Abstract We consider the genome of a sample of n individuals taken at the end of a selective sweep, which is the fixation of an advantageous allele in the population. When the selective advantage is high, the genealogy at a locus under selective sweep can be approximated by a comb with n teeth. However, because of recombinations during the selective sweep, the hitchhiking effect decreases as the distance from the selected site increases, so that far from this locus, the tree can be approximated by a Kingman coalescent tree, as in the neutral case. We first give the distribution of the tree at a given locus. Then we focus on the evolution of this tree along the genome. Since this tree-valued process is not Markovian, we study the evolution of the Ancestral Recombination Graph along the genome in case of selective sweep.

Noname manuscript No.
(will be inserted by the editor)

Evolution of the Ancestral Recombination Graph along the genome in case of selective sweep

Stephanie Leocard · Etienne Pardoux

Received: date / Accepted: date

1 Introduction

We study the impact of selection on partially linked neutral genes when an advantageous allele appears in the population and spreads until fixation, which is called a *selective sweep*. The reduction of diversity imposed by the proximity of a site under selection decreases as the distance from the selected site increases because of recombinations that occur between the site under selection and the neutral genes. Described by Maynard-Smith and Haigh in 1974 [9], this phenomenon is called *genetic hitchhiking* and is the topic of many studies [2,11,13,12,14].

We consider a sample of n individuals taken at the end of the selective sweep and we look at the genealogy of the genes located at various distances from the site under selection. When the selective advantage is high, the advantageous allele fixates very quickly in the population and the genealogy at the site under selection can be approximated by a comb with infinitesimal teeth. This is called the star-like approximation. Conversely, far from this site, the effect of selection becomes negligible and we recover the Kingman coalescent tree as in the neutral case. We are interested in the evolution of the tree along the genome, from the comb to the Kingman coalescent tree. However, as in the neutral case [15], this tree-valued process is not Markovian and we have to take into account all the events that occur between the locus under selection and the locus to study, i.e. consider the Ancestral Recombination Graph (ARG) [3,6,4,1].

We define the scaled selective advantage as follows: while the population size is N, $\alpha = Ns$, where s denotes the selective advantage of the favored allele. For our model to make sense, we need to let s tend to zero and N tend to infinity with $\lim_{N\to+\infty} Ns = \alpha$. Here, we are precisely in this limiting situation of an infinite population size. These definitions are consistent with those in [2] and [11].

In this article, we consider the asymptotic model as the scaled selective advantage α tends to infinity, so that the selective sweep occurs instantaneously. We establish how the ARG evolves as the portion of genome whose genealogy is summed up in the ARG increases. We first give a way of coding an ARG, useful to make precise the evolution of

Supported in part by ANR MAEV under contract ANR-06-BLAN-0113.

Address(es) of author(s) should be given

the ARG under neutrality. This neutral evolution was described by Wiuf and Hein [15] but only the algorithm was proposed, without the transition probabilities. Although we are interested in the evolution of the ARG in the case of selection, the neutral case will also be necessary because we have to take into account the recombinations that impact the ARG during the neutral period that predates the beginning of the selective sweep. We then add selection and calculate the probabilities of the different modifications of the ARG following the recombinations that happen during the selective sweep. We finally obtain the global evolution of the ARG in the presence of a selective sweep.

Let us explain our model. Our assumption of an instantaneous sweep, which is essential for our derivations, in particular for making the process R from section 3.2 a Markov process, corresponds in terms of the approximation of the coalescences to the star-like approximation. This approximation has already been described by many authors, together with other more refined ones [10,14]. Also more crude than the Yule approximation proposed by Schweinsberg and Durrett [14], Etheridge *et al.* [2], and Leocard [7], its numerical performance is not much worse for certain values of the parameters. The numerical results obtained with the Yule approximation are also not very good for a sample of size $n > 5$. We collect a few simulations at the end of section 5, to show how well (or poorly) our model approximates the sweep.

For this instantaneous sweep to be consistent with the appearance of recombinations during the sweep, we need to choose a recombination rate γ during the instantaneous sweep, which is different from the rate λ during the neutral period, since it does not play the same role. λ stands for the mean number of recombinations during a time interval of length one, appearing on a portion of the genome of length one, while γ stands for the mean number of recombinations appearing on a portion of the genome of length one, during the sweep. This is explained in details in section 5 at the end of the paper.

The paper is organised as follows. Section 2 explains the assumptions, defines the notations and describes several results previously established and necessary for the next sections. In Section 3, we determine how the number of hichhiking alleles evolves along the genome when the distance from the site under selection increases (Proposition 4) and we describe the coalescent tree at a given locus (Proposition 6). Section 4 presents the evolution of the ARG under neutrality (Corollary 14) and in case of a selective sweep (Corollary 17). We also obtain the non-Markovian process of the coalescent trees along the genome in the presence of a selective sweep (Proposition 19). Finally we discuss the relevance of our approximation, and the meaning of the two recombination parameters, in the final section 5.

2 Notations, assumptions and useful results

2.1 Notations and hypotheses

The population is assumed to be haploid and of infinite size. We consider the genome (or its subregion) as a single chromosome that we identify with the real line $\mathbb{R}$.

We make the hypothesis that one of these genes is under selective sweep: whereas all the individuals initially carry the wild-type allele b, an advantageous mutation replaces b by B in a single individual and the advantageous allele B spreads in the population until fixation after a finite time. We denote by J the original carrier of the mutation

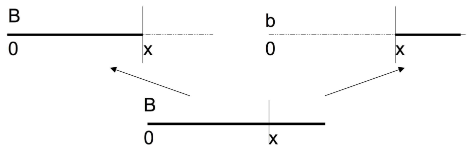


Fig. 1 Event (B$\Rsh$b,x), for $x > 0$. Genes at loci $[x, +\infty)$ originate from an individual carrying b, whereas those at loci $[0, x)$ originate from an individual carrying B.

and by α the scaled selective advantage of B over b. We assume that the location of this gene is known and we use it to initialize the positions of all the genes, denoting by 0 the locus of this gene. We also make the assumptions that no other mutation occurs at this locus and that α is large.

During the selective sweep, along each lineage, recombinations arise on a portion of length dx at rate ρdx, independently of the coalescence events. If the number of recombinations that occur during the selective sweep is negligible, (almost) no variation is expected in the data. If the number of recombination events is excessive, we expect the same evolution as under neutrality. Since the duration of the selective sweep is of the order of $\mathcal{O}\left(\frac{\log\alpha}{\alpha}\right)$ (see [5] or [2], Lemma 3.1.), we choose ρ of the order of $\frac{\alpha}{\log\alpha}$, which is the only way to have a non trivial number of recombinations. More precisely, let $\gamma \in \mathbb{R}_+$ be such that $\rho = \gamma\frac{\alpha}{\log\alpha}$. The parameter γ will be fixed, while α will tend to infinity.

Since the portions of the genome on the left-hand side and on the right-hand side of the site under selection evolve independently and in the same way, we only focus on the right side, which we identify to $\mathbb{R}_+$.

Several kinds of recombinations may occur, according to the position of the recombination and the type of allele at locus 0. For $x > 0$, we define events $(B \Rsh b, x)$ and $(B \Rsh B, x)$ if the individual carries B, $(b \Rsh B, x)$ and $(b \Rsh b, x)$ if it carries b, as follows:

Description of the event $(b_1 \Rsh b_2, x)$, where b_1 and b_2 can be either B or b: For an individual carrying b_1 at locus 0, a recombination happens at locus x and the ancestor for the genes at loci $[x, +\infty)$ (resp. $[0, x)$) carries b_2 (resp. b_1) at locus 0. An example is given in Figure 1.

We consider a sample of n individuals ($n \geq 1$) taken at the end of the sweep. From now on, we reverse time. The end of the selective sweep occurs at time $t = 0$ and the past coalescence and recombination events occur at positive times.

2.2 Results already established

The following results have been established (Proposition 4.1, Corollary 4.2 and Theorem 5.1 in [7]) for an arbitrary number of neutral genes, by extension of the results of [2] and [11]. These results are easily adapted to the continuous version of the genome. Recall that we follow time backward. In other words, an event A happens before an event C whenever its occurrence is closer to the end of the sweep that of C.

Proposition 1 *On any compact subregion $[0, M]$ of the genome $(0 < M < +\infty)$*

1. *The probability of a recombination of type $b \Rsh B$ is $\mathcal{O}\left(\frac{1}{(\log \alpha)^2}\right)$.*
2. *With probability $1 - \mathcal{O}\left(\frac{1}{(\log \alpha)^2}\right)$, all the recombinations of type $B \Rsh B$ happen before those of types $B \Rsh b$ and $b \Rsh b$ and they impact the individuals of the sample independently according to a Poisson process with intensity $\frac{\gamma}{\log(\alpha)} \sum_{\ell=1}^{\lfloor \alpha \rfloor} \frac{1}{\ell}$.*
3. *(Remark 2.6 in [2]) The probability that a recombination of type $B \Rsh b$ happens after the first coalescence event in the sample is $\mathcal{O}\left(\frac{1}{\log \alpha}\right)$.*
4. *We now look at the recombinations of types $B \Rsh b$ or $b \Rsh b$ that impact a given individual before the first coalescence event in the sample.*
   Let U be the number of ancestors of the sample after the recombinations of type $B \Rsh B$. Let F be a random variable definied by its conditional cumulative distribution function, given that $U = u$, as:
   $$\mathbb{P}(F \leq f | U = u) = \frac{(f - (u-1))...(f-1)}{(f + (u-1))...(f+1)}.$$
   With probability $1 - \mathcal{O}\left(\frac{1}{(\log \alpha)^2}\right)$, given that $F = f$, the position of the first recombination (going from locus 0 to $+\infty$), necessarily of type $B \Rsh b$, is exponential with parameter $\frac{\gamma}{\log(\alpha)} \sum_{\ell=f}^{\lfloor \alpha \rfloor} \frac{1}{\ell}$ and the positions of the recombinations of type $b \Rsh b$ at the right of this recombination follow a Poisson process with intensity $\frac{\gamma}{\log(\alpha)} \sum_{\ell=1}^{\lfloor \alpha \rfloor} \frac{1}{\ell}$.

3 Evolution of the number of hitchhiking alleles along the genome

3.1 Asymptotic model when $\alpha \to +\infty$.

Proposition 2 *When $\alpha \to +\infty$:*

1. *There is no recombination of type $b \Rsh B$ and all the recombinations of type $B \Rsh B$ happen before those of type $B \Rsh b$ and $b \Rsh b$.*
2. *There is no recombination of type $B \Rsh b$ after the first coalescence event in the sample.*
3. *The parameters used to model the recombinations of types $B \Rsh B$, $b \Rsh b$ and $B \Rsh b$ tend to γ in quadratic mean.*

Proof From Proposition 1, the probability of a recombination of type $b \Rsh B$ on any compact $[0, M]$ with $M \in \mathbb{N}^*$ is zero when α tends to infinity. Since $\mathbb{R}_+$ is the countable union of the sets $[0, M]$ with $M \in \mathbb{N}^*$, this probability is also null on $\mathbb{R}_+$. The same argument is valid for the probability that a recombination of type $B \Rsh B$ happens after those of type $B \Rsh b$ and $b \Rsh b$ and for the probability that a recombination of type $B \Rsh b$ occurs after the first coalescence event in the sample.
Moreover, the parameter of the Poisson processes that model the recombinations of type $B \Rsh B$ on any compact $[0, M]$ is $\lim_{\alpha \to \infty} \frac{\gamma}{\log \alpha} \sum_{\ell=1}^{\lfloor \alpha \rfloor} \frac{1}{\ell} = \gamma$ so it is also true on $\mathbb{R}_+$.

Before the first coalescence event in the sample, the position of the first recombination of type $B \Rsh b$ at the right of locus 0 on a given individual is exponential with

parameter $\frac{\gamma}{\log\alpha}\sum_{\ell=F}^{\lfloor\alpha\rfloor}\frac{1}{\ell}$.
Let prove that $\lim\limits_{\alpha\to\infty}\mathbb{E}\left(\frac{\gamma}{\log\alpha}\sum_{\ell=F}^{\lfloor\alpha\rfloor}\frac{1}{\ell}\right)=\gamma$ and $\lim\limits_{\alpha\to\infty}Var\left(\frac{\gamma}{\log\alpha}\sum_{\ell=F}^{\lfloor\alpha\rfloor}\frac{1}{\ell}\right)=0$.

$$\begin{aligned}\mathbb{E}\left(\frac{\gamma}{\log\alpha}\sum_{\ell=F}^{\lfloor\alpha\rfloor}\frac{1}{\ell}\right)&=\sum_{f=2}^{\lfloor\alpha\rfloor}\left(\frac{\gamma}{\log\alpha}\sum_{\ell=f}^{\lfloor\alpha\rfloor}\frac{1}{\ell}\right)\mathbb{P}(F=f)\\&=\sum_{f=2}^{\infty}\left(\frac{\gamma}{\log\alpha}\sum_{\ell=f}^{\lfloor\alpha\rfloor}\frac{1}{\ell}\right)1_{\alpha\geq f}\mathbb{P}(F=f)\end{aligned}$$

The terms of the series are non-negative, so we can interchange the order of summation:

$$\begin{aligned}\lim_{\alpha\to+\infty}\mathbb{E}\left(\frac{\gamma}{\log\alpha}\sum_{\ell=F}^{\lfloor\alpha\rfloor}\frac{1}{\ell}\right)&=\sum_{f=2}^{\infty}\lim_{\alpha\to+\infty}\left(\frac{\gamma}{\log\alpha}\sum_{\ell=f}^{\lfloor\alpha\rfloor}\frac{1}{\ell}\right)1_{\alpha\geq f}\mathbb{P}(F=f)\\&=\sum_{f=2}^{\infty}\gamma\,\mathbb{P}(F=f)=\gamma\end{aligned}$$

Moreover,

$$Var\left(\frac{\gamma}{\log\alpha}\sum_{\ell=F}^{\lfloor\alpha\rfloor}\frac{1}{\ell}\right)=\mathbb{E}\left(\left(\frac{\gamma}{\log\alpha}\sum_{\ell=F}^{\lfloor\alpha\rfloor}\frac{1}{\ell}\right)^2\right)-\gamma^2$$

With the same arguments, we obtain $\lim\limits_{\alpha\to+\infty}Var\left(\frac{\gamma}{\log\alpha}\sum_{\ell=F}^{\lfloor\alpha\rfloor}\frac{1}{\ell}\right)=\gamma^2-\gamma^2=0$. The limit of $\frac{\gamma}{\log\alpha}\sum_{\ell=F}^{\lfloor\alpha\rfloor}\frac{1}{\ell}$ in quadratic mean is deterministic and equal to γ.

Finally, given that at least one recombination of type $B \Rsh b$ occurred between loci 0 and $x>0$, the positions of the recombinations of type $b \Rsh b$ after locus x follow a Poisson process with intensity $\lim_{\alpha\to\infty}\frac{\gamma}{\log\alpha}\sum_{\ell=1}^{\lfloor\alpha\rfloor}\frac{1}{\ell}=\gamma$ on any compact $[0,M]$, hence on $\mathbb{R}_+$. □

Since the recombinations of types $B \Rsh b$ and $b \Rsh b$ happen with the same rate and are effective on complementary subsets (i.e. individuals B for $B \Rsh b$, individuals b for $b \Rsh b$), we can merge them into one type $\Rsh b$ with the same rate γ.

Note that during the selective sweep, the sampled individuals are independently impacted by recombinations because no recombination occurs after the first coalescence event in the sample.

The duration of the selective sweep tends to 0 as the selective advantage α goes to infinity, so we approximate the genealogy at the site under selection by a comb with n teeth of infinitesimal length.

Recall that J denotes the individual where the advantageous mutation appeared at the beginning of the sweep.

Proposition 3 *Consider the allele at locus $x > 0$ of an individual taken at the end of the selective sweep. The probability that this allele is inherited from the individual J is* $\exp(-\gamma x)$.

Proof We look at the genealogy going back in time. As explained in Corollary 10 of [7], this individual is first impacted by recombinations of type $B \Rsh B$. Then, each ancestor carries a connected portion of the genetic material of the initial individual. To determine if the allele at locus x is inherited from J, we have to focus on the ancestor that carries the portion containing x. The only recombinations that may now impact this ancestor are of type $\Rsh b$. So the probability that this ancestor inherited its allele at locus x from J is the probability that no recombination of type $\Rsh b$ occured between loci 0 and x on this particular genome, that is $\exp(-\gamma x)$. □

3.2 Distribution of the number of hitchhiking alleles after a selective sweep.

Consider an individual at the end of the selective sweep and suppose that a recombination impacts this individual at position $x > 0$.

First, assume that this individual carries in position $x - \epsilon$ (ϵ chosen small enough so that there is no recombination between the loci $x - \epsilon$ and x) an allele inherited from J. We wish to calculate the probability that, because of this recombination, the allele at locus x is not inherited from J.

There are two ways to obtain this change of ancestor:
- either the recombination is of type $(B \Rsh b, x)$. Then the ancestor at locus x is necessarily other than J (recall that recombinations of type $b \Rsh B$ cannot happen);
- or the recombination is of type $(B \Rsh B, x)$ and the allele at locus x does not come from J, which occurs with probability $1 - \exp(-\gamma x)$, as seen in Proposition 3.

Finally, since the probabilities of the two types of recombinations are the same, the probability that, after that event, the allele at locus x is not inherited from J is $\frac{1}{2}(1 + (1 - \exp(-\gamma x))) = (1 - \frac{1}{2}\exp(-\gamma x))$.

Assume now that the individual carries in position $x - \epsilon$ an allele that does not come from J. Similarly, we wish to calculate the probability that, because of this recombination, the allele at locus x is inherited from J.

With the same arguments, the only possible type of recombination is $(B \Rsh B, x)$ and the probability that the allele at x comes from J is then $\exp(-\gamma x)$. So the probability that, after that event, the allele at position x comes from J is $\frac{1}{2}\exp(-\gamma x)$.

For any distance $x \geq 0$ from the site under selection, we denote by $R(x) \in \{0, ..., n\}$ the number of alleles in the n-sample that are inherited from J. We study the evolution of $R(x)$ as x increases.

The set of the locations of the recombinations impacting the sample is the superposition of $2n$ independent Poisson processes with intensity γ (recombinations of type $B \Rsh B$ and $\Rsh b$ for each of the n individuals), so it is a Poisson process with intensity $2n\gamma$. Since $R(x^-)$ is the number of individuals that inherited a small subregion $[x-\epsilon, x)$ from J, we obtain the following description of the process R:

Proposition 4 *The process $x \in \mathbb{R}_+ \to R(x) \in \{0, ..., n\}$ has the following properties:*

1. $R(0) = n$.
2. *R is a birth and death process.*
3. $\mathbb{P}(R(x + dx) = k - 1 | R(x) = k) = (2 - \exp(-\gamma x))k\gamma dx + o(dx)$,
   $\mathbb{P}(R(x + dx) = k + 1 | R(x) = k) = \exp(-\gamma x)(n - k)\gamma dx + o(dx)$.

In other words, $(R(x), x \geq 0)$ is a non-homogeneous $\{0, 1, ..., n\}$-valued jump Markov process whose jump rates are given as follows:

$$Q_{k,\ell}(x) = \begin{cases} (2 - \exp(-\gamma x))k\gamma & \text{if } \ell = k - 1, \\ \exp(-\gamma x)(n - k)\gamma & \text{if } \ell = k + 1, \\ \quad 0 & \text{if } \ell \notin \{k - 1, k, k + 1\}. \end{cases}$$

Proof of 2. When a recombination occurs, the value of R is modified according to its former value just before the recombination, the type of this recombination and the lineage with which the recombination is made. Since the size of the population is supposed to be infinite, we can assume that all these lineages with which the recombinations happen are different and that their genetic material between the locus 0 and the locus of recombination is not ancestral to the sample. Thus the distributions of $R(x)|R(x^-)$ and of $R(x)|R(y), 0 \leq y < x$ are the same and the process is Markovian. □

Proposition 5 *Almost surely, there exists $x(\omega) > 0$ such that $R(x) = 0$, $\forall x \geq x(\omega)$.*

Proof We can describe the evolution of $(R(x), x \geq 0)$ with the following construction: Let P^1,..., P^n, Q^1,..., Q^n be $2n$ mutually independent Poisson processes, the P^i having the intensity $(2 - e^{-\gamma x})\gamma$ and the Q^j the intensity $e^{-\gamma x}\gamma$.
Let $x > 0$. If $R(x) = k$, we consider $T_1^i < T_2^i < ...$ the jump positions of the process P^i for $1 \leq i \leq k$ and $S_1^j < S_2^j < ...$ the jump positions of the process Q^j for $1 \leq j \leq n - k$ and we look at

$$y = \min\{T_p^i, S_q^j, 1 \leq p, 1 \leq q, 1 \leq i \leq k, 1 \leq j \leq n - k | T_p^i > x, S_q^j > x\}.$$

If this jump is due to a process P^i (resp. Q^j), then $R(y) = R(x) - 1$ (resp. $R(y) = R(x) + 1$).

The mean number of jumps of each process Q^j on $\mathbb{R}_+$ is $\gamma \int_0^{+\infty} e^{-\gamma t} dt = 1$, so there is a finite number of positive jumps for the process R. This means that $\sup_{1 \leq j \leq n, q \geq 1} S_q^j = \bar{x} < \infty$ a.s. Since the process P^1 has infinitely many jumps on the right of $\bar{x}$, the process $R(x)$ hits zero for some $x(\omega) \geq \bar{x}$ and stays there for all $x \geq x(\omega)$. □

3.3 Coalescent tree at a given locus.

Proposition 6 *The coalescent tree at locus x is a tree with n leaves constructed according to the Kingman $(n - R(x) + 1)$-coalescent, one branch ending with a comb of $R(x)$ teeth. This comb gathers the alleles inherited from the individual J.*

Proof Whereas $R(x)$ alleles come from J, the other alleles are inherited from different ancestors, as explained in the proof of Proposition 4. This fact yields the number of leaves and the existence of the comb at the extremity of one of the branches. Then, these branches coalesce under neutrality to get a Kingman coalescent tree. □

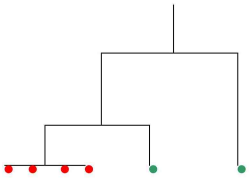


Fig. 2 Example of coalescent tree. Here, $n = 6$ and $R = 4$.

3.4 Evolution of the tree along the genome when a selective sweep occurs.

When a recombination occurs, the recombination point is chosen uniformly on the tree standing for the genealogy at the left of the locus of recombination. Then, the conditional coalescence rate of the recombinant lineage is the number of living lineages in this tree at that time. Thus we obtain the tree at the site of recombination.

As noted in [15] (Figure 6), the process of the coalescent trees along the genome is not Markovian. The fact that the conditional law of the genealogical tree at locus $y > x$, given the genealogical trees at loci $y' \leq x$ differs from the conditional law given the tree at locus x is due to coalescences between lineages with non-overlapping ancestral material. This phenomenon can be seen on Figure 5 (c) below, as explained in the caption of Figure 5. To get a Markov process, we have to work with the Ancestral Recombination Graph (ARG).

We denote by ARG the graph that sums up the genealogy of a sample implied by coalescence and recombination events. It describes a sequence of recombination and coalescence events. It begins at time 0 and goes back in time until a common ancestor appears. This ancestor, called the ultimate ancestor, is common to all the individuals that are present in the genealogy and not just to the genetic material of the sample.

Time goes backwards and to keep the beginning of the sweep at time $t = 0$, this graph is composed of edges of type $[t_i, t_j[$ with $0 \leq t_i < t_j < \infty$. Moreover, because of recombinations, each edge only stands for the history of a region of the genome. The genomic region associated with an edge $[t_i, t_j[$ when t_i is the time of a recombination, is indicated beside this edge. By convention, the ARG will always have a root of length 1.

We denote by M the set of all the ancestral recombination graphs.

The ARG is modified along the genome because of recombinations that occur during the selective sweep and because of those that impact the ARG during the period predating the beginning of the sweep.

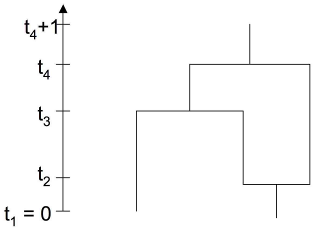


Fig. 3 Example of an ARG. If the position of the recombination is y, we associate to this ARG the set $\mathcal{G} = \Big\{\Big([0,t_2),[0,+\infty)\Big), \Big([0,t_3),[0,+\infty)\Big), \Big([t_3,t_4),[0,+\infty)\Big), \Big([t_2,t_3),[0,y)\Big), \Big([t_2,t_4),[y,+\infty)\Big), \Big([t_4,t_4+1),[0,+\infty)\Big\}$

4 Evolution of the ARG along the genome

4.1 How to code an ARG?

We first consider the neutral case, when the recombination rate along each lineage is $\lambda > 0$.

Let $\mathcal{D} = \{[s,t) \subset \mathbb{R}, 0 \leq s < t \leq +\infty\}$, $\varLambda = \mathcal{D} \times \mathcal{D}$ and $\mathcal{M} = \cup_{n=1}^{\infty} \varLambda^n$.

Consider an ARG, called G. The following coding allows one to indicate the genomic regions associated with the edges: To each edge $A = [t_i, t_j)$ of G, we associate $C = [0,+\infty)$ whenever t_i is a coalescence time or $t_i = 0$; $C = [x,+\infty)$ (resp. $[0,x)$) whenever t_i is the time of a recombination that occurs at position x and the edge describes the genealogy of the sub-region $[x,+\infty)$ (resp. $[0,x)$) of the genome.

Remark 7 Labels only indicate which branch must be followed to establish the genealogy at a given position. They are not the smallest set of positions for which the branch belongs to the corresponding coalescent tree.

We denote by $\mathcal{G} \in \mathcal{M}$ the finite set of pairs $(A,C) \in \varLambda$, where the A's describe all the edges of the G and the C's are given by the coding described above (the order of the pairs is not important). An example is given in Figure 3.

Let $m : M \to \mathcal{M}$ be the mapping that associates to $G \in M$ the element $\mathcal{G} \in \mathcal{M}$ described above.

Proposition 8 *The mapping m is injective.*

Proof We first remark that the probability that several recombination and/or coalescence events occur simultaneously is zero. Indeed, the recombinations happen according to a Poisson process independently of the coalescence events, and the coalescence times are separated by exponential times.

Let $\mathcal{G} \in \mathcal{M}$. We define $\mathcal{G}_1 = \{A \in \mathcal{D}; \exists C \in \mathcal{D}, (A,C) \in \mathcal{G}\}$ and
$\mathcal{G}_2 = \{C \in \mathcal{D}; \exists A \in \mathcal{D}, (A,C) \in \mathcal{G}\}$.

We can reconstruct the ARG G from $\mathcal{G}$ by connecting the edges of type $[t_i, t_j)$ and $[t_j, t_k)$. The labels (elements of $\mathcal{G}_2$) of $[t_i, t_j)$ and $[t_j, t_k)$ indicate which edge we have to follow to obtain the genealogy at a given locus. More precisely:

Step 0. Consider first all the elements of $\mathcal{G}_1$ of the form $[0, t)$, and draw the corresponding vertical segments in increasing order, at a given distance one from the other. The upper ends, the t's, of those segments are the "free ends".

Step 1. Whenever two free ends are at the same level (i. e. the corresponding t's are equal), join them by a horizontal segment (this corresponds to a coalescence event); the two corresponding ends are no longer free; create a new free end on the middle of the coalescence segment.

Step 2. Look for all the still available elements $[t, s)$ of $\mathcal{G}_1$ which are such that t coincides with one of the free ends. Place on the graph those which correspond to a coalescence free end, and the pairs which correspond to some other free end. Those pairs correspond to a recombination; check the corresponding labels to decide which one should be placed on the left, which one on the right.

Step 2 suppresses some free ends, while creating new ones. If there remain available elements of $\mathcal{G}_1$, repeat the above operations by going back to Step 1. □

Remark 9 Since the probability that two events occur simultaneously is zero, the reconstruction is unambiguous and all the internal nodes of the graph that we get are binary.

From now on, we identify each ARG G with $\mathcal{G} = m(G)$.

4.2 Evolution of an ARG under neutrality

We shall describe now the same evolution process as in Wiuf and Hein [15], but with different notations.

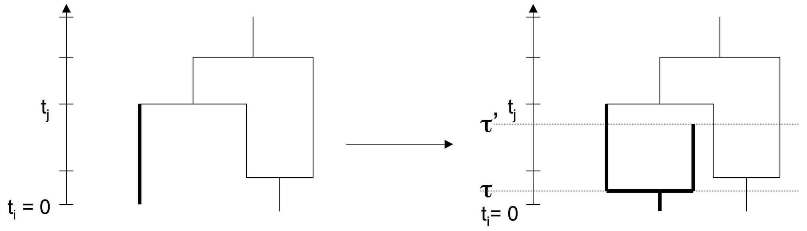
We start with an ARG that describes the genealogy of the genomic region $[0, x)$ and we write $\mathcal{G}(x^-)$ for the associated set in $\mathcal{M}$. We assume that a recombination happens at position x and we write $\mathcal{G}(x)$ for the set in $\mathcal{M}$ associated with the ARG that describes the genealogy on $[0, x]$.

Proposition 10 *There exist* $(T, C) = \Big([t_i, t_j), C\Big) \in \mathcal{G}(x^-)$, $(T', C') = \Big([t_k, t_\ell), C'\Big) \in \mathcal{G}(x^-)$ *and* $\tau, \tau' \in \mathbb{R}_+^*$ *with* $\tau \in T$, $\tau' \in T'$ *and* $\tau < \tau'$, *such that:*
If $T \neq T', \mathcal{G}(x) = \mathcal{G}(x^-) \setminus \{(T, C); (T', C')\} \cup \Big\{\Big([t_i, \tau), C\Big); \Big([\tau, t_j), [0, x)\Big); \Big([\tau, \tau'), [x, +\infty)\Big); \Big([\tau', t_\ell), [0, +\infty)\Big); \Big([t_k, \tau'), C'\Big)\Big\}$.
If $T = T', \mathcal{G}(x) = \mathcal{G}(x^-) \setminus \{(T, C)\} \cup \Big\{\Big([t_i, \tau), C\Big); \Big([\tau, \tau'), [0, x)\Big); \Big([\tau, \tau'), [x, +\infty)\Big); \Big([\tau', t_j), [0, +\infty)\Big)\Big\}$.

Suppose that $\mathcal{G}(x^-)$ is given. As already mentioned in Remark 9, all the elements of $\mathcal{G}_1(x^-)$ are different, so when we choose an edge of $\mathcal{G}(x^-)$, we get its label immediately. We can thus define the $\mathcal{M}$-valued mapping $f_{\mathcal{G}(x^-)}$ such that, for any T, T', τ, τ' such that $\tau \in T, \tau' \in T'$ and $\tau < \tau'$, $f_{\mathcal{G}(x^-)}(T, T', \tau, \tau')$ is the graph $\mathcal{G}(x)$ described above.

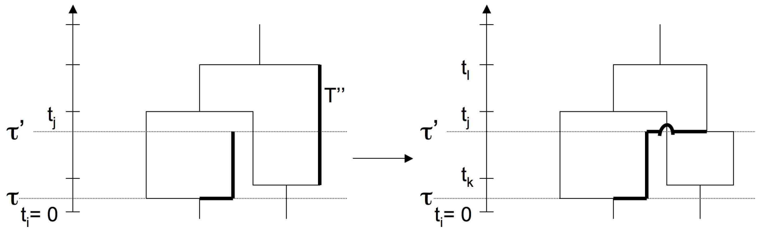
Proof Let τ be the time of the recombination. Let $(T, C) = \Big([t_i, t_j), C\Big)$ be the element of $\mathcal{G}(x^-)$ that describes the edge on which the recombination happens. We have $t_i < \tau < t_j$. This edge disappears from the graph and is replaced by three edges $[t_i, \tau)$, $[\tau, t_j)$ and $[\tau, \tau')$ (τ' will be defined in the next paragraph). $[\tau, t_j)$ (resp. $[\tau, \tau')$) stands for the genealogy for the loci $[0, x)$ (resp. $[x, +\infty)$); we label it with $[0, x)$ (resp. $[x, +\infty)$). Finally, the label associated with $[t_i, \tau)$ is C. We obtain the intermediate graph:

$$\mathcal{G}' = \mathcal{G}(x^-) \setminus \{(T, C)\} \cup \Big\{ \Big([t_i, \tau), C\Big); \Big([\tau, t_j), [0, x)\Big); \Big([\tau, \tau'), [x, +\infty)\Big) \Big\}.$$

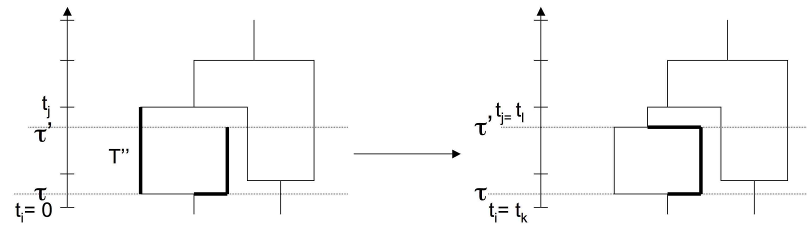


The branch $[\tau, \tau')$ coalesces with an edge of $\mathcal{G}'$. τ' denotes the coalescence time and $(T'', C'') = \Big([t_k, t_\ell), C''\Big)$ the element of $\mathcal{G}'$ which describes the edge with which the branch $[\tau, \tau')$ coalesces. We have $\tau < \tau'$ and $t_k < \tau' < t_\ell$.

If $T'' \neq [\tau, t_j)$, define $(T', C') = (T'', C'')$. Then $T' \in \mathcal{G}_1(x^-)$ and $T' \neq T$. (T', C') disappears from $\mathcal{G}'$ and is replaced by $\Big([\tau', t_\ell), [0, +\infty)\Big)$ and $\Big([t_k, \tau'), C'\Big)$. We then obtain the set $\mathcal{G}(x)$ described in the proposition.



If $T'' = [\tau, t_j)$, define $(T', C') = (T, C)$. $t_i < \tau < \tau' < t_j$. Moreover, $(T'', C'') = \Big([\tau, t_j), [0, x)\Big)$ disappears from $\mathcal{G}'$ and is replaced by $\Big([\tau, \tau'), [0, x)\Big)$, and $\Big([\tau', t_j), [0, +\infty)\Big)$. We then obtain the set $\mathcal{G}(x)$ described in the proposition.



□

Remark 11 The label of the branch above a coalescent event is always $[0,+\infty)$. Indeed, the only branches whose label is different from $[0,+\infty)$ are those that follow a recombination event. Recall that the labels do not indicate the smallest set of positions for which the branch belongs to the corresponding coalescent tree.

Remark 12 A recombination may not imply a modification of genealogy according to the branch it impacts. For example, it is possible that a recombination at position x happens on a branch that only stands for the genealogy of the portion $[0,y)$ of the genome of an individual, with $y<x$. In our model, we authorize this recombination even though its impact on genealogy is null.

Let $x\geq 0$. We denote by $\mathcal{G}(x)$ the ARG that describes the genealogy on $[0,x]$. Let $H(x)$ be its height, that is the waiting time until a common ancestor appears. Recall that this ancestor is common to the genetic material of all the individuals implied in the genealogy.

For any $0\leq t\leq H(x)$, let $A_t(x)$ be the number of lineages in $\mathcal{G}(x)$ at time t. We set $A_t(x)=1$ for $t>H(x)$.

Note that $x\in\mathbb{R}_+\to\mathcal{G}(x)$ is càdlàg, so we can define $\mathcal{G}(x^-)=\lim\limits_{z\to x,z<x}\mathcal{G}(z)$. In the same way, we define $H(x^-)=\lim\limits_{z\to x,z<x}H(z)$ and $A_t(x^-)=\lim\limits_{z\to x,z<x}A_t(z)$.

Proposition 13 *Suppose that $\mathcal{G}(x^-)$ is given and consider a recombination at locus x. Let τ be the time of this recombination; let $T\in\mathcal{G}_1(x^-)$ denote the edge involved in the recombination; let τ' be the coalescence time of the new branch and let $T'\in\mathcal{G}_1(x^-)$ denote the branch chosen to coalesce. We have the following distribution: For any $\mathcal{T}\in\mathcal{G}_1(x^-),\mathcal{T}'\in\mathcal{G}_1(x^-),t\in\mathcal{T},t'\in\mathcal{T}'$ such that $t<t'$,*

$$\begin{aligned}&\mathbb{P}(\tau\leq t,\tau'\leq t',T=\mathcal{T},T'=\mathcal{T}')\\&=\frac{1}{\int_0^{H(x^-)}A_w(x^-)dw}\int_0^t\int_r^{t'}\exp\left(-\int_r^v A_s(x^-)ds\right)dvdr.\end{aligned}$$

Proof We denote by $\mathbb{P}^{(r)}_{\tau'}$ the conditional distribution of τ' given that $\tau = r$.

$$\begin{aligned}
&\mathbb{P}(\tau \leq t, \tau' \leq t', T = \mathcal{T}, T' = \mathcal{T}') \\
&= \int_0^t \mathbb{P}(\tau' \leq t', T' = \mathcal{T}' | \tau = r) \mathbb{P}(T = \mathcal{T} | \tau = r) \mathbb{P}_\tau(dr) \\
&= \int_0^t \int_r^{t'} \mathbb{P}(T' = \mathcal{T}' | \tau' = v) \mathbb{P}^{(r)}_{\tau'}(dv) \times \frac{1}{A_r(x^-)} \times \frac{A_r(x^-)}{\int_0^{H(x^-)} A_w(x^-) dw} dr \\
&= \frac{1}{\int_0^{H(x^-)} A_w(x^-) dw} \int_0^t \int_r^{t'} \frac{1}{A_v(x^-)} A_v(x^-) \exp(-\int_r^v A_s(x^-) ds) dv dr \\
&= \frac{1}{\int_0^{H(x^-)} A_w(x^-) dw} \int_0^t \int_r^{t'} \exp\left(-\int_r^v A_s(x^-) ds\right) dv dr.
\end{aligned}$$

For the first equality, we use the independence of recombination and coalescence events. For the second one, we remark that:
i) The lineage T that recombines at time r is uniformly chosen among the lineages that are present at this moment (their cardinality is $A_r(x^-)$);
ii) the density of τ is $\frac{A_r(x^-)}{\int_0^{H(x^-)} A_w(x^-) dw}$ because the recombination point is uniformly chosen on the ARG.
To get the third equality, we use
i) the fact that the lineage T' that coalesces is chosen uniformly among the lineages present at this moment;
ii) the fact that the distribution of the coalescence time of the lineage is exponential with intensity $A_s(x^-)$. □

Corollary 14 *Under neutrality, the ARG evolves along the DNA sequence in the following way:*
i) The positions of the recombinations are the jump times of a Poisson process with intensity $\lambda \times L(x)$, where $L(x)$ is the total length of the ARG at position x ($L(0) = 1$).
ii) For each jump location x of the process $(\mathcal{G}(y), y \geq 0)$, the conditional law of $\mathcal{G}(x)$ given that $\mathcal{G}(x^-) = G$ is the image by the mapping f_G (introduced just after Proposition 10) of

$$\mathbb{P}_{(\tau,\tau',T,T')}(dt, dt', \mathcal{T}, \mathcal{T}') = \frac{\exp\left(-\int_t^{t'} A_s(x^-) ds\right)}{\int_0^{H(x^-)} A_w(x^-) dw} 1_{t \in \mathcal{T}} 1_{t' \in \mathcal{T}'} 1_{t < t'} dt' dt.$$

4.3 Evolution of the ARG due to a recombination during the selective sweep.

Under neutrality, all the recombinations occur at positive times $\tau > 0$, which does not modify the number of leaves of the ARG. This will not be true anymore when we add a selective sweep whose duration tends to 0. Whenever a recombination happens during the selective sweep, a new branch (or a new tooth of the comb) is created (recall that all the individuals chosen to recombine with are not present on the graph yet).

One of the branches now carries a comb $\mathcal{P}$: each tooth of $\mathcal{P}$ gathers the alleles of one sampled individual, which share the same history during the selective sweep and

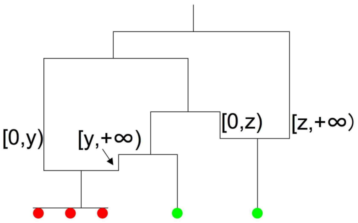


Fig. 4 Example of ARG in case of selective sweep. The edges without indication carry the label $[0, +\infty)$.

which are inherited from the individual J. These alleles correspond to a subregion $[x, y)$ of the genome, where $0 \leq x < y \leq +\infty$. We label each tooth with such an interval to make precise the loci for which the genealogy is obtained starting at this tooth. We denote by $|\mathcal{P}|$ the number of teeth of $\mathcal{P}$.

The ARG is now equivalent to the data of the comb $\mathcal{P}$ and a set $\mathcal{G} \in \mathcal{M}$ where one of the pairs of the form $\Big([0,t), C\Big)$ carries the comb, that we consider as a leaf of the graph.

For $x \geq 0$, denote by $\mathcal{P}(x)$ the comb used to describe the genealogy of the segment $[0, x]$. The evolution of $\mathcal{P}$ along the genome is càdlàg, so we can define $\mathcal{P}(x^-)$. Note that contrary to Section 3.3 where we considered the coalescent tree at a single locus, the sum of the number of leaves of $\mathcal{G}$ other than the comb plus the number of teeth of the comb is not necessarily equal to n anymore. More precisely, $|\mathcal{P}(x)| + |\mathcal{G}(x)| - 1 = n$ + the number of recombinations during the selective sweep, between loci 0 and x. This is related to the fact that in our model, the sweep is instantaneous.

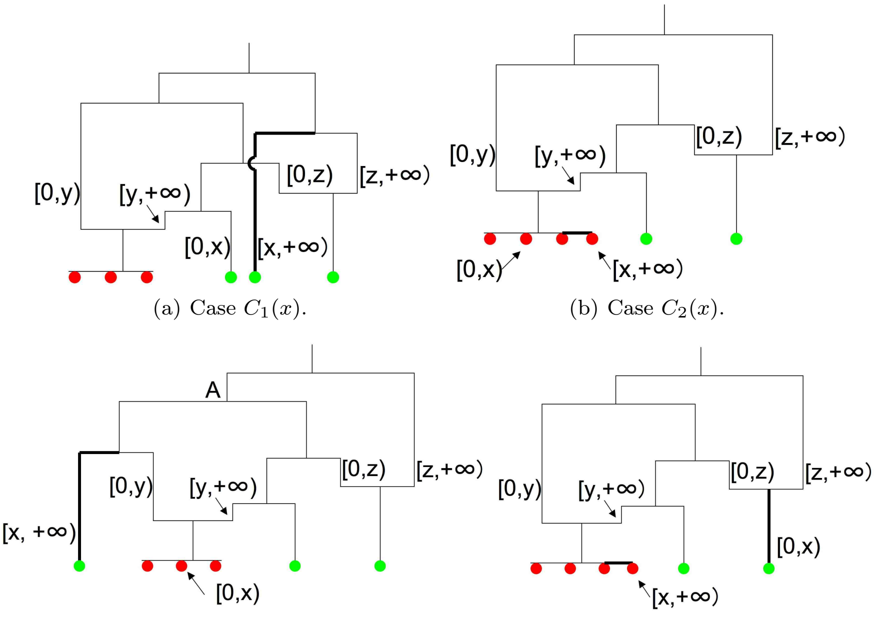
Notations In the next proposition, we will use the following notations:
$\mathcal{I} \in \mathcal{P}(x^-)$ will be a tooth of $\mathcal{P}(x^-)$, identified with its label;
$\mathcal{T} = [0, a)$ will be an edge of $\mathcal{G}_1(x^-)$ attached to a leaf;
$\mathcal{T}'$ and t' will be an edge of $\mathcal{G}_1(x^-)$ and a positive number such that $t' \in \mathcal{T}'$.

Proposition 15 *Suppose that a recombination happens at position $x > 0$ during the selective sweep. Let $\epsilon > 0$ be such that the impacted individual inherited the portion $[x - \epsilon, x)$ from an single ancestor at the beginning of the sweep. The recombination event is of one of the types $C_i(x)$, $1 \leq i \leq 4$.*

- *$C_1(x)$: If $R(x) = R(x^-)$ and the portion $[x - \epsilon, x)$ does not descend from J, then $\mathcal{P}(x) = \mathcal{P}(x^-)$ and there exist $(T, C) = \Big([0, t_i), C\Big) \in \mathcal{G}(x^-)$ a branch of the graph attached to a leaf other than the comb,*
  $(T', C') = \Big([t_k, t_\ell), C'\Big) \in \mathcal{G}(x^-)$ and $\tau' \in T'$ such that

$$\mathcal{G}(x) = \mathcal{G}(x^-) \setminus \{(T, C), (T', C')\}$$
$$\cup \Big\{ \Big(T, C \cap [0, x)\Big); \Big([0, \tau'), [x, +\infty)\Big), \Big([\tau', t_\ell), [0, +\infty)\Big); \Big([t_k, \tau'), C'\Big) \Big\}$$

Moreover, τ' is the first jump of a Poisson process with intensity $A_s(x^-)$. The conditional law of T, T', τ' given that a recombination of type $C_1(x)$ happens at



(a) Case $C_1(x)$.

(b) Case $C_2(x)$.

(c) Case $C_3(x)$.

(d) Case $C_4(x)$.

Fig. 5 Examples for each type of event. $z < y < x$. We can observe on (c) that the process of the coalescent trees along the genome is not Markovian. Indeed, the probability that A is the most recent common ancestor to the recombinant lineage and to the individuals that form the comb is zero if we only consider the tree at the left of x. On the contrary, when we consider the ARG (that is when we add all the "past"), this probability is positive.

position x is specified by:

$$
\begin{aligned}
&\mathbb{P}(T = \mathcal{T}, T' = \mathcal{T}', \tau' \le t' | C_1(x)) \\
&= \frac{1}{A_{0^+}(x^-) - 1} \int_0^{t'} \frac{1}{A_t(x^-)} A_t(x^-) \exp\left(-\int_0^t A_s(x^-) ds\right) dt \\
&= \frac{1}{A_{0^+}(x^-) - 1} \int_0^{t'} \exp\left(-\int_0^t A_s(x^-) ds\right) dt
\end{aligned}
$$

(We note that the number of branches starting at 0 from a leaf other than the comb is $A_{0^+}(x^-) - 1$.)

- *$C_2(x)$: If $R(x) = R(x^-)$ and the portion $[x - \epsilon, x)$ descends from J, then the recombination is of type $B \Rsh B$, $\mathcal{G}(x) = \mathcal{G}(x^-)$ and there exists a tooth $I \in \mathcal{P}(x^-)$ such that $\mathcal{P}(x) = (\mathcal{P}(x^-) \setminus I) \cup (I \cap [0, x)\big) \cup (I \cap [x, \infty)\big)$. Moreover, I is chosen uniformly chosen among the teeth of the comb.*

- *$C_3(x)$: If $R(x) = R(x^-) - 1$, then the involved recombination is either of type $B \Rsh b$ or of type $B \Rsh B$; there exists $I \in \mathcal{P}(x^-)$ such that $\mathcal{P}(x) = \Big(\mathcal{P}(x^-) \setminus I) \cup (I \cap [0, x)\Big)$ and there exist $(T', C') = \Big([t_k, t_\ell), C'\Big) \in \mathcal{G}(x^-)$ and $\tau' \in T'$ such that*

$$
\mathcal{G}(x) = \mathcal{G}(x^-) \setminus (T', C') \cup \Big\{ \Big([0, \tau'), [x, +\infty)\Big); \Big([\tau', t_\ell), [0, +\infty)\Big); \Big([t_k, \tau'), C'\Big) \Big\}.
$$

Moreover, the conditional law of I, T', τ' given that a recombination of type $C_3(x)$ happens at position x is specified by:

$$\mathbb{P}(I = \mathcal{I}, T' = \mathcal{T}', \tau' \leq t' | C_3(x)) = \frac{1}{|\mathcal{P}(x^-)|} \int_0^{t'} \exp\left(-\int_0^t A_s(x^-)ds\right) dt$$

– *$C_4(x)$: If $R(x) = R(x^-) + 1$, then the involved recombination is of type $B \Rsh B$ and there exists a branch $(T, C) = \Big([0, a), C\Big) \in \mathcal{G}(x^-)$ not attached to the comb such that $\mathcal{G}(x) = \mathcal{G}(x^-) \setminus (T, C) \cup (T, C \cap [0, x))$ and $\mathcal{P}(x) = \mathcal{P}(x^-) \cup [x, \infty)$. Moreover, T is chosen uniformly among the branches attached to one of the leaves of the graph other than the comb.*

Proof The condition $R(x) = R(x^-)$ implies that both of the portion $[x - \epsilon, x)$ carried by the impacted individual and the portion $[x, +\infty)$ carried by the other recombinant come from J ($C_2(x)$) or do not come from J ($C_1(x)$).

Whenever a recombination is of type $\Rsh b$ at locus x, the portion $[x, \infty)$ inherited from the recombinant always comes from an individual carrying b at the beginning of the sweep. Otherwise, a recombination $b \Rsh B$ would be necessary, but this event cannot happen (Proposition 2). This is why in $C_2(x)$ the recombination is necessarily of type $B \Rsh B$.

With the same arguments, if $R(x) = R(x^-) + 1$ ($C_4(x)$) (resp. $R(x^-) - 1$ ($C_3(x)$)), then the individual of the sample impacted by the recombination inherited the portion $[x - \epsilon, x)$ from an individual carrying b (resp. B) at the beginning of the sweep, whereas the portion $[x, +\infty)$ given by the recombinant comes (resp. does not come) from individual J. □

For $1 \leq i \leq 4$, denote by $\nu_i(x)$ the conditional probability that the recombination is of type $C_i(x)$, given that a recombination happens during the selective sweep at locus x.

Lemma 16

$$\begin{aligned}
\nu_1(x) &= \left(1 - \frac{R(x^-)}{n}\right)\left(1 - \frac{\exp(-\gamma x)}{2}\right) \\
\nu_2(x) &= \frac{R(x^-)}{n} \frac{1}{2} \exp(-\gamma x) \\
\nu_3(x) &= \frac{R(x^-)}{n} \left(1 - \frac{\exp(-\gamma x)}{2}\right) \\
\nu_4(x) &= \left(1 - \frac{R(x^-)}{n}\right) \frac{1}{2} \exp(-\gamma x)
\end{aligned}$$

Proof Let A be the event "The allele in position $x - \epsilon$ (ϵ chosen small enough so that there is no recombination between the loci $x - \epsilon$ and x) carried by the impacted individual is inherited from the individual J". $\mathbb{P}(A) = \frac{R(x^-)}{n}$.
All the following probabilities are easily deduced from Section 3.2.

$$\begin{aligned}
\nu_1(x) &= \mathbb{P}(\{R(x) = R(x^-)\} \cap A^c) = \mathbb{P}(R(x) = R(x^-)|A^c)\mathbb{P}(A^c), \\
\nu_2(x) &= \mathbb{P}(\{R(x) = R(x^-)\} \cap A) = \mathbb{P}(R(x) = R(x^-)|A)\mathbb{P}(A), \\
\nu_3(x) &= \mathbb{P}(R(x) = R(x^-) - 1 | \text{a recombination happens at locus}\, x), \\
\nu_4(x) &= \mathbb{P}(R(x) = R(x^-) + 1 | \text{a recombination happens at locus}\, x),
\end{aligned}$$

where A^c is the complementary set of A. □

Corollary 17 *The process $x \to (\mathcal{P}(x), \mathcal{G}(x))$ evolves in the following way:*

1. *$\mathcal{P}(0) = \mathbb{R}_+^n$ and $\mathcal{G}(0) = \Big([0,1), [0,+\infty)\Big)$ (by convention).*

2. *The recombination positions follow a Poisson process with parameter $2n\gamma + \lambda L(x)$, where $L(x)$ is the total length of $\mathcal{G}(x)$.*

3. *At the position x of a jump:*
   - *With probability $\frac{\lambda L(x^-)}{2n\gamma+\lambda L(x^-)}$, the recombination predates the beginning of the selective sweep. Then $\mathcal{P}(x) = \mathcal{P}(x^-)$ and the law of $\mathcal{G}(x)$ is the image by the mapping $f_{\mathcal{G}(x^-)}$ of*

$$\mathbb{P}_{(\tau,\tau',T,T')}(dt,dt',\mathcal{T},\mathcal{T}') = \frac{\exp\left(-\int_t^{t'} A_s(x^-)ds\right)}{\int_0^{H(x^-)} A_w(x^-)dw} 1_{t\in\mathcal{T}} 1_{t'\in\mathcal{T}'} 1_{t<t'} dt' dt.$$

   - *With probability $\frac{2n\gamma\times\nu_1(x)}{2n\gamma+\lambda L(x^-)}$, we obtain $\mathcal{P}(x) = \mathcal{P}(x^-)$ and $\mathcal{G}(x)$ of the form $\mathcal{G}(x^-) \setminus \{(\mathcal{T}, C), (\mathcal{T}', C')\} \cup \Big\{\Big(\mathcal{T}, C \cap [0,x)\Big); \Big([0,\tau'), [x,+\infty)\Big),$ $\Big([\tau', t_\ell), [0,+\infty)\Big); \Big([t_k,\tau'), C'\Big)\Big\}$ with $\mathcal{T}' = [t_k, t_\ell)$ and*

$$\mathbb{P}_{(\tau',T,T')}(dt',\mathcal{T},\mathcal{T}') = \frac{\exp\left(-\int_0^{t'} A_s(x^-)ds\right)}{A_{0^+}(x^-)-1} 1_{t'\in\mathcal{T}'} dt'.$$

   - *With probability $\frac{2n\gamma\times\nu_2(x)}{2n\gamma+\lambda L(x^-)}$, we obtain $\mathcal{G}(x) = \mathcal{G}(x^-)$ and a comb $\mathcal{P}(x)$ of the form $(\mathcal{P}(x^-) \setminus \mathcal{I}) \cup (\mathcal{I} \cap [0,x)) \cup (\mathcal{I} \cap [x,\infty))$ where the tooth $\mathcal{I}$ is chosen uniformly on the comb $\mathcal{P}(x^-)$.*

   - *With probability $\frac{2n\gamma\times\nu_3(x)}{2n\gamma+\lambda L(x^-)}$, we obtain $\mathcal{P}(x)$ of the form $(\mathcal{P}(x^-) \setminus \mathcal{I}) \cup (\mathcal{I} \cap [0,x)\Big)$ and $\mathcal{G}(x)$ of the form $\mathcal{G}(x^-) \setminus (\mathcal{T}', C') \cup \Big\{\Big([0,\tau'), [x,+\infty)\Big);$ $\Big([\tau', t_\ell), [0,+\infty)\Big); \Big([t_k,\tau'), C'\Big)\Big\}$ with $\mathcal{T}' = [t_k, t_\ell)$ and*

$$\mathbb{P}_{(I,\tau',T')}(\mathcal{I}, dt', \mathcal{T}') = \frac{1}{|\mathcal{P}(x^-)|} \exp\left(-\int_0^{t'} A_s(x^-)ds\right) 1_{t'\in\mathcal{T}'} dt'.$$

   - *With probability $\frac{2n\gamma\times\nu_4(x)}{2n\gamma+\lambda L(x^-)}$, we obtain $\mathcal{P}(x) = \mathcal{P}(x^-) \cup [x,\infty)$ and $\mathcal{G}(x)$ of the form $\mathcal{G}(x^-) \setminus (\mathcal{T}, C) \cup (\mathcal{T}, C \cap [0,x))$ where $\mathcal{T}$ is chosen uniformly among the $A_{0^+}(x^-) - 1$ branches attached to one of the leaves of the graph other than the comb.*

Remark 18 In this paper, the sample is drawn at the end of the selective sweep. We could easily extend these results to a sample drawn later, after a neutral period has followed the sweep. Indeed, we only have to extend the neutral recombination process to the whole graph. The influence of selection would decrease, even close to the site under selection, because of recombinations and new mutations happening after the end of the sweep.

4.4 Evolution of the coalescent tree along the genome in the presence of a selective sweep

As mentioned in Section 3.4, the evolution of the coalescent tree along the genome is not Markovian. Nevertheless, the positions of the recombinations along the genome constitute a Poisson process and given the tree at the left of a recombination point, we can obtain the conditional law of the tree at the right of this point. The neutral case was described in the first paragraph of Section 3.4. We now study the conditional distribution when a selective sweep occurs.

Denote by $\tilde{L}(x)$ the total length of the coalescent tree at position $x \geq 0$ and by $\tilde{A}_t(x)$ the number of living lineages in the tree at time $t \geq 0$.

If the recombination occurs during the neutral period, the recombination point is chosen uniformly on the tree. The impacted branch is replaced by a new one that will coalesce with another branch of the tree. The coalescence rate at time t is $\tilde{A}_t(x)$ and the branch chosen to coalesce is drawn uniformly among the existing ones.

If the recombination occurs during the selective sweep, we can deduce the evolution of the tree for each of the four cases from Proposition 15.

$C_1(x)$: A branch ending by a leaf other than the comb is suppressed. A new branch starts from 0 and coalesces at rate $\tilde{A}_t(x)$ with a branch chosen uniformly among the existing ones at the coalescence time.

$C_2(x)$: The coalescent tree is unchanged.

$C_3(x)$: One tooth of the comb is suppressed and a new branch starts from 0 and coalesces at rate $\tilde{A}_t(x)$ with a branch chosen uniformly among the existing ones at the coalescence time.

$C_4(x)$: A branch ending by a leaf other that the comb is suppressed and the tooth is added to the comb.

For $x \leq 0$, let $\nu(x) = \nu_1(x) + \nu_3(x) + \nu_4(x) = 1 - \frac{R(x^-)}{n}\frac{1}{2}\exp(-\gamma x)$.
We obtain the following evolution:

Proposition 19 *The positions where the tree is modified follow a Poisson process with intensity $2n\gamma\nu(x) + \lambda\tilde{L}(x)$. At the position x of a modification:*

- *With probability $\frac{\lambda\tilde{L}(x^-)}{2n\gamma\nu(x)+\lambda\tilde{L}(x^-)}$, the recombination predates the beginning of the selective sweep. The recombination point is chosen uniformly on the tree. The impacted branch is replaced by a new one that will coalesce with another branch of the tree. The coalescence rate at time t is $\tilde{A}_t(x)$ and the branch that coalesces is chosen uniformly among the existing ones at the coalescence time.*

- *With probability $\frac{2n\gamma\times\nu_1(x)}{2n\gamma\nu(x)+\lambda\tilde{L}(x^-)}$, a branch ending by a leaf other than the comb is chosen uniformly and is suppressed. Then a new branch starts from 0 and coalesces at rate $\tilde{A}_t(x)$ with a branch chosen uniformly among the existing ones at the coalescence time.*

– *With probability* $\frac{2n\gamma\times\nu_3(x)}{2n\gamma\nu(x)+\lambda\tilde{L}(x^-)}$*, one tooth of the comb is chosen uniformly and is suppressed. Then a new branch starts from 0 and coalesces at rate* $\tilde{A}_t(x)$ *with a branch chosen uniformly among the existing ones at the coalescence time.*

– *With probability* $\frac{2n\gamma\times\nu_4(x)}{2n\gamma\nu(x)+\lambda\tilde{L}(x^-)}$*, a branch ending by a leaf other than the comb is chosen uniformly and is suppressed. Moreover, one tooth is added to the comb.*

Remark 20 Compared to the ARG-valued process along the genome, some jumps have been suppressed to obtain the process of coalescent trees. Indeed, recombinations of type C_2 do not modify the tree and those impacting a branch of the ARG that does not belong to the coalescent tree are not taken into account. Thus the coupling of these two jump processes is easy to deduce.

5 Discussion

In this final section, we want to discuss the relevance of our results, and how they should be understood by Biologists. More precisely, our results are based of an assumption of infinite population (which of course should be understood in practise as a situation of a large popultion), and an instantaneous sweep, which is of course unrealistic. What can we say to the Biologists about that ? Also, our results are expressed in term of two recombination parameters, λ and γ. Why two distinct parameters ? Which values should be given to them in a specific situation ? We first discuss the second question.

5.1 The two parameters λ and γ

These two parameters are quite different, and should a priori be given two distinct values. We have introduced γ in section 2.1, from another parameter ρ. But we do not need to use this third parameter in order to understand how to interpret our two parameters.

During the neutral period, going back in time, recombinations happen at rate λ. What does it mean ? Consider first a single individual. Recombinations can possibly hit that individual, according to the following law. A recombination will hit one particular locus of his genome (which has been identified with the real line), at a particular time. The assumption that this happens at rate λ means exactly that recombinations hit the finite portion $[x, y]$ of the genome during a period of time $[s, t]$ according to a Poisson process of rate λ, i. e. the number of recombinations in disjoint subsets of $[s, t] \times [x, y]$ are mutually independent, and the number recombinations which hit $[z, z']$ during the time interval $[r, r']$ follows a Poisson distribution with parameter $\lambda|z' - z| \times |r' - r|$. There are two possible descriptions of these recombinations hitting $[x, y]$ during the time–interval $[s, t]$ in terms of one–dimensional Poisson process. One is to say that if we follow the genome from x to y, we meet points of recombinations according to a Poisson process of rate $\lambda(t - s)$, and each of those reombinations happens at a time which is chosen uniformly in $[s, t]$, independently of the others. The other one is to say that when following the time evolution from s to t, we encounter recombination times according to a Poisson process of rate $\lambda(y - x)$, and each of those recombinations is placed at a locus chosen uniformly on $[x, y]$, independently of the others. Note that the recombinations hit all lineages (or individuals) alive during a given time interval in an "i. i. d." manner (i. e. recombination hit various individuals independently and at the same rate), and that the addition of mutually independent Poisson processes yields a Poisson process with a rate which sums up the rates of each of those Poisson processes.

This is consistent with the formulation of Corollary 14 i), which says that the position of the jumps locations follow a Poisson process with rate $\lambda L(x)$, if $L(x)$ denotes the length of the ARG at location x.

Let us now explain what is the meaning of the parameter γ, concerning the rate of recombinations during the sweep. In reality, recombinations happen during the sweep, essentially similarly as during the neutral period. However, since we have wrongly assumed that the sweep is instantaneous, we need to model the recombinations during the sweep differently from the neutral period. Above we had a Poison process in two dimensions. Here there is no time interval, so we model the recombinations according to a one–dimensional Poisson process along the genome, with the intensity γ. This γ in itself has no biological meaning. In our model, it is the rate at which each of two types of recombinations happen along the genome of each individual in the sample during the sweep, which is a time interval of length 0. If one thinks that the rcombination rate is the same during the sweep and during the neutral period, it would then be natural to set $2\gamma = \lambda \mathit{\Delta}$, if $\mathit{\Delta}$ denotes the actual duration of the sweep (the factor 2 comes from the fact that we sum up two types of recombinations). However that would mean that γ is small with $\mathit{\Delta}$, which in fact contradicts the basic assumptions of this paper.

5.2 How does our model fit with reality ?

Our "instantaneous sweep" does not exclude recombinations during the sweep, as we explained it in the previous subsection. However, our approximation excludes coalescences during the sweep. In particular, the individuals carrying the advantageous allele (those who have not escaped the sweep as an effect of recombinations) all coalesce, i. e. we have a starlike tree, a "comb" at the time of the sweep. Consequently the approximation implied by our model, at least as far as the individuals which originate from the initial bearer of the advantageous allele are concerned, is exactly the approximation of the coalescent tree during the sweep by a starlike coalescence. This approximation is well known. It is rather crude, but has the advantage of making the computations easy. In order to give an idea of how well (or poorly) it approximates reality, we report below results of several simulations.

Consider the joint law of the length of the hitch-hiked set of neutral genes located on one side of the locus where the advantageous allele has appeared, in n individuals of a sample taken at the end of the sweep. How well that joint law is approximated, that is what we have chosen as criterion for the quality of our approximation. We compute an approximation of the "true" joint law via a Monte Carlo method, with 4×10^4 simulations. We simulated recombinations and coalescence during the sweep. The two types of coalescence are simulated in non–constant populations, the proportion of individuals carrying the advantageous allele evolving according to the Wright–Fisher diffusion conditioned upon fixation of the advantageous allele, i. e. upon eventual completion of the sweep. The diffusion was simulated with a time step of 10^{-5}.

$(h_1,\dots,h_n)$	Simulations	EPW approx.	Our approx.
$(2,2,2)$	0.3963	0.4414	0.3198
$(2,2,3)$	0.2688	0.3108	0.2187
$(3,3,3)$	0.1558	0.1609	0.1023
$(3,3,4)$	0.1052	0.0954	0.0699
$(2,4,4)$	0.1011	0.0941	0.0699
$(2,2,2,2,2)$	0.2626	0.2735	0.1496
$(2,2,2,2,3)$	0.1793	0.1662	0.1023
$(2,2,2,2,4)$	0.1227	0.0795	0.0699
$(2,2,2,3,3)$	0.1349	0.0956	0.0699
$(2,2,2,3,4)$	0.0922	0.0222	0.0478
$(1,1,1,1,1,1,1,2)$	0.7180	0.6095	0.6838
$(1,1,1,1,1,2,2,2)$	0.4391	0.2639	0.3198
$(1,1,1,1,2,2,3,3)$	0.1959	0	0.1023
$(1,1,1,1,2,3,3,3)$	0.1603	0	0.0699
$(1,1,1,1,3,3,3,3)$	0.1257	0	0.0478

Table 1 $\alpha = 2000$, $\gamma = 0.38$

$(h_1,\dots,h_n)$	Simulations	EPW approx.	Our approx.
$(2,2,2)$	0.4372	0.4868	0.3198
$(2,2,3)$	0.3018	0.3585	0.2187
$(3,3,3)$	0.1728	0.2030	0.1023
$(3,3,4)$	0.1057	0.1290	0.0699
$(2,4,4)$	0.0965	0.1269	0.0699
$(2,2,2,2,2)$	0.3107	0.3174	0.1496
$(2,2,2,2,3)$	0.2050	0.2018	0.1023
$(2,2,2,2,4)$	0.1380	0.1031	0.0699
$(2,2,2,3,3)$	0.1578	0.1230	0.0699
$(2,2,2,3,4)$	0.1023	0.0350	0.0478

Table 2 $\alpha = 1000$, $\gamma = 0.36$

The same law in our approximation has the well–known explicit formula

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

Finally, we can compute the same law in the approximate model of Etheridge, Pfaffelhuber, Wakolbinger [2], using the results of the first author [7]. This approximation (called here "EPW approximation") is of order $[\log(\alpha)]^{-2}$, while our approximation is of order $[\log(\alpha)]^{-1}$. Note that the formulas for probabilities in [7] can take negative values, whenever the value of that probability is smaller than $[\log(\alpha)]^{-2}$. In those cases, we put 0 as approximate value of that probability.

We note that our approximation produces in most cases results which are worse than that of the EPW approximation, as expected. However, our results are not always much worse, and when the EPW approximation is very poor (this happens when $n > 5$ and the true probability is small), the star–like approximation ("our approximation") is at least slightly better.

Acknowledgements The authors would like to thank the anonymous Referees for their useful remarks, and their insistence that the paper should be made as much as possible accessible to a wide audience. We also thank Franọis Ezanno for his help with the numerical simulations.

References

1. Durrett R (2008) Probability Models for DNA Sequence Evolution. 2nd ed. Springer
2. Etheridge A, Pfaffelhuber P, Wakolbinger A (2006) An approximate sampling formula under genetic hitchhiking. Ann Appl Prob 16:685-729.
3. Griffiths RC (1981) Neutral two-locus multiple allele models with recombination. Theor Popul Biol 19:169-186.
4. Griffiths RC and Marjoram P (1997) An ancestral recombination graph. In Progress in PopulationGenetics and Human Evolution, P. Donnelly ad S. Tavaré, eds., IMA Volumes in Mathematics and its Applications, Springer, New-York, pp. 257-270.
5. van Herwaarden O and van der Wal N (2002) Extinction time and age of an allele in a large finite population. Theor Pop Biol 61:311-318.
6. Hudson RR (1983) Properties of the neutral model with intragenic recombination. Theor Pop Biol 23.2:213-201.
7. Leocard S (2009) Selective sweep and the size of the hitchhiking set, to appear in Adv. in Applied Probab. 41:731-764.
8. Leocard S (2009) Modèles probabilistes du balayage sélectif et auto-stop génétique. Dissertation, Aix-Marseille Université. http://www.latp.univ-mrs.fr/~leocard/these.pdf
9. Maynard Smith J, Haigh J (1974) The hitch-hiking effect of a favorable gene. Gen Res 23:23-35.
10. Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG and Bustamante C (2005) Genomic scans for selective sweeps using SNP data. Genome res. 15(11):1566-1575.
11. Pfaffelhuber P, Studeny A (2007) Approximating genealogies for partially linked neutral loci under selective sweep. J Math Biol 55:299-330.
12. Stephan W, Song YS, Langley CH (2006) The hitchhiking effect on linkage disequilibrium between linked loci. Genetics 172:2647-2663.
13. McVean G (2007) The Structure of Linkage Disequilibrium Around a Selective Sweep. Genetics 175:1395-1406.
14. Schweinsberg J, Durrett R (2005) Random partitions approximating the coalescence of lineages during a selective sweep, Ann Appl Probab 15:1591-1651.
15. Wiuf C, Hein J (1999) Recombination as a Point Process along Sequences. Theor Popul Biol 55:248-259.