)], \mathbb{V}[(\cdot)] = \mathcal{O}(\cdot) \tag{5.14}$$

as $\epsilon \rightarrow 0$. Denote the distribution function of the Poisson distribution with parameter λ by $\text{Poi}(\lambda)$. We claim that for $k, l \in \mathbb{N}_0$

$$\mathbb{E}[\text{Poi}(\lambda)(k)] = \text{Poi}_{\mathbb{E}[\lambda]}(k) + \mathcal{O}(\epsilon) \tag{5.15}$$

$$\mathbb{E}[\text{Poi}(\lambda)(k) \cdot \text{Poi}(\lambda)(l)] = \mathbb{E}[\text{Poi}(\lambda)(k)] \cdot \mathbb{E}[\text{Poi}(\lambda)(l)] + \mathcal{O}(\epsilon) \tag{5.16}$$

Note that by a Taylor series approximation, for a random variable \mathbb{R}_+ with second moments and some satisfying $|\tilde{\mu} - \mathbb{E}[\cdot]| \leq |\mu - \mathbb{E}[\cdot]|$,

$$\begin{aligned} & \left| \mathbb{E} \left[e^{-\frac{\cdot}{k!}} \right] - e^{-\mathbb{E}[\cdot]} \frac{\mathbb{E}[\cdot]^k}{k!} \right| \\ &= \frac{1}{2} \left| \mathbb{E} \left[\frac{d^2}{d^2} \left(e^{-\frac{\cdot}{k!}} \right) \Big|_{\tilde{\mu}} \cdot (\mu - \mathbb{E}[\cdot])^2 \right] \right| \\ &\leq \frac{1}{2} \mathbb{E} \left[e^{-\tilde{\mu}} \left| \left\{ \frac{\tilde{\mu}^{k-2}}{(k-2)!} - 2 \frac{\tilde{\mu}^{k-1}}{(k-1)!} + \frac{\tilde{\mu}^k}{k!} \right\} \cdot (\mu - \mathbb{E}[\cdot])^2 \right| \right] \\ &\leq \mathbb{V}[\cdot] \end{aligned} \tag{5.17}$$

where the terms in $\{\dots\}$ only show up if the denominators are non-zero and the last step follows from the fact that the Poisson weights $\{ \dots \}$ lie in $[0, 1]$. As this holds for every (\cdot) , (5.15) follows immediately from 5.14. Moreover, by a calculation similar to (5.17),

$$\mathbb{V}[\text{Poi}(\cdot)(k)] = \mathbb{E} \left[e^{-2(\cdot)} \frac{(\cdot)^{2k}}{(k!)^2} \right] - \mathbb{E} \left[e^{-\cdot} \frac{(\cdot)^k}{k!} \right]^2 = \mathcal{O}(\mathbb{V}[\cdot]) = \mathcal{O}(\cdot).$$

Additionally, (5.16) follows easily from the fact that

$$\begin{aligned} & \left| \mathbb{E} [\text{Poi}(\cdot)(k) \cdot \text{Poi}(\cdot)(l)] - \mathbb{E} [\text{Poi}(\cdot)(k)] \cdot \mathbb{E} [\text{Poi}(\cdot)(l)] \right| \\ &= |\text{Cov}[\text{Poi}(\cdot)(k) \cdot \text{Poi}(\cdot)(l)]| \leq \sqrt{\mathbb{V}[\text{Poi}(\cdot)(k)] \cdot \mathbb{V}[\text{Poi}(\cdot)(l)]} = \mathcal{O}(\cdot) \end{aligned}$$

by the Cauchy–Schwarz inequality.

Step 5 (Green function estimates)

Set $\epsilon = \frac{1}{\log}$ where $\epsilon = \epsilon_{LR}$ for geometry (i) and $\epsilon = \epsilon_{LS} + \epsilon_{SR}$ for geometry (ii). Using our approximations from Step 4 we will show next

$$\mathbb{P} [|\mathcal{I}_5| = k] = \text{Poi}_{\mathbb{E}[r \int_0^T X_s ds]}(k) + \mathcal{O} \left(\frac{1}{(\log)^2} \right) \tag{5.18}$$

$$\begin{aligned} & \mathbb{P} [|(\mathcal{I}_3 \cup \mathcal{I}_4) \cap [T, T] | = k, |\mathcal{I}_5| = l] \\ &= \mathbb{P} [|(\mathcal{I}_3 \cup \mathcal{I}_4) \cap [T, T] | = k] \cdot \mathbb{P} [|\mathcal{I}_5| = l] + \mathcal{O} \left(\frac{1}{(\log)^2} \right) \end{aligned} \tag{5.19}$$

as $\epsilon \rightarrow \infty$. To see this, set $\epsilon = \frac{1}{(\log)^2}$ and define

$$(\cdot) = r \int_0^T X_s ds, \quad (\cdot) = (\cdot + r) \int_T^T (1 - X_s) ds$$

Observe that $k \in \{0, 1, 2, \dots\}$

$$\begin{aligned} \mathbb{P}[|\mathcal{T}_5| = k] &= \mathbb{E}[\text{Poi}(\lambda)(k)] \\ \mathbb{P}[|(\mathcal{T}_3 \cup \mathcal{T}_4) \cap [\mathcal{T}, \mathcal{T}]| = k] &= \mathbb{E}[\text{Poi}(\lambda)(k)] \end{aligned} \tag{5.20}$$

because $\mathcal{T}_3, \mathcal{T}_4, \mathcal{T}_5$ are randomly time-changed Poisson processes (5.9) and (5.10), (5.18) and (5.19) follow once we have shown

$$\mathbb{E} \left[\int_{\mathcal{T}}^{\mathcal{T}} (1 - X_s) ds \right] \leq \mathbb{E} \left[\int_0^{\mathcal{T}} X_s ds \right] \leq 2 + \mathcal{O} \left(\frac{1}{\lambda} \right) \tag{5.21}$$

$$\mathbb{V} \left[\int_{\mathcal{T}}^{\mathcal{T}} (1 - X_s) ds \right] \leq \mathbb{V} \left[\int_0^{\mathcal{T}} X_s ds \right] = \mathcal{O} \left(\frac{1}{(\log \lambda)^2} \right) \tag{5.22}$$

as $\lambda \rightarrow \infty$.

First observe that $(X_t)_{0 \leq t \leq \mathcal{T}}$ has the same distribution as $(1 - X_{\mathcal{T}-t})_{0 \leq t \leq \mathcal{T}}$ by time-reversibility (see e.g. [12]). Hence the inequalities on the left hand side (5.21) and (5.22) follow. Second, we verify the expressions on the right hand side (5.21) and (5.22) by an application of the Green function $G(\cdot, \cdot)$ of the diffusion $(X_t)_{0 \leq t \leq \mathcal{T}}$. This function satisfies

$$\mathbb{E}_x \left[\int_0^{\mathcal{T}} g(X_t) dt \right] = \int_0^1 G(x, y) g(y) dy$$

where $\mathbb{E}_x[\cdot]$ refers to the path $(X_t)_{0 \leq t \leq \mathcal{T}}$ with $X_0 = x$ and $\mathbb{E}[\cdot] := \mathbb{E}_0[\cdot]$. The Green function is given by

$$G(x, y) = \begin{cases} \frac{(1 - e^{-\lambda(1-y)})(1 - e^{-\lambda y})}{y(1-y)(1 - e^{-\lambda})} & \text{if } x \leq y \\ \frac{(e^{-\lambda x} - e^{-\lambda}) (e^{\lambda y} - 1)(1 - e^{-\lambda y})}{y(1-y)(1 - e^{-\lambda})(1 - e^{-\lambda x})} & \text{if } x \geq y, \end{cases}$$

see e.g. [1, 12]. More generally, $G(\cdot, \cdot)$ satisfies

$$\begin{aligned} &\mathbb{E}_x \left[\int_0^{\mathcal{T}} \int_{t_1}^{\mathcal{T}} \dots \int_{t_{k-1}}^{\mathcal{T}} g_k(X_{t_k}) \dots g_1(X_{t_1}) dt_k \dots dt_1 \right] \\ &= \int_0^1 \dots \int_0^1 G(x, x_1) \dots G(x_{k-1}, x_k) g_1(x_1) \dots g_k(x_k) dx_k \dots dx_1 \end{aligned}$$

for all $k = 1, 2, \dots$ which can be proved by induction. We may thus write, because $G(x, y) = G(0, y)$ for $y \geq x$,

$$\begin{aligned} & \mathbb{V} \left[\int_0^T X_s ds \right] \\ &= 2 \left(2 \int_0^1 \int_0^1 G(0, x)G(x, y)xydydx - 2 \int_0^1 \int_x^1 G(0, x)G(0, y)xydydx \right) \\ &= 2^2 \int_0^1 \int_0^x G(0, x)G(x, y)xydydx \leq 2^2 \int_0^1 \int_0^x G(0, x)G(x, y)dydx \\ &= 2^2 \mathbb{V}[T] = \mathcal{O} \left(\frac{1}{(\log)^2} \right) \end{aligned}$$

by (3.1) which gives 5.22.

Step 6 (Approximate independence)

As we have seen in 5(8), the distribution of T^X is determined by the distribution of W from Step 1 and the distribution of the order of events in the Poisson processes $\mathcal{T}_3, \mathcal{T}_4$ and \mathcal{T}_5 . The calculations in Step 3 allow us to make the assumptions

$$\mathcal{T}_1 \cap [T, T] = \emptyset, \quad \mathcal{T}_5 \cap [0, T] = \emptyset, \quad \max(\mathcal{T}_3 \cup \mathcal{T}_4) < \min \mathcal{T}_5$$

on the ordering of events in these Poisson processes as these events have probability $1 - \mathcal{O}((\log)^{-2})$. Furthermore, we know from 5(9) that events in \mathcal{T}_3 and \mathcal{T}_4 may be exchanged without changing the distribution of T^X . Moreover, given $(\mathcal{T}_3 \cup \mathcal{T}_4) \cap [T_e, T] = k$, the distribution of $|\mathcal{T}_3 \cap [T, T]|$ is binomial with parameters $(k + T)$ and k . Hence, the distribution of T^X is determined once the joint distribution of

$$\mathcal{T}_1 \cap [0, T], \quad \mathcal{T}_3 \cap [0, T], \quad \mathcal{T}_4 \cap [0, T], \quad |(\mathcal{T}_3 \cup \mathcal{T}_4) \cap [T, T]|, \quad |\mathcal{T}_5|$$

is known. To approximate the joint distribution of these objects, define

$$\mathcal{T}_i := \mathcal{T}_i \cap [0, T], \quad i = 1, 3, 4 \quad \text{and} \quad K_{3,4} := |(\mathcal{T}_3 \cup \mathcal{T}_4) \cap [T, T]|, \quad K_5 := |\mathcal{T}_5|.$$

We will prove

$$\begin{aligned} & \mathbb{P} \circ (\mathcal{T}_1, \mathcal{T}_3, \mathcal{T}_4, K_{3,4}, K_5)^{-1} \\ &= \mathbb{P} \circ (\mathcal{T}_1, \mathcal{T}_3, \mathcal{T}_4, K_{3,4})^{-1} \otimes \text{Poi}_{\mathbb{E}[r \int_0^T X_s ds]} + \mathcal{O} \left(\frac{1}{(\log)^2} \right) \quad (5.23) \end{aligned}$$

where $\mathbb{P} \circ X^{-1}$ is the image measure of the random variable X under \mathbb{P} and the Landau symbol in this context gives the order in variation distance of the distributions.

Once (5.23) is shown we conclude that \mathbf{K}_5 is approximately independent of all other events. Furthermore, its distribution may be interpreted as the sum of Poisson distributions with parameter $\mathbb{E} \left[\int_0^T X_s ds \right]$. These determine the number of split events on all partition elements $\in r$ with $r \cap \mathbf{K}_5 \neq \emptyset$. A partition element splits, if it is hit by at least one split event. The probability for a split of a partition element is thus given, using (5.18) and (5.20) for $k = 0$, by

$$1 - \exp\left(-\mathbb{E} \left[\int_0^T X_s ds \right]\right) = 1 - \mathbb{E} \left[\exp\left(-\int_0^T X_s ds\right) \right] + \mathcal{O}\left(\frac{1}{(\log \lambda)^2}\right).$$

with $\mathbf{K}_5 = L_R$ for geometry (i) and $\mathbf{K}_5 = L_S + S_R$ for geometry (ii). Observe that \mathbf{K}_5 is determined by the distribution $(\mathcal{T}_1, \mathcal{T}_3, \mathcal{T}_4, K_{3,4})$ if K_5 is known. The random partition \mathbf{K}_5 is determined by the distribution $(\mathcal{T}_1, \mathcal{T}_3, \mathcal{T}_4, K_{3,4})$ independently of K_5 . So, Proposition 5.2 is a consequence of the approximate independence of $(\mathcal{T}_1, \mathcal{T}_3, \mathcal{T}_4, K_{3,4})$ and K_5 given by (5.23).

We write

$$\begin{aligned} & \mathbb{P} \circ (\mathcal{T}_1, \mathcal{T}_3, \mathcal{T}_4, K_{3,4}, K_5)^{-1} \\ &= \int \mathbb{P}_{\mathcal{X}} \circ (\mathcal{T}_1, \mathcal{T}_3, \mathcal{T}_4, K_{3,4}, K_5)^{-1} \mathbb{P}[d\mathcal{X}] \\ &= \int \mathbb{P}(\mathcal{X}_t)_{0 \leq t \leq T} \circ (\mathcal{T}_1, \mathcal{T}_3, \mathcal{T}_4)^{-1} \mathbb{P}[d(\mathcal{X}_t)_{0 \leq t \leq T}] \\ & \quad \otimes \int \mathbb{P}(\mathcal{X}_t)_T \leq t \leq T \circ (K_{3,4}, K_5)^{-1} \mathbb{P}[d(\mathcal{X}_t)_T \leq t \leq T] + \mathcal{O}\left(\frac{1}{(\log \lambda)^2}\right) \end{aligned}$$

where we have used the fact that T is a stopping time and the strong Markov property of the process \mathcal{X} . Note that by (5.11) we may assume $K_5 = |\mathcal{T}_5 \cap [T, T]|$ which gives an error of $\mathcal{O}\left(\frac{1}{(\log \lambda)^2}\right)$ in probability. From Steps 4 and 5 we get

$$\begin{aligned} & \int \mathbb{P}(\mathcal{X}_t)_T \leq t \leq T \circ (K_{3,4}, K_5)^{-1} \mathbb{P}[d(\mathcal{X}_t)_T \leq t \leq T] \\ &= \text{Poi}_{\mathbb{E}[(+r) \int_T^T (1-X_s) ds]} \otimes \text{Poi}_{\mathbb{E}[r \int_T^T X_s ds]} + \mathcal{O}\left(\frac{1}{(\log \lambda)^2}\right) \end{aligned}$$

Rewriting

$$\text{Poi}_{\mathbb{E}[(+r) \int_T^T (1-X_s) ds]} = \int \mathbb{P}(\mathcal{X}_t)_T \leq t \leq T \circ (K_{3,4})^{-1} \mathbb{P}[d(\mathcal{X}_t)_T \leq t \leq T],$$

and using the strong Markov property of a second time we get

$$\begin{aligned}
 & \mathbb{P} \circ \left(\mathcal{T}_1, \mathcal{T}_3, \mathcal{T}_4, \mathbf{K}_{3,4}, \mathbf{K}_5 \right)^{-1} \\
 &= \int \mathbb{P}_{(X_t)_{0 \leq t \leq T}} \circ \left(\mathcal{T}_1, \mathcal{T}_3, \mathcal{T}_4 \right)^{-1} \mathbb{P} [d(X_t)_{0 \leq t \leq T}] \\
 & \quad \otimes \int \mathbb{P}_{(X_t)_T \leq t \leq T} \circ (\mathbf{K}_{3,4})^{-1} \mathbb{P} [d(X_t)_T \leq t \leq T] \\
 & \quad \otimes \text{Poi}_{\mathbb{E}} \left[r \int_0^T X_s ds \right] + \mathcal{O} \left(\frac{1}{(\log)^2} \right) \\
 &= \mathbb{P} \circ \left(\mathcal{T}_1, \mathcal{T}_3, \mathcal{T}_4, \mathbf{K}_{3,4} \right)^{-1} \otimes \text{Poi}_{\mathbb{E}} \left[r \int_0^T X_s ds \right] + \mathcal{O} \left(\frac{1}{(\log)^2} \right)
 \end{aligned}$$

and we are done.

By Proposition 5.2, events (5) can be generated independently of the frequency path and of all other events. The rates of the recombination events (4), (6) at time t are all proportional to $(1 - X_{T-})$. This is reminiscent of the case of only one neutral locus, studied in [3], where a line carrying one neutral locus in recombination distance t recombines to the wild-type background with rate $(1 - X_{T-})$. As a consequence we can use the same techniques used there, especially their Proposition 3.6. which states that a marked Yule tree approximately gives the same partition as the structured coalescent.

Definition 5.3 Define a $\mathcal{P}'_{\ell \cup r}$ -valued random variable γ' as follows: For all partition elements $\ell \in \gamma'$ which $\ell \cap \ell \neq \emptyset, \ell \cap r \neq \emptyset$, i.e., ℓ carries both left and right loci, split the partition element in its left and right loci, $\ell \cap \ell, \ell \cap r$ according to (5.1). Denote the resulting partition by γ' .

Let Y be an infinite Yule tree with branching rate λ . Moreover, consider the random tree $Y_{|\gamma'}$ which arises by sampling $|\gamma'|$ lines from Y at infinity. Identify each of the $|\gamma'|$ partition elements of γ' with one sampled line. Between the root of the Yule tree Y starts and the time it has $|\gamma'|$ lines, mark all lines by the following procedure:

For geometry (i), the tree is marked by Poisson processes with rates λ_L and λ_R . These marks are relabelled such that each branch is hit by at most one mark. Call the corresponding marks SL-, LR- and SLR-marks. The following rules are applied:

- (a) If the Poisson process with rate λ_L puts the first (backward in time) mark at time t from the root, start a Poisson process with rate λ_R and run it for time t . If an event occurs during this time, the branch is marked by an LR-mark, otherwise by an SL-mark.
- (b) If the Poisson process with rate λ_R puts the first (backward in time) mark distinguish the following two cases: if the Poisson process with rate λ_L hits the branch as well, it obtains an SLR-mark. Otherwise, it obtains an LR-mark.

For geometry (ii), mark the tree by two independent Poisson processes with rates λ_{LS} and λ_{SR} . If a branch is hit by one or more events of the Poisson process with rate

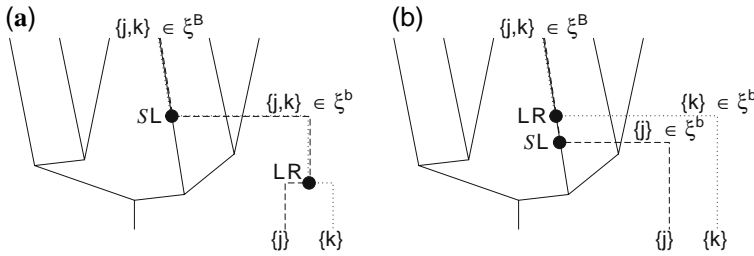


Fig. 6 There are two possibilities how an SL R-mark may occur. Here SL and LR refer to points in the Poisson processes with rates λ_{SL} and λ_{LR} . See text for further explanation

λ_{SL} , it gets an LR S-mark. If it is hit by one or more events with rate λ_{LR} , it additionally gets an SR R-mark.

The result of this procedure is a marked Yule tree \mathcal{Y}_1 . Given ℓ and the marked Yule tree \mathcal{Y}_1 we use the same equivalence relation as given in (3.6) and (3.7) to define $\pi \in \mathcal{P}'_{\ell \cup r}$. Furthermore, we use (3.8) and (3.9) to define the random partition

$$\pi := (\{ \ell \}, \pi \setminus \{ \ell \}).$$

Example 5.4 The two cases in which an SL R-mark occurs for geometry (i) are illustrated in Fig. 6. Consider the line in the sample Yule tree which can be identified with the partition element $\{j, k\}$ where $j \in \ell$ and $k \in r$. Consider case (a) first, shown on the left side of Fig. 6: The SL-mark hitting a branch in \mathcal{Y}_1 leads to a jump of the partition element into the wild-type background. We now have to consider the additional Poisson process at rate λ_{LR} to determine whether or not the line will split within the wild-type background. If an event with rate λ_{LR} occurs, the ℓ - is separated from the R-locus on this line. Case (b) is illustrated on the right side of Fig. 6: Here, the line which refers to the partition element $\{j, k\}$ is first (backward in time) hit by an LR-mark, bringing the R-locus into the wild-type background, and after that an additional SL-mark hits the same branch, which additionally brings the ℓ -locus into the wild-type background. In both cases the ℓ and k end up separated in the wild-type background. This is summarized in Definition 5.3 by an SL R-mark.

As a next step in the proof of Theorem 1 we now show that $\pi \approx \pi'$.

Proposition 5.5 Let $\pi \in \mathcal{P}'_{\ell \cup r}$ and $\pi' \in \mathcal{P}'_{\ell \cup r}$ be as in Definition 5.1 and 5.3. Then,

$$\sup_{\pi \in \mathcal{P}'_{\ell \cup r}} |\mathbb{P}[\pi = \pi] - \mathbb{P}[\pi' = \pi]| = \mathcal{O}\left(\frac{1}{(\log \ell)^2}\right).$$

Proof As the mechanism to generate splits in the beneficial background is the same for both random partitions, π and π' , we concentrate on all other events.

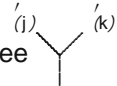
The proof follows along the lines of the Yule approximation in the case of only one neutral locus, given in [3, Definition 3.3. and Section 4.3.]. The crucial observation is that by a random time change $t \rightarrow \tau$ given by $d\tau = (1 - X_t)dt$ the frequency path X , given by (2.1), is taken to the solution $\tilde{X} = (Z_t)_{t \geq 0}$ of

$$dZ = Z \coth(Z)dt + \sqrt{Z}dW \tag{5.24}$$

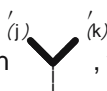
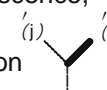
with a standard Brownian motion W and $Z_0 = 0$. This is an α -supercritical Feller branching process conditioned on non-extinction. It was shown in [20] that the genealogy of the α -supercritical branching process is a Yule process with branching rate α . Observe that the time-transformation $t \mapsto \tau(t)$ only works until the supercritical branching process has reached frequency 1. From 4.5(3) we see that at this time the number of lines in the Yule process is Poisson distributed with mean $2/\alpha$ (The additional factor of 2 arises because we made the assumption that the individual offspring variance in the underlying Cannings model is 1 rather than 2. See 2.1.) However, as typical deviations in this Poisson distribution are of the order $\sqrt{2/\alpha}$ we may instead assume that the Yule process has $2/\alpha$ lines. This was made precise in the proof of Proposition 4.7. in [3].



Moreover, for geometries (i) and (ii) the rates in the process change at time τ from $s_L(1 - X_{T-\tau}), l_R(1 - X_{T-\tau})$ to s_L, l_R and from $s_R(1 - X_{T-\tau}), l_L(1 - X_{T-\tau})$ to s_R, l_L , respectively. Especially, the time-changed rates are constant. Under the random time change the coalescence rate (1) changes at time τ from $1/X_{T-\tau}$ to $1/(X_{T-\tau}(1 - X_{T-\tau}))$. However, it was shown in [3, Proposition 4.2.] that the change of these rates can only produce an error in probability of order $O((\log \tau)^{-2})$. This fact was used in [3, Lemma 4.5., Proposition 4.7.] to prove that the marked Yule process gives an accurate approximation in the case of one neutral locus. However, this result carries over to the present situation because all Poisson processes along the Yule process have constant rates.

It remains to check whether the equivalence relation coincides with \sim given the change in the coalescence rate has no effect. First of all, realize the splits in the beneficial background according to Definition 5.1. Then, take $j, k \in \ell \cup r$ and trace their partition elements backwards up to time 0, $t = T$. We only consider geometry (i) and $j \in \ell, k \in r$, since the other cases $j, k \in \ell$ and $j, k \in r$ and all cases for geometry (ii) are similar. If we consider the process from Definition

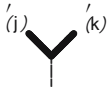
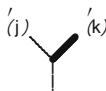
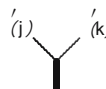
5.1 without any recombination events we would obtain a tree  for the

genealogy relating j and k . However, recombination events may cause the locus j and the R -locus k to end up in different partition element in the random partitions. This will be the case if and only if one of the following events occurs in the process

- (a) a recombination event τ_1 with rate $s_L(1 - X)$ on , which takes either j or k to the wild-type background before coalescence,
- (b) a recombination event τ_2 with rate $l_R(1 - X)$ on , which takes k to the wild-type background before coalescence τ with

- (c) an event (4_i) with rate $\lambda_{LR}(1 - X)$ on  before (backward in time) an event with rate $\lambda_{SL}(1 - X)$ happens on that branch; in this case j and k have coalesced, but a recombination event brings j to the wild-type background without
- (d) an event (3_i) with rate $\lambda_{SL}(1 - X)$ on  before (backward in time) an event with rate $\lambda_{LR}(1 - X)$ happens on that branch, which brings both j and k to the wild-type background. Here, an event (6_i) at rate $\lambda_{LR}(1 - X)$ happens which splits j and k in the wild-type background.

The trees in events (a)–(d) refer to trees generated by the random time change and our assumption that the change in coalescence rate does not alter random partitions we can as well take trees generated by the Yule process and change the rates λ_{SL} and $\lambda_{LR}(1 - X)$ to λ_{SL} and λ_{LR} . Hence we are dealing with a Yule tree with branching rates marked by Poisson processes with rates λ_{SL} and λ_{LR} which is the exact situation of Definition 5.3. Using the definition of the SL-, LR- and SLR-marks, we note that

- (a) produces either an SL- or an SLR-mark on ,
- (b) produces an LR-mark on ,
- (c) and (d) produce either an R- or an SLR-mark on .

If none of these marks occur, j and k are in the same partition element of \mathcal{P} by (3.6). Hence \mathcal{P}_j and \mathcal{P}_k coincide with high probability.

We conclude the proof of Theorem 1 by showing that \mathcal{P}_j from Definition 5.3 and \mathcal{P}_k from Definition 3.3 are close in variation distance.

Proposition 5.6 Let $\mathcal{P} \in \mathcal{P}'_{\ell \cup r}$ and \mathcal{P}_j and \mathcal{P}_k be as in Definition 5.3 and 3.3. Then,

$$\sup_{\mathcal{P} \in \mathcal{P}'_{\ell \cup r}} |\mathbb{P}[\mathcal{P}_j = \mathcal{P}] - \mathbb{P}[\mathcal{P}_k = \mathcal{P}]| = \mathcal{O}\left(\frac{1}{(\log \ell)^2}\right).$$

Proof We will only consider geometry (i). The proof for geometry (ii) is analogous. After realizing the splits in the beneficial background first according to the probabilities given in (5.1) and (3.3), respectively, \mathcal{P}_j and \mathcal{P}_k are determined by the same equivalence relations (3.6) using the marks which hit the tree according to Definition 5.3 and Table 1. Hence our proof consists of two steps. First, we show that the probabilities given in (5.1) and (3.3) differ only by $\mathcal{O}((\log \ell)^{-2})$. Second, we show that

the error caused by generating LR- and SLR-marks using Definitions 5.3 is $\mathcal{O}((\log n)^{-2})$.

Both assertions rely on the same calculation. Assume a line in the Yule tree starts when the full Yule tree has i_1 lines for the last time and ends when the full Yule tree has $i_2 > i_1$ lines for the last time. Additionally, the line is hit by a Poisson process with rate $\lambda = \frac{1}{\log n}$. The probability that the line is not hit by the Poisson process during the time the Yule process has $i_1 < i \leq i_2$, is

$$\frac{i}{i + \lambda}$$

because of competing exponential clocks. Analogously, the probability that the whole line is not hit, is, by a Taylor approximation,

$$\begin{aligned} \prod_{i=i_1+1}^{i_2} \frac{i}{i + \lambda} &= \exp\left(\sum_{i=i_1+1}^{i_2} \log\left(1 - \frac{\lambda}{i + \lambda}\right)\right) \\ &= \exp\left(-\frac{\lambda}{\log n} \sum_{i=i_1+1}^{i_2} \frac{1}{i + \lambda}\right) + \mathcal{O}\left(\frac{\lambda^2}{(\log n)^2}\right) \\ &= \exp\left(-\frac{\lambda}{\log n} \sum_{i=i_1+1}^{i_2} \frac{1}{i}\right) + \mathcal{O}\left(\frac{\lambda^2}{(\log n)^2}\right) = p_{i_1}^{i_2}(\lambda) + \mathcal{O}\left(\frac{\lambda^2}{(\log n)^2}\right), \end{aligned} \tag{5.25}$$

since the neglected terms in the Taylor series are of order $\lambda^2 / (\log n)^2 = \mathcal{O}((\log n)^{-2})$ and higher.

To prove that (5.1) and (3.3) coincide approximately, observe that

$$\mathbb{E} \left[\exp\left(-\lambda \int_0^T X_s ds\right) \right] = \mathbb{E} \left[\exp\left(-\lambda \int_0^T (1 - X_s) ds\right) \right]$$

by the time-reversibility of \mathcal{X} . Additionally, the right hand side gives the probability that a Poisson process with rate $\lambda(1 - X)$ does not hit a line by time T . By the random time change $d\tau = (1 - X_t)dt$ this is approximately the same as the probability that a Poisson process with rate λ does not hit one line in a Yule tree until it has $\lfloor \lambda T \rfloor$ lines and is hence given by $p_0^{\lfloor \lambda T \rfloor}(\lambda)$.

Next, we are considering the generation of LR- and SLR-marks along the Yule tree. The probability that more than one event with rate λ and λ_{LR} hits the Yule tree during the time it has i lines is

$$\frac{\lambda^2}{(i + \lambda)^2} = \mathcal{O}\left(\frac{\lambda^2}{(\log n)^2}\right).$$

Hence we can ignore this event. Together with the Markov property of the Poisson process we see that the marks on different lines in a sample tree may be generated independently once the topology and the total number of lines in the full Yule tree is known.

Consider a branch which starts when the full Yule tree has i_1 lines and ends when it has i_2 lines. Using Definition 5.3 this line is hit by an SL-mark if it is hit by the Poisson process at rate s_L and an independent Poisson process with rate r_L produces no mark between time 0 and the time the Yule tree has i_2 lines. Hence the probability for an SL-mark in τ_{i_1, i_2} is approximately given by

$$\left(1 - \prod_{i=i_1+1}^{i_2} \frac{i}{i + s_L}\right) \left(\prod_{i=1}^{i_2} \frac{i}{i + r_L}\right) = (1 - p_{i_1}^{i_2}(s_L)) p_0^{i_2}(r_L) + \mathcal{O}\left(\frac{1}{(\log \gamma)^2}\right)$$

If a branch is hit by the Poisson process with rate s_L but did not obtain an SL-mark, it obtains an SLR-mark. Hence the probability for such a mark is given by

$$\begin{aligned} &\left(1 - \prod_{i=i_1+1}^{i_2} \frac{i}{i + s_L}\right) \left(1 - \prod_{i=1}^{i_2} \frac{i}{i + r_L}\right) \\ &= \left(1 - p_{i_1}^{i_2}(s_L)\right) \left(1 - p_0^{i_2}(r_L)\right) + \mathcal{O}\left(\frac{1}{(\log \gamma)^2}\right) \end{aligned}$$

The branch is hit by an LR-mark if it is hit by the Poisson process at rate r_L but not by the Poisson process with rate s_L . Hence the probability for an LR-mark is

$$\prod_{i=i_1+1}^{i_2} \frac{i}{i + s_L} \left(1 - \prod_{i=i_1+1}^{i_2} \frac{i}{i + r_L}\right) = p_{i_1}^{i_2}(s_L) \left(1 - p_{i_1}^{i_2}(r_L)\right) + \mathcal{O}\left(\frac{1}{(\log \gamma)^2}\right)$$

As a consequence, the marks \mathcal{Y}_η and \mathcal{Y}_η coincide approximately (cf. Table 1) and we are done.

Acknowledgment We thank Bernhard Haubold, Joachim Hermisson, Etienne Pardoux and Stephanie Leocard for comments on the manuscript and Anton Wakolbinger and Franz Merkl for fruitful discussion. We are grateful to Andy Lehnert, not only for help with Figure 4.

References

1. Barton, N.: The effect of hitch-hiking on neutral genealogies. *Gen.* **72**, 123–133 (1998)
2. International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**(7063), 1299–1320 (2005)
3. Etheridge, A., Pfaffelhuber, P., Wakolbinger, A.: An approximate sampling formula under genetic hitchhiking. *Ann. Appl. Probab.* **16**, 685–729 (2006)
4. Evans, S.N., O’Connell, N.: Weighted occupation time for branching particle systems and a representation for the supercritical superprocess. *Can. Math. Bull.* **37**(2), 187–196 (1994)

5. Ewens W.J.: *Mathematical population genetics. I. theoretical introduction*. 2nd edn. Springer, Heidelberg (2004)
6. Fay, J.C., Wu, C.-I.: Hitchhiking under positive darwinian selection. *Genetics* **155**, 405–413 (2000)
7. Griffiths, R.C., Marjoram, P.: An ancestral recombination graph. In: *Progress in population genetics and human evolution, IMA volumes in mathematics and its applications*, 87, pp. 257–270, Springer, Berlin (1997)
8. Griffiths, R.C.: The frequency spectrum of a mutation and its age, in a general diffusion model. *Theo. Pop. Biol.* **64**(2), 241–251 (2003)
9. Hudson, R.R.: Properties of a neutral allele model with intragenic recombination. *Theo. Pop. Biol.* **23**, 183–201 (1983)
10. Kaplan, N.L., Darden, T., Hudson, R.R.: The coalescent process in models with selection. *Genetics* **120**, 819–829 (1988)
11. Kaplan, N.L., Hudson, R.R., Langley, C.H.: The 'Hitchhiking effect' revisited. *Genetics* **123**, 887–899 (1989)
12. Karlin, S., Taylor, H.M.: *A second course in stochastic processes*. Academic, London (1981)
13. Kim, Y., Stephan, W.: Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* **160**, 765–777 (2002)
14. Lehnert, A., Stephan, W., Pfaffelhuber, P.: A stochastic analysis of linkage disequilibrium under selective sweeps. submitted (2006)
15. Lewontin, R.C.: The interaction of selection and linkage. I. General considerations; Heterotic models. *Genetics* **49**, 49–67 (1964)
16. Li, H., Stephan, W.: Maximum-likelihood methods for detecting recent positive selection and localizing the selected site in a genome. *Genetics* **173**, 377–384 (2005)
17. Maynard Smith, J., Haigh, J.: The hitch-hiking effect of a favorable gene. *Genetics* **23**, 28–35 (1974)
18. Nielsen, R., Williamson, S., Kim, Y., Hubisz, M.J., Clark, A.G., Bustamante, C.: Genomic scans for selective sweeps using SNP data. *Genome Res.* **15**(1), 1566–1575 (2005)
19. Nurminsky, D.: *Selective sweep*. Kluwer, Dordrecht (2005)
20. O'Connell, N.: Yule process approximation for the skeleton of a branching process. *J. Appl. Prob.* **30**, 725–729 (1993)
21. Pfaffelhuber, P., Haubold, B., Wakolbinger, A.: Approximate genealogies under genetic hitchhiking. *Genetics* (2006, in press)
22. Reed, F.A., Tishkoff, S.A.: Positive selection can create false hotspots of recombination. *Genetics* **172**(3), 2011–2014 (2006)
23. Saunders, I.W., Tavaré, S., Watterson, G.A.: On the genealogy of nested subsamples from a haploid population. *Adv. Appl. Probab.* **16**, 471–491 (1984)
24. Schweinsberg, J., Durrett, R.: Random partitions approximating the coalescence of lineages during a selective sweep. *Ann. Appl. Probab.* **15**, 1591–1651 (2005)
25. Stephan, W., Song, Y., Langley, C.: The hitchhiking effect on linkage disequilibrium between linked neutral loci. *Genetics* **172**, 2647–2663 (2006)
26. Stephan, W., Wiehe, T., Lenz, M.: The effect of strongly selected substitutions on neutral polymorphism: analytical results based on diffusion theory. *Theo. Pop. Biol.* **51**, 237–254 (1992)