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

# Single-crossover recombination and ancestral recombination trees

**Ellen Baake · Ute von Wangenheim**

Received: date / Accepted: date

**Abstract** We consider the Wright-Fisher model for a population of $N$ individuals, each identified with a sequence of a finite number of sites, and single-crossover recombination between them. We trace back the ancestry of single individuals from the present population. In the $N \to \infty$ limit without rescaling of parameters or time, this ancestral process is described by a random tree, whose branching events correspond to the splitting of the sequence due to recombination. With the help of a decomposition of the trees into subtrees, we calculate the probabilities of the topologies of the ancestral trees. At the same time, these probabilities lead to a semi-explicit solution of the deterministic single-crossover equation. The latter is a discrete-time dynamical system that emerges from the Wright-Fisher model via a law of large numbers and has been waiting for a solution for many decades.

**Keywords** population genetics · recombination · segmentation process · ancestral trees · subtree decomposition

**Mathematics Subject Classification (2000)** MSC 92D10 · MSC 60J28

## 1 Introduction

Recombination happens during sexual reproduction and refers to the combination of the genetic material of two parents into the 'mixed' type of an offspring individual. More precisely, the recombined offspring results from a reciprocal exchange of maternal and paternal gene sequences via so-called *crossovers*. Due to the interaction of individuals and due to dependencies between the positions at which recombination may take place, the process is difficult to handle. This applies even to the limit of infinite population size, where a law of large numbers turns the dynamics of gene frequencies

Ellen Baake
Faculty of Technology
Bielefeld University
33594 Bielefeld
Tel.: +49-521-1064896
Fax: +49-521-1066411
E-mail: ebaake@TechFak.Uni-Bielefeld.DE

Ute von Wangenheim
Faculty of Technology
Bielefeld University
33594 Bielefeld
E-mail: uvonwang@TechFak.Uni-Bielefeld.DE

into a deterministic, nonlinear system of difference or differential equations, which has challenged population geneticists since its first formulation by Geiringer in 1944; see Bennett (1954); McHale and Ringwood (1983); Dawson (2000, 2002); and Baake and Baake (2003) for a sample of subsequent work. Under the so-called single-crossover assumption, where at most one crossover occurs in any gene sequence in every generation, the deterministic model can be solved explicitly (and in an astonishingly simple way) in *continuous* time (Baake and Baake 2003). But the corresponding discrete-time dynamics, which is prevalent in the biological literature, is more difficult; its solution has, so far, required nontrivial transformations and recursions that have not yet been solved in closed form (Bennett 1954; Dawson 2000, 2002; von Wangenheim et al. 2010).

In this paper, we will present a *semi-explicit* solution to the *discrete-time single-crossover population model* by considering the *ancestry of single individuals*. The original *deterministic forward-time dynamics* is thus considered in terms of a *stochastic process backward in time*, whose solution leads to that of the original system. In the backward process, one gains a certain conditional independence of gene segments, which will allow for a solution. In this sense, a probabilistic representation provides the necessary understanding to solve the original deterministic problem.

More precisely, we proceed as follows. In Section 2, we start from the stochastic (i.e., finite-population) version of the discrete-time single-crossover model, that is, the Wright-Fisher model with single-crossover recombination (Hein et al. 2005, Chap. 5.4). In the limit of population size tending to infinity (without rescaling of parameters or time), a law of large numbers (established here explicitly) leads to the corresponding deterministic dynamical system. We recall some general properties of this system and discuss the various dependencies (between individuals and between gene segments) that have, so far, obstructed an explicit solution. In Section 3, we take the backward point of view and consider the ancestry of the genetic material of single individuals. In the limit of infinite population size, this ancestry is a random tree for any finite time horizon, that is, segments that have been separated once do not come together again in the same individual (with probability one). The law for this ancestral tree may be formulated explicitly in terms of a (stochastic) segmentation process, which involves conditional independence between segments once they appear. As a consequence, the time evolution of the ancestral process may be calculated via a decomposition into subtrees. This solution is semi-explicit in the sense that it is a sum of well-defined terms, where summation is over certain tree topologies, which must be enumerated in a recursive way. In the same sense, this yields a solution of the deterministic forward-in-time model. We will discuss our results in the context of related approaches in Section 4, in particular, the ancestral recombination graph (the usual approach to recombination in finite populations).

## 2 The recombination model forward in time

### 2.1 The model

In this section, we describe the basic setting, the *Wright-Fisher model with single–crossover recombination*, as well as the dynamical system (from von Wangenheim et al. 2010) that arises as its infinite-population limit. A chromosome is described by a linear arrangement of, say, $n+1$ *sites*, namely, the elements of the set $S := \{0, 1, \ldots, n\}$. Sites represent discrete positions on a chromosome that may be interpreted as gene or nucleotide positions. Thus, each site $i \in S$ can be occupied by an *allele* (or *letter*) $x_i \in X_i$, where we restrict ourselves to *finite* $X_i$. A *type* $x$ is then defined as a sequence $x = (x_0, x_1, \ldots, x_n) \in X_0 \times X_1 \times \cdots \times X_n =: X$, where $X$ denotes the (finite) *type space*. Neighbouring sites are connected by *links*, the entities where recombination events may occur. They are collected into the set $L = \{\frac{1}{2}, \frac{3}{2}, \ldots, \frac{2n-1}{2}\}$, where link $\alpha = \frac{2i+1}{2}$ denotes the link between sites $i$ and $i+1$. We will only be concerned with single crossovers, i.e., the case where recombination occurs

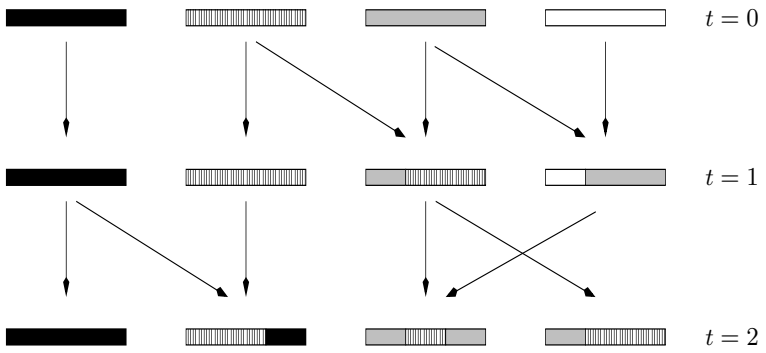


**Fig. 1** Wright-Fisher model with single–crossover recombination for $N = 4$.

at a single link $\alpha \in L$ and results in a mixed type composed of the sites before $\alpha$ from the first parent, and those after $\alpha$ from the second parent. Explicitly, if recombination involves the ordered pair of types $x = (x_0, \ldots, x_n)$ and $y = (y_0, \ldots, y_n)$, the outcome of recombination at link $\frac{2i+1}{2}$ is the recombined type $(x_0, \ldots, x_i, y_{i+1}, \ldots, y_n)$. The dynamics of a finite population that evolves under single-crossover recombination can be described by the following version of the Wright-Fisher model (cf. Hein et al. 2005, Chap. 5.4):

Each link is equipped with a crossover probability $\varrho_\alpha > 0$ (with $\sum_{\alpha \in L} \varrho_\alpha \leqslant 1$). Each generation is of constant size $N$. In each generation, the current population is replaced by its offspring, where each offspring individual chooses its parent(s) independently according to the following scheme (see Figure 1):

- With probability $\varrho_\alpha > 0$, $\alpha \in L$, two parents are chosen uniformly with replacement. They recombine at link $\alpha$, which gives rise to the corresponding recombined offspring with the leading segment (the sites $0, \ldots, \lfloor \alpha \rfloor$) from the first and the trailing segment (the sites $\lceil \alpha \rceil, \ldots, n$) from the second parent, where $\lfloor \alpha \rfloor$ ($\lceil \alpha \rceil$) denotes the largest integer below (the smallest above) $\alpha$; if the same parent is chosen twice, it is effectively transmitted unchanged.
- With probability $0 \leqslant 1 - \sum_{\alpha \in L} \varrho_\alpha < 1$, a single parent is selected uniformly and with replacement from the previous generation.

We denote the population at time $t$ by

$$Z_t = (Z_t(x))_{x \in X} \in E := \{\nu \text{ counting measure on } X \mid \|\nu\| = N\},$$

where $\|.\|$ denotes total variation norm and $Z_t(x)$ is the number of individuals of type $x$ at time $t$. We will also need the corresponding normalised quantity $\widehat{Z}_t := Z_t/N$, which is a probability vector (or measure) for the population at time $t$. In order to formalise the stochastic process, let $\pi_J \colon X \longrightarrow X_J := \bigtimes_{i \in J} X_i$, $\pi_J(x) = (x_i)_{i \in J} =: x_J$, be the canonical projection to the sites in $J$ ($J \subseteq S$). We specifically need $\pi_{<\alpha} := \pi_{\{0,\ldots,\lfloor \alpha \rfloor\}}$ and $\pi_{>\alpha} := \pi_{\{\lceil \alpha \rceil,\ldots,n\}}$. For $p \in \mathcal{P}(X)$, the set of probability measures on $X$, we denote by $\pi_J.p := p \circ \pi_J^{-1}$ (where $\pi_J^{-1}$ denotes the preimage of $\pi_J$) the marginal distribution of $p$ with respect to the sites in $J$. Furthermore,

$$R_\alpha(p) := (\pi_{<\alpha}.p) \otimes (\pi_{>\alpha}.p) \tag{1}$$

is the product measure of the two marginals (before and after $\alpha$); $R_\alpha$ is known as the recombination operator (or *recombinator* for short), cf. Baake and Baake (2003). It is clear that an individual that recombines at link $\alpha \in L$ in generation $t$ draws its type from $R_\alpha(\widehat{Z}_{t-1})$, and a non-recombining individual draws its type from $\widehat{Z}_{t-1} = R_\varnothing(\widehat{Z}_{t-1})$, with $R_\varnothing := \mathbb{1}$ (the reason for this notation will become clear later).

The discrete-time Markov chain $\{\widehat{Z}_t\}_{t \in \mathbb{N}_0}$ on $\mathcal{P}(X)$ may therefore be formulated as follows:

– Let $N_\alpha(t)$, $\alpha \in L$, denote the random number of individuals generated in generation $t$ via recombination at link $\alpha \in L$. Analogously, $N_\varnothing(t)$ is the number of individuals that are sampled without recombining. Clearly, they follow a multinomial distribution:

$$(N_\varnothing(t), N_{\frac{1}{2}}(t), \ldots, N_{\frac{2n-1}{2}}(t)) \sim \mathcal{M}(N, (1 - \sum_{\alpha\in L} \varrho_\alpha, \varrho_{\frac{1}{2}}, \ldots, \varrho_{\frac{2n-1}{2}}))\,, \quad \text{i.i.d for all } t\,. \tag{2}$$

– According to the previous step, $Z_t$ consists of subpopulations $Y_\beta(t)$, $\beta \in L \cup \{\varnothing\}$, where $Y_\beta(t)$ consists of those individuals that, in generation $t$, experience recombination at $\beta$ (where $\beta = \varnothing$ indicates no recombination). Clearly,

$$Y_\beta(t) \sim \mathcal{M}(N_\beta(t), R_\beta(\widehat{Z}_{t-1})), \quad \beta \in L \cup \{\varnothing\}. \tag{3}$$

– Finally, we obtain $\widehat{Z}_t$ via

$$\widehat{Z}_t = \frac{1}{N}(Y_\varnothing(t) + \sum_{\alpha\in L} Y_\alpha(t))\,. \tag{4}$$

Obviously, the resampling-recombination mechanism is independent of the types. So, the Wright-Fisher model may, alternatively, be constructed as an independent superposition of the two processes, that is,

**(F1)** It is first determined, for each time point and for each individual, which of the sites come from which parental individual (resampling/recombination without types).

**(F2)** Letters are then attached to the sites at time $t = 0$ and are then propagated through the model to time $t$ according to the relations decided in **(F1)**.

### 2.2 Law of large numbers.

Let us first consider the Wright-Fisher model in the so-called *infinite population limit (IPL)*, where we let $N \to \infty$ without rescaling any other parameters. This may be considered as a limit of *strong recombination*, in which the stochastic effects of resampling (also known as genetic drift) are lost. This is in contrast to the more frequently used weak recombination limit, which leads to a diffusion process, compare Ewens (2004, Chap. 6.6) and Section 4 below.

More precisely, we consider the family of processes $\{\widehat{Z}_t^{(N)}\}_{t\in\mathbb{N}_0}$, $N \in \mathbb{N}$ (where we temporarily add an upper index $N$ to denote population size) and compare it with the deterministic recombination dynamics, where we identify the population at time $t \in \mathbb{N}_0$ with $p_t = (p_t(x))_{x\in X} \in \mathcal{P}(X)$. Here $p_t(x)$ denotes the relative frequency of type $x \in X$ at time $t$, and $p_0$ is the initial population. The population is described by the dynamical system

$$p_t = \Phi(p_{t-1}), \quad \text{where } \Phi(p) := \Big(1 - \sum_{\alpha\in L} \varrho_\alpha\Big)p + \sum_{\alpha\in L} \varrho_\alpha R_\alpha(p)\,, \tag{5}$$

which is usually obtained by direct deterministic modelling (von Wangenheim et al. 2010). The following result shows that, indeed, (5) describes the infinite population limit of the stochastic process, more precisely for the family of processes $\{\widehat{Z}_t^{(N)}\}_{t\in\mathbb{N}_0}$, $N \in \mathbb{N}$. We will use $\Phi^{t+1} = \Phi \circ \Phi^t$ for the composition of the nonlinear mapping $\Phi$.

**Proposition 1 (Infinite Population Limit)** *Let $\{\widehat{Z}_t^{(N)}\}_{t\in\mathbb{N}_0}$ with $N \in \mathbb{N}$ be a family of Wright-Fisher models with single-crossover recombination (as defined by (2)–(4)) with initial states such that $\lim_{N\to\infty} \widehat{Z}_0^{(N)} = p_0$. Then, for every given $t \in \mathbb{N}_0$, one has*

$$\lim_{N\to\infty} \widehat{Z}_t^{(N)} = p_t \quad \textit{in mean square}, \tag{6}$$

*where $p_t = \Phi^t(p_0)$ denotes the solution of (5).*

The corresponding situation in *continuous* time and with *almost sure convergence* (see Remark 1) is covered by the general law of large numbers of Ethier and Kurtz (1986, Theorem 11.2.1), but no such general result seems to be available in discrete time. We therefore include a proof.

*Proof (of Prop. 1)* We employ induction over $t$. By assumption, the claim holds for $t = 0$. Now assume that it holds for $t-1$, for some $t \geqslant 1$. We then have

$$\frac{N_\alpha^{(N)}(t)}{N} \xrightarrow{N\to\infty} \varrho_\alpha \,, \alpha \in L, \quad \text{and} \quad \frac{N_\varnothing^{(N)}(t)}{N} \xrightarrow{N\to\infty} 1 - \sum_{\alpha\in L} \varrho_\alpha \tag{7}$$

by the mean square law of large numbers (cf. Grimmett 2001, Chap. 7.4). Furthermore, for $\beta \in L$, $N_\beta^{(N)}(t) \to \infty$ as $N \to \infty$ with probability one (because $\varrho_\beta > 0$) and thus

$$\frac{Y_\beta^{(N)}(t)}{N_\beta^{(N)}(t)} \xrightarrow{N\to\infty} R_\beta(p_{t-1}) \quad \text{in mean square,} \tag{8}$$

since $Y_\beta^{(N)}(t)/N_\beta^{(N)}(t) - R_\beta(\widehat{Z}_{t-1}^{(N)}) \xrightarrow{N\to\infty} 0$ due to the mean-square law of large numbers (except on the set of measure 0 where $N_\beta^{(N)}(t) \nrightarrow \infty$, and thus altogether in mean square), and $R_\beta(\widehat{Z}_{t-1}^{(N)}) \xrightarrow{N\to\infty} R_\beta(p_{t-1})$ by the induction hypothesis. Analogously, for $1 - \sum_{\alpha\in L} \varrho_\alpha > 0$,

$$\frac{Y_\varnothing^{(N)}(t)}{N_\varnothing^{(N)}(t)} \xrightarrow{N\to\infty} p_{t-1} \quad \text{in mean square.} \tag{9}$$

Since, by (4),

$$\widehat{Z}_t^{(N)} = \frac{N_\varnothing^{(N)}(t)}{N} \cdot \frac{Y_\varnothing^{(N)}(t)}{N_\varnothing^{(N)}(t)} + \sum_{\alpha\in L} \frac{N_\alpha^{(N)}(t)}{N} \cdot \frac{Y_\alpha^{(N)}(t)}{N_\alpha^{(N)}(t)},$$

(7)–(9) together tell us that

$$\widehat{Z}_t^{(N)} \xrightarrow{N\to\infty} \Phi(p_{t-1}) \quad \text{in mean square,} \tag{10}$$

which proves the claim. □

*Remark 1* Note that Prop. 1 automatically implies that, for every given $t$,

$$\lim_{N\to\infty} \max_{s\leqslant t} |\widehat{Z}_s^{(N)} - p_s| = 0 \quad \text{in mean square,}$$

which is reminiscent of the continuous-time result (Ethier and Kurtz 1986, Theorem 11.2.1). Note, however, that the latter result is a *strong* law of large numbers; we have established the mean-square version here (which, of course, implies a weak law of large numbers since convergence in mean square implies convergence in probability) since the construction of a sequence of processes on the *same* probability space in the discrete-time setting is beyond the scope of this paper. Note also that the convergence in (6) applies for any fixed $t \in \mathbb{N}_0$, but need *not* hold as $t \to \infty$. Indeed, the asymptotic behaviour of the stochastic system is radically different from that of the deterministic one: Due to resampling, the Markov chain is absorbing (in fact, it experiences fixation of a single type with probability one in the long run). In contrast, the deterministic system never loses any type, and the complete product measure with respect to all links in $L$ is obtained as the stationary distribution, see Geiringer (1944) and von Wangenheim et al. (2010). Let us emphasise that it is the short time scale, not the long-term behaviour, that we are interested in here; see Section 4 for the discussion of the biological context.

### 2.3 Structure of the deterministic solution

Based upon an initial population $p_0$, every individual in the population at time $t = 1$ is either an unaltered copy of an individual from $p_0$ or it is composed of exactly two recombined segments, hence the population $p_1$ is a mixture of $p_0$ and the $R_\alpha(p_0)$, $\alpha \in L$, in line with (5). For $t > 1$, the population will contain individuals that consist of several segments pieced together from the sequences in the initial population due to various recombination events at different times. To describe these, we use the *composite* recombinators $R_G$, $G \subseteq L$, which act on probability vectors as

$$R_G := \prod_{\alpha\in G} R_\alpha, \tag{11}$$

where we set $R_{\{\alpha\}} = R_\alpha$. Here, the product is to be read as composition. It is, indeed, a matrix product if the recombinators are written in their matrix representation, which is available in the case of finite types considered here, provided the problem is embedded into a larger space (Baake 2001). This definition is consistent since all $R_\alpha$ are idempotents and commute with each other, compare Baake and Baake (2003). Clearly, $R_G(p)$ is the product measure derived from $p$ with respect to all links in $G$. We thus expect the population at any time to be a convex combination of the $R_G(p_0)$ with $G \subseteq L$. This means

$$p_t = \Phi^t(p_0) = \sum_{G\subseteq L} a_G(t) R_G(p_0)\,, \tag{12}$$

with $a_G(0) = \delta_{G,\varnothing}$, $a_G(t) \geqslant 0$ for all $G \subseteq L$, and $\sum_{G\subseteq L} a_G(t) = 1$. It has been proved by von Wangenheim et al. (2010) that the solution indeed has this form, but plausibility arguments go back to Geiringer (1944). The difficulty consists in determining the coefficient functions $a_G(t)$. Let us introduce the following abbreviations,

$$G_{<\alpha} := \{\beta \in G \mid \beta < \alpha\}\,, \quad G_{>\alpha} := \{\beta \in G \mid \beta > \alpha\}\,,$$
$$G_{\leqslant\alpha} := \{\beta \in G \mid \beta \leqslant \alpha\}\,, \quad G_{\geqslant\alpha} := \{\beta \in G \mid \beta \geqslant \alpha\}\,.$$

Let us recall the recursion for the coefficient functions from von Wangenheim et al. (2010):

**Theorem 1** *For all $G \subseteq L$ and $t \in \mathbb{N}_0$, the coefficient functions $a_G(t)$ evolve according to*

$$a_G(t+1) = \Big(1 - \sum_{\alpha\in L} \varrho_\alpha\Big)\, a_G(t) + \sum_{\alpha\in G} \varrho_\alpha \Big(\sum_{H\subseteq L_{\geqslant\alpha}} a_{G_{<\alpha}\cup H}(t)\Big)\Big(\sum_{K\subseteq L_{\leqslant\alpha}} a_{K\cup G_{>\alpha}}(t)\Big), \tag{13}$$

*with initial condition $a_G(0) = \delta_{G,\varnothing}$.* □

A verbal description of this iteration can already be found in Geiringer (1944). It will become clear later that we may interpret $a_G(t)$ as the proportion of the population whose types have been pieced together by recombination at *exactly* the links of $G$.

Due to its nonlinearity, the recursion does not allow for an immediate solution (at least from four sites onwards). The nonlinearity comes from the *dependence* of links: Due to the single-crossover assumption, a crossover event forbids any other recombination events in the same time step. In sharp contrast, and quite surprisingly, the analogous (deterministic) single–crossover model in *continuous* time has a very simple explicit solution (Baake and Baake 2003; Baake 2005). The main reason for this is the fact that simultaneous crossover events are *automatically* excluded in continuous time. This implies an *effective independence* of links, which turns the dynamics corresponding to Theorem 1 into a *linear* one. For a detailed investigation of the differences between single-crossover dynamics in continuous and in discrete time, the reader is referred to von Wangenheim et al. (2010).

The conventional way (Bennett 1954; Dawson 2000, 2002) to overcome the obstacles of nonlinearity in recombination models lies in finding an appropriate transformation of the dynamics

to a solvable diagonalised system, but this usually involves a new set of coefficients that must be constructed in a recursive manner. We have performed this for the single-crossover model (von Wangenheim et al. 2010), but the solution still requires recursions and does not lead to closed-form expressions for the $a_G(t)$. In contrast, we will pursue the stochastic perspective here and look at recombination *backward* in time, which will lead us to the coefficient functions in semi-explicit form.

## 3 Ancestral recombination process

### 3.1 The ancestral process

In the *ancestral recombination process*, we follow the ancestry of the genetic material of a selected individual from a population that evolved according to the Wright-Fisher model with single-crossover recombination of Section 2.1. To this end, we start with an individual in the present population at time $t$ and let time run backwards, as illustrated in Figure 2 for two individuals from the realisation of the Wright-Fisher model in Figure 1. Let us first describe the resulting partitioning of sites into parents, keeping in mind that this happens independently of the types, in analogy with step **(F1)** in the forward model.

We denote by $\tau$ the time backward from the present at time $t$, i.e., backward time $\tau$ corresponds to forward time $t-\tau$. (Note that, altogether, we use the symbol $t$ both for the variable of time and for the fixed number of generations for which the (forward-time) dynamics is considered. In the latter sense, $t$ stands for 'today'.) We capture the partitioning by a process $\{\Sigma_\tau\}_{\tau\in\mathbb{N}_0}$ on $\Pi(S)$, the set of partitions of $S$. Here, the parts of $\Sigma_\tau$ correspond to the parents at (backward) time $\tau$ of our individual at (forward) time $t$; sites in the same part correspond to sites that go back to the same parent. In view of the forward Wright-Fisher model, it is clear that $\{\Sigma_\tau\}_{\tau\in\mathbb{N}_0}$ is declared as follows.

Start with $\Sigma_0 = \{S\}$. Assume now that for some $\tau > 0$, $\Sigma_\tau = \sigma := \{\sigma_1, \ldots, \sigma_k\}$, where $\sigma_j = \{\sigma_{j1}, \ldots, \sigma_{jn_j}\}$ and we imply $\sigma_{j1} < \sigma_{j2} < \cdots < \sigma_{jn_j}$, $1 \leqslant j \leqslant k$. Referring back to the Wright-Fisher model, $\Sigma_{\tau+1}$ is obtained in two steps:

(S) Splitting: Every part $\sigma_j$ of $\Sigma_\tau$, $1 \leqslant j \leqslant k$, independently of the others, either remains unchanged (probability $1 - \sum_{\sigma_{j1}<\alpha<\sigma_{jn_j}} \varrho_\alpha$), or, for every $\sigma_{j1} < \alpha < \sigma_{jn_j}$, it may split into $\{\sigma_{j1}, \ldots, \sigma_{j\lfloor\alpha\rfloor}\}$ and $\{\sigma_{j\lceil\alpha\rceil}, \ldots, \sigma_{jn_j}\}$ (probability $\varrho_\alpha$). Note that two or more $\alpha$'s can lead to the same split if $\sigma_j$ is not contiguous, where 'contiguous' means an uninterrupted run of sites. The resulting refined partition is denoted by $\Sigma'_\tau$. This step corresponds to the splitting of the ancestral material into smaller segments due to recombination, where we do not yet decide which segment ends up in which parent.

(C) Coalescence: Each part of $\Sigma'_\tau$ now chooses one out of $N$ parents, uniformly and with replacement. Parts that end up in the same parent are united; otherwise, nothing happens. The resulting partition is $\Sigma_{\tau+1}$. Figure 2 illustrates this: If all parts are assigned to different parents, then no coalescence takes place, that is, $\Sigma_{\tau+1} = \Sigma'_\tau$ (as in Figure 2, left). If two or more parts go back to the same parent, we have a coalescence event, see Figure 2 (right).

A closely related process describing the ancestry of single individuals in *continuous time* and on a *continuous chromosome* was investigated by Wiuf and Hein (1997), but in the weak-recombination limit, and with a different purpose; we will come back to this in Section 4.

Our aim is now to determine the law for the ancestry and the type of a random individual at time $t$ *without* constructing a realisation of the forward Wright-Fisher model first. Such an individual, together with its ancestry, may be constructed in a three-step procedure, see Figure 3.

(A1) Run $\{\Sigma_\tau\}_{\tau\in\mathbb{N}_0}$ until $\tau = t$. $\Sigma_t$ tells us how the ancestral material of our individual is partitioned into parents at forward time 0 (in Figure 3, this is the top of the tree).

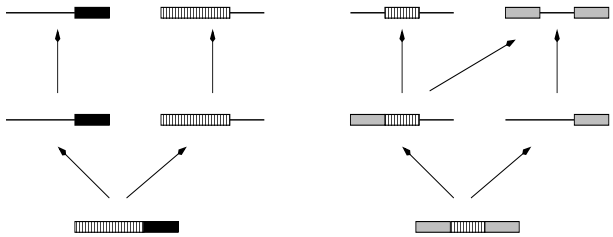

**Fig. 2** Ancestries for two individuals from the Wright-Fisher population of Figure 1. We trace back the ancestry of the segments present at $t = 2$; the thin black lines indicate nonancestral material whose history is not relevant. The left graph refers to the second individual from Figure 1; here the two segments go back to two different ancestors. The right graph corresponds to the third individual from Figure 1. Two of its three segments have the same parent at $t = 0$ due to a coalescence event. In the infinite population limit, such a situation does not occur; rather, all possible ancestries are binary trees.

(A2) Assign a different colour to each part of $\Sigma_t$. The colours are for illustration; each colour corresponds to one individual from the initial population (at $t = 0$), chosen uniformly *without* replacement. Equivalently, one may sample a parent from the initial population *with* replacement for every part of $\Sigma'_{t-1}$. (The latter is more convenient and will be favoured in what follows.) In any case, every site receives a colour, which is propagated downwards. This results in the present individual pieced together from segments of different colours that correspond to different parental individuals.

(A3) Assign a letter to every site at $t = 0$ (i.e., $\tau = t$). By (A2), this entails that the type for part $\sigma_j$ of $\Sigma'_{t-1}$ is drawn from $\pi_{\sigma_j}.\widehat{Z}_0$, independently for every element of the partition. Like the colours, the letters are attached to the sites once and for all, and thus propagated downwards, i.e., down to $\Sigma_0$.

As a consequence of (A2) and (A3), conditional on $\Sigma'_{t-1} = \sigma = \{\sigma_1, \ldots, \sigma_k\}$, the type distribution at present (that is, at forward time $t$) is $:(\pi_{\sigma_1}.\widehat{Z}_0) \otimes \cdots \otimes (\pi_{\sigma_k}.\widehat{Z}_0):$, where $:\ldots:$ means that the factors are ordered as in $X$. Denoting by $\Xi_t$ the type at forward time $t$, we thus have

$$\mathbb{P}(\Sigma'_{t-1} = \{\sigma_1, \ldots, \sigma_k\}, \Xi_t = x) = \mathbb{P}(\Sigma'_{t-1} = \{\sigma_1, \ldots, \sigma_k\}) : (\pi_{\sigma_1}.\widehat{Z}_0) \otimes \cdots \otimes (\pi_{\sigma_k}.\widehat{Z}_0) : (x) \,. \qquad (14)$$

Eq. (14) gives the marginal distribution (of partition and type) for every *single* individual in a sample, or in the entire population. Due to coalescence events, however, the individuals in a finite population are correlated, and the joint distribution is a difficult matter. (This is investigated within the framework of the *ancestral recombination graph*, which traces back the genealogy of a *sample* of individuals; compare Wakeley 2008, Chap. 7.2, Durrett 2008, Chap. 3.4, and Section 4).

Our goal here is a somewhat simpler one, namely, the distribution of types and ancestries in the $N \to \infty$ limit (under *strong* recombination). In this limit, the partitioning process simplifies substantially due to the following result.

**Lemma 1** *Let $\Omega_t$ be the event that $\Sigma_\tau = \Sigma'_{\tau-1}$ for $1 \leqslant \tau \leqslant t$; that is, no coalescence occurs until (backward) time $t$, or, equivalently, $\{\Sigma_\tau\}_{0 \leqslant \tau \leqslant t}$ is a process of progressive refinements of ordered partitions. For every fixed finite $t$, one has $\mathbb{P}(\Omega_t) \geqslant 1 - n(n+1)t/2N + \mathcal{O}(1/N^2)$.*

*Proof* For every $\tau$, in step (S), $\Sigma'_\tau$ is obtained from $\Sigma_\tau$ as a refinement. It is thus clear that $\{\Sigma_\tau\}_{0 \leqslant \tau \leqslant t}$ is a process of progressive refinements (and hence of ordered partitions) if and only if $\Sigma_\tau = \Sigma'_{\tau-1}$ for $1 \leqslant \tau \leqslant t$. If $\Sigma'_{\tau-1}$ has $k$ parts, then the probability that each is assigned to a different parent in the coalescence step leading to $\Sigma_\tau$ is

$$q_k := 1 \cdot \left(1 - \frac{1}{N}\right) \cdot \left(1 - \frac{2}{N}\right) \cdot \ldots \cdot \left(1 - \frac{k-1}{N}\right) . \qquad (15)$$

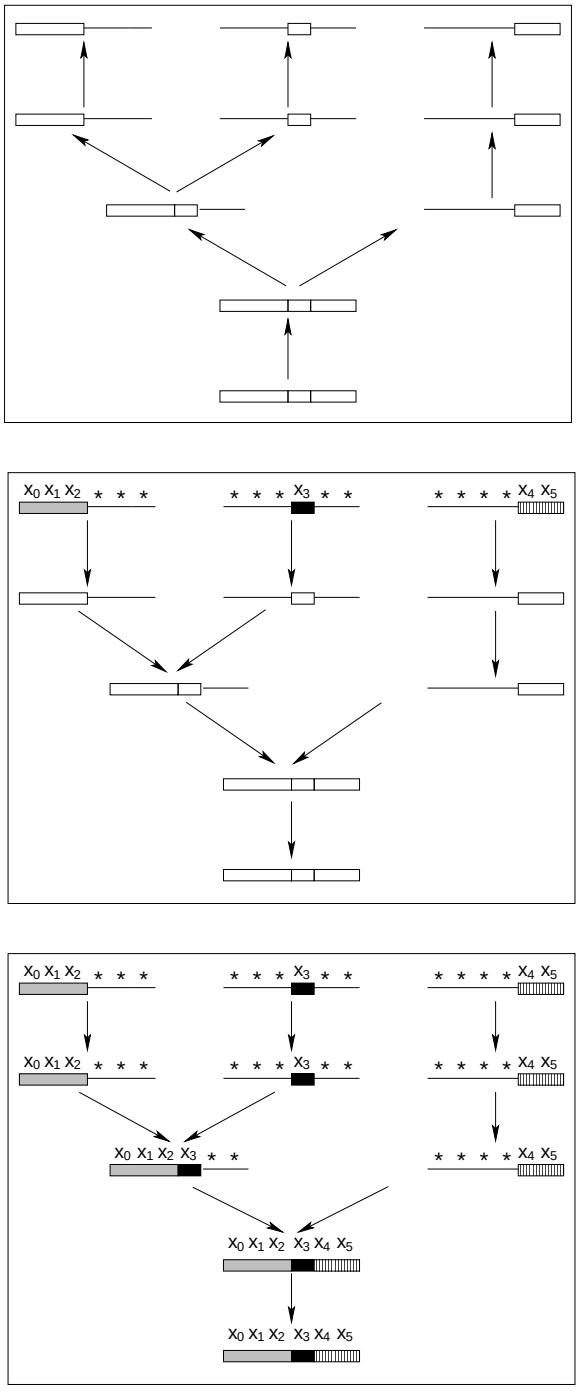


**Fig. 3** Construction of a random individual at time $t$, together with its ancestry. Top: partitioning of sites (backward; step (A1)); middle: assignment of colours and letters at the top; bottom: propagation of colours and letters downwards. The middle and bottom panels together correspond to the simultaneous performance of steps (A2) and (A3). In this example, there are no coalescence events, so all partitions are ordered. As a consequence, this realisation of the partitioning process $\{\Sigma_\tau\}_{0\leqslant\tau\leqslant t}$ is a tree and, at the same time, a realisation of the segmentation process $\{F_\tau\}_{0\leqslant\tau\leqslant t}$ of Section 3.2.

Obviously,

$$q_k \geqslant q_{n+1} = 1 - \frac{n(n+1)}{2N} + \mathcal{O}(1/N^2) \tag{16}$$

because $k \leqslant |S| = n+1$. As a consequence,

$$\mathbb{P}(\Omega_t) \geqslant q_{n+1}^t = 1 - \frac{n(n+1)t}{2N} + \mathcal{O}(1/N^2)$$

for every fixed finite $t$. □

Note that Lemma 1 implies that, for any finite $t$, coalescence events are absent in the $N \to \infty$ limit and the ancestry is a tree – in line with intuition, and as in Figure 2 (left), and in Figure 3. Note also that Lemma 1 holds for any finite $t$, but not for $t \to \infty$, in the same spirit as the law of large numbers in Prop. 1.

### 3.2 Segments and the segmentation process

Since, as we have just seen, we only have to deal with ordered partitions (with probability one for any finite $t$ as $N \to \infty$), we can introduce a simplifying notation for the partitions that is based on links rather than on sites. This is because ordered partitions are in one-to-one correspondence with the subsets of $L$ as follows. As in Baake (2005), let $G = \{\alpha_1, \ldots, \alpha_{|G|}\} \subseteq L$, with $\alpha_1 < \alpha_2 < \cdots < \alpha_{|G|}$, an ordering which we will assume implicitly from now on. Let then $\mathcal{S}(\varnothing) := \{S\}$ and, for $G \neq \varnothing$, let $\mathcal{S}(G) := \{\sigma_1, \sigma_2, \ldots, \sigma_{|G|+1}\}$ denote the ordered partition of $S$ with parts

$$\sigma_1 := \{0, \ldots, \lfloor \alpha_1 \rfloor\},\ \sigma_2 := \{\lceil \alpha_1 \rceil, \ldots, \lfloor \alpha_2 \rfloor\}, \ldots, \sigma_{|G|+1} := \{\lceil \alpha_{|G|} \rceil, \ldots, n\}. \tag{17}$$

In particular, $\mathcal{S}(L) = \big\{\{0\}, \ldots, \{n\}\big\}$. It is clear that $\mathcal{S}(H)$ is a refinement of $\mathcal{S}(G)$ if and only if $G \subseteq H$. It is also obvious that $\mathcal{S}$ defines a bijection; its inverse, $\psi := \mathcal{S}^{-1}$, associates with every *ordered* partition of $S$ the corresponding subset of $L$, so that $\psi(\mathcal{S}(G)) = G$ for all $G \subseteq L$.

We now define the associated ordered partitions $\mathcal{L}_G$ of $L \setminus G$. Let $\mathcal{L}_\varnothing = \{L\}$ and, for $G \neq \varnothing$, set (cf. Figure 4):

$$\begin{aligned} \widetilde{\mathcal{L}}_G &:= \Big\{ \big\{\alpha \in L : \tfrac{1}{2} \leqslant \alpha < \alpha_1\big\}, \big\{\alpha \in L : \alpha_1 < \alpha < \alpha_2\big\}, \ldots, \big\{\alpha \in L : \alpha_{|G|} < \alpha \leqslant \tfrac{2n-1}{2}\big\} \Big\}, \\ \mathcal{L}_G &:= \widetilde{\mathcal{L}}_G \setminus \{\varnothing\}\,. \end{aligned} \tag{18}$$

That is, $\mathcal{L}_G$ is the ordered partition of $L \setminus G$ that holds the segments (in the sense of contiguous sets of links) that arise when recombination has occurred at all links in $G$; in particular, $\mathcal{L}_L = \{\}$.

Let us now consider the following process of progressive segmentation (which will turn out to coincide with the $N \to \infty$ limit of the partitioning process for any finite time).

**Definition 1 (Segmentation process)** The *segmentation process* is the discrete-time Markov chain $\{F_\tau\}_{\tau \in \mathbb{N}_0}$, where $F_\tau$ takes values in the power set of $L$ according to the following rules. Start with $F_0 = \varnothing$ and recall that $\mathcal{L}_\varnothing = \{L\}$. If $F_\tau = G$, choose either none or one link in every segment, according to the following rule. From segment $I$ of $\mathcal{L}_G$, independently of all other segments, either no link is chosen (probability $1 - \sum_{\alpha \in I} \varrho_\alpha$), or a single link is chosen, namely link $\alpha \in I$ with probability $\varrho_\alpha$. Then $F_{\tau+1}$ is the union of $G$ with the set of all newly chosen links. That is,

$$F_{\tau+1} = F_\tau \cup A, \quad \text{where} \quad A = \big( \bigcup_{I \in \mathcal{L}_G} A_I \big) \tag{19}$$

and

$$A_I = \begin{cases} \varnothing, & \text{with probability } 1 - \sum_{\alpha \in I} \varrho_\alpha\,, \\ \{\alpha\}, & \text{with probability } \varrho_\alpha, \text{ for all } \alpha \in I\,. \end{cases} \tag{20}$$

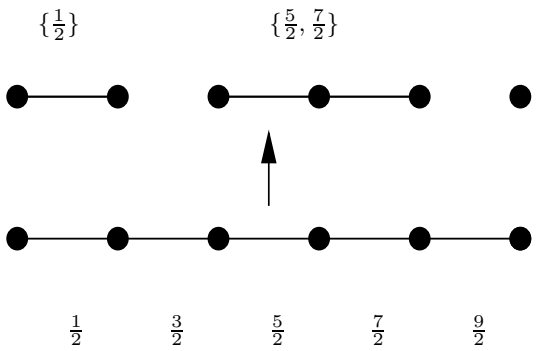


**Fig. 4** The segments induced by $G = \{\frac{3}{2}, \frac{9}{2}\}$ in the case $L = \{\frac{1}{2}, \ldots, \frac{9}{2}\}$ (i.e., $\widetilde{\mathcal{L}}_G = \{\{\frac{1}{2}\}, \{\frac{5}{2}, \frac{7}{2}\}, \varnothing\}$, $\mathcal{L}_G = \{\{\frac{1}{2}\}, \{\frac{5}{2}, \frac{7}{2}\}\}$). $\mathcal{L}_G$ corresponds to the ordered partion $\mathcal{S}(G) = \{\{0,1\}, \{2,3,4\}, \{5\}\}$ of $S = \{0,1,2,3,4,5\}$.

Clearly, picking a link corresponds to recombination, and $F_\tau$ is the set of links that have been cut until time $\tau$. Note that, as in the Wright-Fisher model with recombination (and its deterministic limit), the links are not, in general, independent: At most one link in a given segment may be cut in one time step; cutting of one link prevents cutting of any other link in the same generation. However, the backward point of view adopted here reveals *(conditional) independence of the individual segments* once they arise. Put differently, links are independent as soon as they are on different segments. This is analogous to the conditional independence of offspring inviduals in branching processes and will turn out as the golden key to the solution.

The connection of the segmentation process with the partitioning process can now be clarified (we use upper indices once more to denote the dependence on population size):

**Proposition 2** *Let $t \geqslant 0$ be arbitrary but fixed. The law of $\{\Sigma_\tau^{(N)}\}_{0\leqslant\tau\leqslant t}$ then agrees with that of $\{F_\tau\}_{0\leqslant\tau\leqslant t}$ up to $\mathcal{O}(1/N)$, provided $\{\Sigma_\tau^{(N)} = \sigma\}$ is put in bijective correspondence with $\{F_\tau = \psi(\sigma)\}$ for all ordered partitions $\sigma$ and $0 \leqslant \tau \leqslant t$. For individual time points $\tau \leqslant t$, this implies specifically that*

$$\mathbb{P}(\Sigma_\tau^{(N)} = \sigma) = \begin{cases} \mathbb{P}(F_\tau = \psi(\sigma)) + \mathcal{O}(1/N), & \text{if } \sigma \text{ is an ordered partition,} \\ \mathcal{O}(1/N), & \text{otherwise} \end{cases} \tag{21}$$

*for every $\sigma \in \Pi(S)$, with $\psi = \mathcal{S}^{-1}$ as defined after* (17).

*Proof* It is clear that, under the above identification, the initial conditions ($\Sigma_0^{(N)} = \{S\}$ and $F_0 = \psi(\{S\}) = \varnothing$) agree. It is also clear that, if $\Sigma_\tau^{(N)} = \mathcal{S}(G)$ for some $G \subseteq L$, then $(\Sigma^{(N)})'_\tau$ follows the same law as $F_{\tau+1}$ given $F_\tau = G$ (by step (S) and Def. 1).

Consider now $\{\Sigma_\tau^{(N)}\}_{0\leqslant\tau\leqslant t}$ conditional on $\Omega_t$. By the above observation together with Lemma 1, the law of the conditional process may be understood as follows. Run $\{F_\tau\}_{0\leqslant\tau\leqslant t}$, but in every step $\tau$, kill the process with probability $1 - q_k$ (from (15)) if $|F_\tau| = k$. The law of the surviving process then is the law of $\{\Sigma_\tau^{(N)} \mid \Omega_t\}_{0\leqslant\tau\leqslant t}$. Since the killing probability up to time $t$ is $1 - \mathbb{P}(\Omega_t) = \mathcal{O}(1/N)$ (see Lemma 1), the law of $\{\Sigma_\tau^{(N)} \mid \Omega_t\}_{0\leqslant\tau\leqslant t}$ agrees with that of $\{F_\tau\}_{0\leqslant\tau\leqslant t}$ up to $\mathcal{O}(1/N)$. Finally, using $1 - \mathbb{P}(\Omega_t) = \mathcal{O}(1/N)$ once more yields the claim.

□

Let us remark that (21) (evaluated at time $\tau - 1$) also entails that

$$\mathbb{P}((\Sigma^{(N)})'_{\tau-1}) = \sigma) = \begin{cases} \mathbb{P}(F_\tau = \psi(\sigma)) + \mathcal{O}(1/N), & \text{if } \sigma \text{ is an ordered partition,} \\ \mathcal{O}(1/N), & \text{otherwise} \end{cases} \tag{22}$$

since $(\Sigma^{(N)})'_{\tau-1}$ is obtained from $\Sigma^{(N)}_{\tau-1}$ according to the same law as $F_\tau$ from $F_{\tau-1}$, provided $\Sigma^{(N)}_{\tau-1}$ is an ordered partition. Let us also remark that, for *finite* $N$, $\{\Sigma^{(N)}_\tau \mid \Omega_t\}_{0\leqslant\tau\leqslant t}$ only agrees with $\{F_\tau\}_{0\leqslant\tau\leqslant t}$ up to a probability of $\mathcal{O}(1/N)$. This is due to a bias, in a finite population, towards partitions with fewer parts, since these bear less risk of coalescence.

Before we proceed, let us note another elementary, but crucial property of the segmentation process. Consider the segmentation process $\{F^{(\tilde{L})}_\tau\}_{\tau\in\mathbb{N}_0}$ on a *contiguous* subset $\tilde{L}$ of $L$. Here $\{F^{(\tilde{L})}_\tau\}_{\tau\in\mathbb{N}_0}$ is defined in the same way as $\{F_\tau\}_{\tau\in\mathbb{N}_0}$ but with $L$ replaced by $\tilde{L}$, and based on the recombination probabilities $\varrho_\alpha$, $\alpha\in\tilde{L}$, alone. Here, the upper index now indicates dependence on the (sub-) set of links, which we may omit if $\tilde{L}=L$; that is, $F^{(L)}_\tau = F_\tau$. Likewise, we will denote by $\mathcal{L}^{(\tilde{L})}_G$, $G\subseteq\tilde{L}$, the partition of $\tilde{L}\setminus G$ defined in analogy with (18), with $L$ replaced by $\tilde{L}$. We then have the following important fact.

**Proposition 3 (marginalisation property)** *Let $\tilde{L}$ be a contiguous subset of $L$. The process $\{F^{(\tilde{L})}_\tau\}_{\tau\in\mathbb{N}_0}$ then is the marginal version of $\{F^{(L)}_\tau\}_{\tau\in\mathbb{N}_0}$ with respect to the links in $\tilde{L}$, that is, for all $G\subseteq\tilde{L}$ and all $\tau\in\mathbb{N}_0$, we have*

$$\mathbb{P}(F^{(\tilde{L})}_\tau = G) = \sum_{H\subseteq L\setminus\tilde{L}} \mathbb{P}(F^{(L)}_\tau = G\cup H).$$

*Proof* We will prove the claim by showing that, for every $G\subseteq\tilde{L}$ and $H\subseteq L\setminus\tilde{L}$, the set $A^{(\tilde{L})}$ of links picked from the segments of $\mathcal{L}^{(\tilde{L})}_G$ if $F^{(\tilde{L})}_\tau = G$ follows the same law as $A^{(L)}\cap\tilde{L}$, where $A^{(L)}$ is the set of links picked from the segments of $\mathcal{L}^{(L)}_{G\cup H}$ if $F^{(L)}_\tau = G\cup H$ (according to (19) and (20), and likewise for $\tilde{L}$). Due to the independence of the segments (within $\mathcal{L}^{(\tilde{L})}_G$ and within $\mathcal{L}^{(L)}_{G\cup H}$), we may consider these segments separately. Segments that are contained in both $\mathcal{L}^{(\tilde{L})}_G$ and $\mathcal{L}^{(L)}_{G\cup H}$ contribute identically to $A^{(\tilde{L})}$ and $A^{(L)}\cap\tilde{L}$ by construction. Segments $I\in\mathcal{L}^{(L)}_{G\cup H}$ with $I\cap\tilde{L}=\varnothing$ do not contribute to $A^{(L)}\cap\tilde{L}$ and are independent of those in $\mathcal{L}^{(\tilde{L})}_G$ and thus of $A^{(\tilde{L})}$. We are left to consider segments $\tilde{I}\in\mathcal{L}^{(\tilde{L})}_G$ with $\tilde{I}\subsetneq I\in\mathcal{L}^{(L)}_{G\cup H}$ for some $I$. But here the probability to pick any $\alpha\in\tilde{I}$ (for $A^{(\tilde{L})}$ ) is the same as picking this $\alpha$ from $I$ for $A^{(L)}\cap\tilde{L}$, namely, $\varrho_\alpha$, which completes the proof. □

We are now ready to state the main result of this section.

**Theorem 2 (type distribution via ancestral process)** *Consider a sequence of Wright-Fisher models with single-crossover recombination and increasing population size $N$. Let the initial states be such that $\lim_{N\to\infty}\widehat{Z}^{(N)}_0 = p_0$. The type distribution of any given individual for any finite $t\in\mathbb{N}_0$ converges to $\sum_{G\subseteq L}\mathbb{P}(F_t=G)R_G(p_0)$ as $N\to\infty$. For the composition of the population, we have*

$$\lim_{N\to\infty}\widehat{Z}^{(N)}_t = \sum_{G\subseteq L}\mathbb{P}(F_t=G)R_G(p_0) \quad \textit{in mean square.} \tag{23}$$

Clearly, (23) is again a law of large numbers, analogous to the infinite-population limit of Prop. 1, but this time expressed in terms of the *backward* process; the connection will be exploited in the next section. As in Remark 1, the result again carries over to finite time intervals.

*Proof (of Theorem 2)* We prove the theorem by considering the joint distribution of partitions and types, $\Sigma'_{t-1}$ and $\Xi_t$, as in (14). We will omit the dependence on $N$ for ease of notation. For a given single individual, we first rewrite the joint probabilities as

$$\mathbb{P}(\Sigma'_{t-1}=\sigma, \Xi_t = x) = \mathbb{P}(\Sigma'_{t-1}=\sigma)\,\mathbb{P}(\Xi_t = x \mid \Sigma'_{t-1}=\sigma). \tag{24}$$

As to the first term on the right-hand side, recall that, by (22), the only partitions that survive in the limit are $\mathcal{S}(G)$, $G\subseteq L$, for which $\mathbb{P}(\Sigma'_{t-1}=\mathcal{S}(G)) = \mathbb{P}(F_t=G)+\mathcal{O}(1/N)$. As to the second term, (14) tells us that the type distribution corresponding to $\mathcal{S}(G)$ is

$$(\pi_{\sigma_1}.\widehat{Z}_0)\otimes\cdots\otimes(\pi_{\sigma_{|G|+1}}.\widehat{Z}_0) = R_G(\widehat{Z}_0)$$

with $\sigma_1, \dots, \sigma_{|G|+1}$ of (17). But $R_G(\widehat{Z}_0)$ converges to $R_G(p_0)$ by assumption. By Lemma 1, in the limit $N \to \infty$, (24) thus becomes

$$\lim_{N\to\infty} \mathbb{P}(\Sigma'_{t-1} = \sigma, \Xi_t = x) = \begin{cases} \mathbb{P}(F_t = G)(R_G(p_0))(x) & \text{if } \sigma = \mathcal{S}(G) \\ 0 & \text{otherwise,} \end{cases} \tag{25}$$

from which the type distribution for single individuals follows by marginalisation over $\Sigma'_{t-1}$.

As to the population, let individuals be numbered $1, 2, \dots, N$, and let $\Sigma'_{t-1,i}$ and $\Xi_{t,i}$ be the partition at (backward) time $t-1$ (after splitting), and the type at (forward) time $t$, respectively, of individual $i$, $1 \leqslant i \leqslant N$ (so that the above $\Sigma'_{t-1}$ and $\Xi_t$ may be identified with $\Sigma'_{t-1,1}$ and $\Xi_{t,1}$). The $\Sigma'_{t-1,i}$ are identically distributed (across $i$), but not independent (they are correlated due to common ancestry); the same holds for the $\Xi_{t,i}$. Clearly,

$$\widehat{Z}_t(x) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{\Xi_{t,i} = x\}, \quad x \in X,$$

where $\mathbb{1}\{\dots\}$ denotes the indicator function for the event in question. We will show that, for all $x \in X$ and $\sigma \in \Pi(S)$,

$$\lim_{N\to\infty} \Big( \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{\Sigma'_{t-1,i} = \sigma, \Xi_{t,i} = x\} - \mathbb{P}(\Sigma'_{t-1,1} = \sigma, \Xi_{t,1} = x) \Big) = 0 \tag{26}$$

in mean square. Eq. (23) then follows via summation over all $\sigma \in \Pi(S)$, together with the result for single indviduals, which tells us that $\mathbb{P}(\{\Xi_{t,1} = x\}) \xrightarrow{N\to\infty} \mathbb{P}(F_t = G)\big(R_G(p_0)\big)(x)$.

To establish (26), it is sufficient to show that the covariance of $\mathbb{1}\{\Sigma'_{t-1,1} = \sigma, \Xi_{t,1} = x\}$ and $\mathbb{1}\{\Sigma'_{t-1,2} = \sigma, \Xi_{t,2} = x\}$ is $\mathcal{O}(1/N)$; due to exchangeability, this then carries over to arbitrary pairs of individuals. Let $\widetilde{\Omega}_t$ be the event that no coalescence happens between ancestors of individual 1 and those of individual 2 until time $t$ (while coalescences between ancestors of the same individual are allowed). In analogy with (16), the probability of no such coalescence in a single time step is bounded from below by

$$\tilde{q} := \Big(1 - \frac{n+1}{N}\Big)^{n+1} = 1 - \frac{(n+1)^2}{N} + \mathcal{O}(1/N^2) \tag{27}$$

since each individual has at most $|S| = n + 1$ ancestors. Thus

$$\mathbb{P}(\widetilde{\Omega}_t) \geqslant \tilde{q}^t = 1 - \frac{(n+1)^2 t}{N} + \mathcal{O}(1/N^2)$$

for every finite $t$ as $N \to \infty$. We now consider $\{\Sigma'_{\tau,1}, \Sigma'_{\tau,2} \mid \widetilde{\Omega}_{t-1}\}_{0\leqslant\tau\leqslant t-1}$. Arguing in a similar way as in the proof of Prop. 2, we find that the law of the conditional joint process agrees with that of two *independent* copies of $\{\Sigma'_\tau\}_{0\leqslant\tau\leqslant t-1}$ as long as there is no common ancestry between parts of $\Sigma'_{\tau,1}$ and those of $\Sigma'_{\tau,2}$. This is the case with probability $\mathbb{P}(\tilde{\Omega}_{t-1})$, which deviates from 1 by $\mathcal{O}(1/N)$, so that

$$\mathbb{P}(\Sigma'_{t-1,1} = \sigma, \Sigma'_{t-1,2} = \sigma \mid \widetilde{\Omega}_{t-1}) = \big(\mathbb{P}(\Sigma'_{t-1,1} = \sigma)\big)^2 + \mathcal{O}(1/N). \tag{28}$$

Next, on $\widetilde{\Omega}_{t-1}$, the parts of $\Sigma'_{t-1,1}$ and of $\Sigma'_{t-1,2}$ pick their types *independently*, so that

$$\mathbb{P}(\Xi_{t,1} = x, \Xi_{t,2} = x \mid \Sigma'_{t-1,1} = \sigma, \Sigma'_{t-1,2} = \sigma, \widetilde{\Omega}_{t-1}) = \big(\mathbb{P}(\Xi_{t,1} = x \mid \Sigma'_{t-1,1} = \sigma)\big)^2. \tag{29}$$

Taking together $1-\mathbb{P}(\widetilde{\Omega}_{t-1})=\mathcal{O}(1/N)$, (28) and (29), we get

$$\begin{aligned}\mathbb{P}(\Sigma'_{t-1,1}=\sigma,&\Sigma'_{t-1,2}=\sigma,\Xi_{t,1}=x,\Xi_{t,2}=x)\\ &=\mathbb{P}(\Sigma'_{t-1,1}=\sigma,\Sigma'_{t-1,2}=\sigma,\Xi_{t,1}=x,\Xi_{t,2}=x\mid\widetilde{\Omega}_{t-1})+\mathcal{O}(1/N)\\ &=\mathbb{P}(\Xi_{t,1}=x,\Xi_{t,2}=x\mid\Sigma'_{t-1,1}=\sigma,\Sigma'_{t-1,2}=\sigma,\widetilde{\Omega}_{t-1})\\ &\qquad\times\mathbb{P}(\Sigma'_{t-1,1}=\sigma,\Sigma'_{t-1,2}=\sigma\mid\widetilde{\Omega}_{t-1})+\mathcal{O}(1/N)\\ &=\big(\mathbb{P}(\Xi_{t,1}=x,\Sigma'_{t-1,1}=\sigma)\big)^2+\mathcal{O}(1/N),\end{aligned}$$

so that indeed

$$\mathrm{Cov}(\mathbb{1}\{\Sigma'_{t-1,1}=\sigma,\Xi_{t,1}=x\},\mathbb{1}\{\Sigma'_{t-1,2}=\sigma,\Xi_{t,2}=x\})=\mathcal{O}(1/N),$$

which establishes (26) and proves the claim. □

### 3.3 Connection with the deterministic dynamical system

A main result now is

**Theorem 3** *For all $G\subseteq L$ and all $\tau\geqslant 0$, we have*

$$\mathbb{P}(F_\tau=G)=a_G(\tau)\,.$$

We give two proofs that result in different and mutually complementary insight. The first uses a general argument, the second a concrete calculation.

*Proof (First proof of Theorem 3)* Compare the two laws of large numbers, Prop. 1 and Theorem 2. The claim is obvious via comparison of coefficients. The latter is justified by the following observation (cf. the argument in the proof of Theorem 3 in von Wangenheim et al. 2010). For generic $p_0$ and generic $X_i$, the vectors $R_G(p_0)$ with $G\subseteq L$ are the extremal vectors of the closed simplex $\mathrm{conv}\{R_K(p_0)\mid K\subseteq L\}$, where conv denotes the convex hull. They are the vectors that (generically) cannot be expressed as non-trivial convex combination within the simplex, and hence the vertices of the simplex (in cases with degeneracies, one reduces the simplex in the obvious way). □

*Proof (Second proof of Theorem 3)* We simply show that $a_G(\tau)$ and $\mathbb{P}(F_\tau=G)$ follow the same recursions (with the same initial values). To this end, recall that we have implied

$$\mathbb{P}(F_\tau=G)=\mathbb{P}(F_\tau=G\mid F_0=\varnothing).$$

We then decompose the $\tau+1$ time steps into the initial step, followed by an interval of $\tau$ steps (in the spirit of the Kolmogorov backward equation) to obtain

$$\begin{aligned}\mathbb{P}(F^{(L)}_{\tau+1}=G)&=\mathbb{P}(F^{(L)}_{\tau+1}=G\mid F^{(L)}_0=\varnothing)=\sum_{H\subseteq G}\mathbb{P}(F^{(L)}_1=H\mid F^{(L)}_0=\varnothing)\mathbb{P}(F^{(L)}_\tau=G\mid F^{(L)}_0=H)\\ &=\mathbb{P}(F^{(L)}_1=\varnothing\mid F^{(L)}_0=\varnothing)\mathbb{P}(F^{(L)}_\tau=G\mid F^{(L)}_0=\varnothing)\\ &\quad+\sum_{\alpha\in G}\mathbb{P}(F^{(L)}_1=\{\alpha\}\mid F^{(L)}_0=\varnothing)\mathbb{P}(F^{(L)}_\tau=G\mid F^{(L)}_0=\{\alpha\})\\ &=(1-\sum_{\alpha\in L}\varrho_\alpha)\mathbb{P}(F^{(L)}_\tau=G)+\sum_{\alpha\in G}\varrho_\alpha\mathbb{P}(F^{(L_{<\alpha})}_\tau=G_{<\alpha})\mathbb{P}(F^{(L_{>\alpha})}_\tau=G_{>\alpha})\,.\end{aligned}\tag{30}$$

In the last step, we have used $\mathbb{P}(F^{(L)}_\tau=G|F^{(L)}_0=\{\alpha\})=\mathbb{P}(F^{(L_{<\alpha})}_\tau=G_{<\alpha})\mathbb{P}(F^{(L_{>\alpha})}_\tau=G_{>\alpha})$, which is due to the conditional independence of segments according to Def. 1. Knowing Prop. 3 (for

$L' = L_{<\alpha}$ and hence $L \setminus L' = L_{>\alpha}$), the recursion (30) is identical to the one in (13); together with the identity of the initial conditions,

$$a_G^{(L)}(0) = \mathbb{P}(F_0^{(L)} = G) = \delta_{G,\varnothing},$$

this proves the claim. □

It is important to note that the second proof does not rely on Theorem 2. Theorem 3 could thus be used to establish Theorem 2 by simply invoking Prop. 1 and the deterministic time evolution (12). However, the independent proof of Theorem 2 bears the great advantage that it only requires the stochastic arguments derived in the current paper, thus making the argument self-contained and independent of the knowledge of the deterministic dynamics developed in previous work via quite a different toolbox.

### 3.4 Towards an explicit solution – preparation and example

Before we proceed, let us consider an important aspect of the segmentation process, namely, the probability that nothing happens in one time step given the current state is $G$. For any contiguous $\tilde{L} \subseteq L$, let

$$\lambda_G^{(\tilde{L})} := \mathbb{P}(F_{\tau+1}^{(\tilde{L})} = G \mid F_\tau^{(\tilde{L})} = G) = \prod_{I \in \mathcal{L}_G^{(\tilde{L})}} (1 - \sum_{\alpha \in I} \varrho_\alpha), \quad \text{for } G \subseteq \tilde{L}\,. \tag{31}$$

As before, we will omit the upper index in the case of $\tilde{L} = L$, that is, $\lambda_G^{(L)} = \lambda_G$. Since for every $I \in \mathcal{L}_G^{(\tilde{L})}$ one has $\mathcal{L}_\varnothing^{(I)} = \{I\}$, and thus $\lambda_\varnothing^{(I)} = 1 - \sum_{\alpha \in I} \varrho_\alpha$, we can rewrite (31) as

$$\lambda_G^{(\tilde{L})} = \prod_{I \in \mathcal{L}_G^{(\tilde{L})}} \lambda_\varnothing^{(I)}\,. \tag{32}$$

The coefficients $\lambda_G$ have already been identified by Bennett (1954), Lyubich (1992), and Dawson (2000, 2002), as well as by von Wangenheim et al. (2010) as the generalised eigenvalues of the linearised deterministic dynamics.

Our aim now is to find a closed-form expression for $\mathbb{P}(F_\tau = G)$ for all $\tau$. Clearly, $\mathbb{P}(F_\tau = G)$ is the sum over the probabilities for all paths that lead to the state $F_\tau = G$. Each path of the segmentation process may be represented by a tree, which we will call *ancestral recombination tree* or ART for short; we thus have to sum over the corresponding trees. Considering the trees carefully will be the key to the solution. Let us illustrate this by means of an example.

*Example 1* For four sites $S = \{0, 1, 2, 3\}$ with the corresponding links $L = \{\frac{1}{2}, \frac{3}{2}, \frac{5}{2}\}$, we consider $\mathbb{P}(F_\tau = \{\frac{1}{2}, \frac{3}{2}\})$, as illustrated in Figure 5. That is, we are concerned with all paths of the segmentation process that lead to $F_\tau = \{\frac{1}{2}, \frac{3}{2}\}$. The left tree captures the path where link $\frac{1}{2}$ is the first to be cut, the second that link $\frac{3}{2}$ is cut first. The $\lambda_G^j$ are the probabilities that nothing happens to any of the current segments for $j$ time steps. In the case where $\frac{3}{2}$ is the first event, the additional factor $\lambda_\varnothing^{(\{\frac{5}{2}\})} = (1 - \varrho_{\frac{5}{2}})$ is required to guarantee that at the time of the second segmentation event (at link $\frac{1}{2}$), the segment that belongs to link $\frac{5}{2}$ remains unchanged (the corresponding term in the other case is $\lambda_\varnothing^{(\varnothing)} = 1$). Finally, summing over all possible time combinations, one obtains

$$\begin{aligned}\mathbb{P}(F_\tau = \{\tfrac{1}{2}, \tfrac{3}{2}\}) = {} & \varrho_{\frac{1}{2}} \varrho_{\frac{3}{2}} \sum_{k=0}^{\tau-2} \lambda_\varnothing^k \sum_{i=0}^{\tau-2-k} \lambda_{\frac{1}{2}}^i \lambda_{\{\frac{1}{2}, \frac{3}{2}\}}^{\tau-2-k-i} \\ & + \varrho_{\frac{1}{2}} \varrho_{\frac{3}{2}} (1 - \varrho_{\frac{5}{2}}) \sum_{k=0}^{\tau-2} \lambda_\varnothing^k \sum_{i=0}^{\tau-2-k} \lambda_{\frac{3}{2}}^i \lambda_{\{\frac{1}{2}, \frac{3}{2}\}}^{\tau-2-k-i},\end{aligned} \tag{33}$$

where the first double sum belongs to the left, the second to the right tree of Figure 5. Unsurprisingly, the same result is obtained by explicitly solving (13) (by the method established by von Wangenheim et al. 2010), which demonstrates once more that $\mathbb{P}(F_\tau = \{\frac{1}{2}, \frac{3}{2}\}) = a_{\{\frac{1}{2},\frac{3}{2}\}}(\tau)$, in line with Theorem 3.

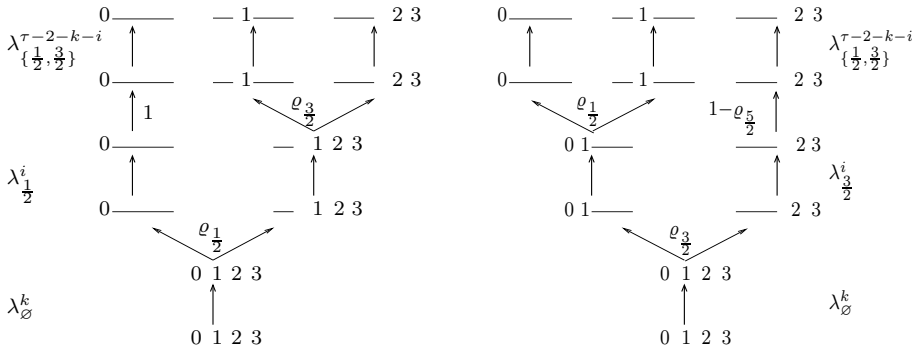


**Fig. 5** The two possible paths of the segmentation process of Example 1 that lead to $F_\tau = \{\frac{1}{2}, \frac{3}{2}\}$. The left panel refers to the first double sum of (33), the right one to the second.

$\mathbb{P}(F_\tau = G)$ may be understood as a sum over both tree topologies and branch lengths, i.e. we are concerned with all possible *ultrametric binary trees* that can be produced by the segmentation process. (The trees will be explained in more detail later. For the moment, recall that, in a binary tree, each internal node has at most two offspring nodes. An ultrametric tree is a tree whose branches are assigned lengths such that all leaves have the same distance from the root. For a review of metric trees, see Semple and Steel 2003, Chap. 7). In our case, the branch length corresponds to the number of time steps between consecutive nodes, and each internal node with two offspring nodes corresponds to a recombination event. We will now show that (and how) it is sufficient to deal with the corresponding *tree topologies* instead, which are obtained by contracting consecutive edges connected by a node with a single offspring into a *single* edge and removing the branch lengths. The result of this (many-to-one) operation is the topology of a *full binary tree*, that is, every internal node has exactly two offspring nodes (which may be internal nodes or leaves). The probability for each topology then is the sum of all probabilities of all the underlying original (ultrametric) trees, that is, the probability for all possible combinations of branch lengths (cf. the double sums in (33).) It will turn out that these sums may be evaluated explicitly for each topology, which is the reason that this approach is useful. For Example 1, this will (after a simple but lengthy calculation) result in

$$\begin{aligned}\mathbb{P}(F_\tau = \{\tfrac{1}{2}, \tfrac{3}{2}\}) &= \mathbb{P}(\text{tree 1}) + \mathbb{P}(\text{tree 2}) \\ &= \Big((\lambda^\tau_{\{\frac{1}{2},\frac{3}{2}\}} - \lambda^\tau_\varnothing)\frac{\varrho_{\frac{1}{2}}}{\lambda_{\{\frac{1}{2},\frac{3}{2}\}} - \lambda_\varnothing} - (\lambda^\tau_{\frac{1}{2}} - \lambda^\tau_\varnothing)\Big) \\ &\quad + \Big((\lambda^\tau_{\{\frac{1}{2},\frac{3}{2}\}} - \lambda^\tau_\varnothing)\frac{\varrho_{\frac{3}{2}}}{\lambda_{\{\frac{1}{2},\frac{3}{2}\}} - \lambda_\varnothing} - (\lambda^\tau_{\frac{3}{2}} - \lambda^\tau_\varnothing)\frac{\varrho_{\frac{3}{2}}}{\lambda_{\frac{3}{2}} - \lambda_\varnothing}\Big),\end{aligned} \tag{34}$$

where tree 1 refers to the left and tree 2 to the right panel in Figure 5.

### 3.5 Ancestral tree topologies

Our aim now is to assign probabilities to each of the possible topologies that have the elements of a given set $G$ as their internal nodes. Once the probabilities are known, $\mathbb{P}(F_\tau = G)$ is obtained by

summing over all compatible topologies. Let us begin with a suitable definition for our tree topologies (see Figure 6 for an illustration).

**Definition 2** For $\varnothing \neq G \subseteq L$, a *tree topology* is defined as $T := (G, m)$, where $G$ signifies the set of internal nodes, $\gamma \in G$ designates the *initial branching point* of the tree, and, in addition, $r$ is the *root*. The function $m$ is given by

$$\begin{aligned} m:\ G &\longrightarrow G \cup \{r\} \\ \alpha &\mapsto m(\alpha)\,, \end{aligned}$$

and $m(\alpha)$ denotes the (unique) ancestor of the internal node $\alpha \in G$. $m(\alpha)$ is an internal node except for $\alpha = \gamma$, where $m(\gamma) = r$. We will assume throughout that $m$ is *tree-consistent*, that is, the resulting structure is a full binary tree topology. For $G = \varnothing$, the only tree topology is the empty tree (with no internal nodes).

Thus, $T$ has the internal nodes $\alpha \in G$ and the set of internal edges

$$\{(m(\alpha), \alpha) \mid \alpha \in G, m(\alpha) \neq r\},$$

as well the external edge $(r, \gamma)$. Note that we have not included the (external) leaves (and corresponding external edges) in our definition since they will never be required explicitly. Note also that, in the context of phylogeny, the (canonical) root of the tree is what we call initial branching point, cf. Semple and Steel (2003). For an excellent account of terminology and properties of trees, see Gross and Yellen (1999, Chap. 3).

We will use the standard partial order on $T$, namely, $\alpha \preccurlyeq \beta$ means that $\alpha$ is on the path from $\gamma$ to $\beta$, i.e. $\alpha = m^i(\beta)$ for some $i \in \{0, \ldots, |G|\}$. Obviously, $\gamma$ is the minimal element of $G$ with respect to $\preccurlyeq$. Furthermore, $\alpha \prec \beta$ means that $\alpha \preccurlyeq \beta$ with $\alpha \neq \beta$.

Note that the topology $T$ as such does not depend on $L$ except via the requirement $G \subseteq L$; it may likewise represent a realisation of a process $\{F_s^{(\tilde{L})}\}_{0 \leqslant s \leqslant \tau}$, restricted to a (contiguous) subset $\tilde{L}$ of $L$, provided $G \subseteq \tilde{L}$. If we also specify the set of links, say $\tilde{L}$, then each edge of a given tree topology can be associated with a particular segment. Namely, for $T = (G, m)$ and $\alpha \in G$, we associate with the edge $(m(\alpha), \alpha)$ the segment

$$I_\alpha^{(\tilde{L})}(T) := K \in \mathcal{L}^{(\tilde{L})}_{\{\beta : \beta \prec \alpha\}} \ \text{ s.t. } \alpha \in K\,.$$

In words, $I_\alpha^{(\tilde{L})}(T)$ is the segment that will receive its next cut at link $\alpha$ (given the topology $T$). In particular, $I_\gamma^{(\tilde{L})}(T) = \tilde{L}$ (independently of $T$). An example is given in Figure 6. From now on we will suppress the dependence on $T$ and $\tilde{L}$ throughout and write $I_\alpha$ instead of $I_\alpha^{(\tilde{L})}(T)$. Next, we define *subtrees*.

**Definition 3 (Subtrees and subtree decomposition)** Consider $T = (G, m)$ with $\varnothing \neq G \subseteq L$. Then, for any $\gamma \in H \subseteq G$ and $\alpha \in G$, a subtree of $T$ is defined via $T_\alpha(H) = (G_\alpha(H), m|_{G_\alpha(H)})$, where

$$G_\alpha(H) := \{\beta \in G | \alpha \preccurlyeq \beta \ \text{ and } h \not\prec \beta \quad \forall\, h \in H \ \text{ with } \alpha \prec h\}\,,$$

and $m|_{G_\alpha(H)}$ is the restriction of $m$ to $G_\alpha(H)$. Specifically, we set $m|_{G_\alpha(H)}(\alpha) =: r_\alpha$ for the initial branching point of the respective subtree (so that $r = r_\gamma$ for consistency). The collection $\{T_\beta(H)\}_{\beta \in H}$ describes a *decomposition* of $T$ into subtrees, where $T_\beta(H)$ has initial branching point $\beta$ and internal nodes $G_\beta(H)$.

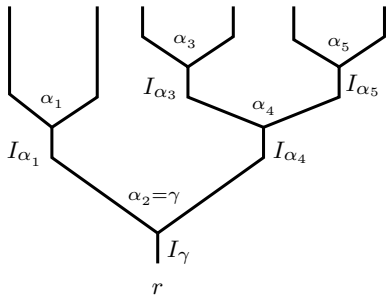


**Fig. 6** Example of a tree topology $T$ for $G = \{\alpha_1, \ldots, \alpha_5\}$, with root $r$ and initial branching point $\gamma = \alpha_2$. Each (internal and root) edge is identified with a certain segment. Here, we have $I_{\alpha_1} = L_{<\gamma}$, $I_{\alpha_2} = I_\gamma = L$, $I_{\alpha_3} = \{\gamma+1, \ldots, \alpha_3, \ldots, \alpha_4 - 1\}$, $I_{\alpha_4} = L_{>\gamma}$ and $I_{\alpha_5} = L_{>\alpha_4}$.

Intuitively, the decomposition is obtained by 'cutting the tree below each element of $H$'. The tree then disintegrates into the subtrees $\{T_\beta(H)\}_{\beta\in H}$, and each element of $H$ appears as the initial branching point of one of the subtrees; Figure 7 provides an example. The $T_\alpha(H), \alpha \in G \setminus H$, are, in turn, subtrees of these subtrees; they will also be required in what follows.

Obviously, $G_\alpha(H)$ depends on the topology $T$ (via the partial order), but again we omit this for economy of notation. Note that the subtrees inherit the segments from the original tree, i.e., $I_\beta(T_\alpha(H)) = I_\beta(T) = I_\beta$. Let us mention that similar subtree decompositions appear in the context of molecular phylogeny, for example, *Tuffley's poset* (Gill et al. 2008).

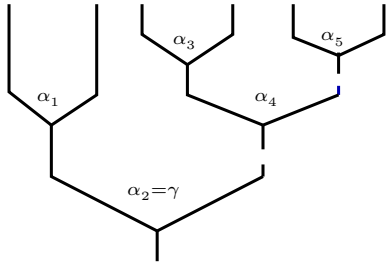


**Fig. 7** Decomposition of a tree topology into three subtrees via $H = \{\gamma, \alpha_4, \alpha_5\}$.The subtrees are labelled with their initial branching points, so the decomposition consists of $T_\gamma(H)$, $T_{\alpha_4}(H)$, and $T_{\alpha_5}(H)$, with node sets $G_\gamma(H) = \{\alpha_1, \gamma\}$, $G_{\alpha_4}(H) = \{\alpha_3, \alpha_4\}$, and $G_{\alpha_5}(H) = \{\alpha_5\}$, respectively.

### 3.6 ART probabilities and explicit solution of the segmentation process

Let us now assign probabilities to tree topologies. To this end, consider the *augmented segmentation process* $\{\widetilde{F}_\tau\}_{\tau\in\mathbb{N}_0}$ with values in the set of all possible tree topologies $T = (G, m)$ (rather than the sets $G$ alone); $\widetilde{F}_\tau = (G, m)$ means that $F_\tau = G$, and the segmentation events have occured according to the partial order implied by $m$ (as in (34) of Example 1). We will abbreviate $\mathbb{P}(\widetilde{F}_\tau = T)$ as $\mathbb{P}_\tau(T)$. Let us now state the central result for these tree probabilities.

**Theorem 4 (ART probabilities)** *Under the segmentation process, the probability for the tree topology $T = (G, m)$ at time $\tau$ is given by $\mathbb{P}_\tau(T) = (\lambda_\varnothing^{(L)})^\tau$ for $G = \varnothing$, and, for $\varnothing \neq G \subseteq L$ and initial branching*

*point $\gamma \in G$, by*

$$\mathbb{P}_\tau(T) = \sum_{\gamma \in H \subseteq G} (-1)^{|H|-1} \left[ (\lambda^{(L)}_{G_\gamma(H)})^\tau - (\lambda^{(L)}_{\varnothing})^\tau \right] f(T,H). \tag{35}$$

*Here, for $H \subseteq G$,*

$$f(T,H) := \prod_{\alpha \in G} g(T_\alpha(H)) \tag{36}$$

*with*

$$g(T_\alpha(H)) := \frac{\varrho_\alpha}{\lambda^{(I_\alpha)}_{G_\alpha(H)} - \lambda^{(I_\alpha)}_{\varnothing}} \quad \text{for } \alpha \in G,\ H \subseteq G\,. \tag{37}$$

*Remark 2* A few remarks are in order:

1. Note that the dependence on $\tau$ is solely due to the term in square brackets in (35), whereas $f$ is independent of time.
2. The same term implies that $\mathbb{P}_0((G,m)) = 0$ for all $G \neq \varnothing$ and all tree-consistent mappings $m$.
3. The $g(T_\alpha(H))$ are well-defined and strictly positive since $G_\alpha(H) \neq \varnothing$ implies that $\lambda_{G_\alpha(H)} > \lambda_\varnothing$ (cf. Lemma 5 in von Wangenheim et al. 2010).
4. Eq. (35) implies a sum over all possible subtree decompositions of $T$; for every given decomposition, the subtree containing $\gamma$ plays a special role.
5. We have implied here that the underlying set of links is $L$, i.e., $\mathbb{P}_\tau(T) = \mathbb{P}(\widetilde{F}^{(L)}_\tau = T)$. However, due to Prop. 3, the result carries over if $L$ is replaced by any contiguous $\tilde{L} \subseteq L$ that contains $G$.

Before we embark on the proof, let us briefly comment on the general strategy. Like every Markov chain, the segmentation process can be viewed in forward or in backward direction (with respect to the time increment on the $\tau$ time scale): If the increment is at the end of the time interval, then the corresponding ultrametric tree grows at its top (i.e. the external branches are extended or split up); otherwise it grows at its base (i.e. the root branch is extended or the two corresponding subtrees coalesce). Where the original formulation (Definition 1) is in the bottom-up direction, the advantage of the top-down approach is that one only has to deal with two objects in every step, namely the left and the right subtrees that emerge via the first segmentation event on the $\tau$ timescale (and that are joined when looking back), instead of a possibly large number of smaller segments at the top. This point of view has already been used in the proof of Theorem 3 and will again serve in the following proof.

*Proof (of Theorem 4.)* We will prove the claim via induction in the top-down direction by progressively merging pairs of subtrees. To do so, we first need some properties related to the corresponding tree decomposition, see Figure 8. Consider a tree topology $T = (G,m)$, with $\varnothing \neq G \subseteq L$. The initial branching point of $T$ is $\gamma \in G$ as before. If $\gamma$ has two (internal) offspring nodes, these are denoted by $\gamma' \in G_{<\gamma} \subseteq L_{<\gamma}$ and $\gamma'' \in G_{>\gamma} \subseteq L_{>\gamma}$. We then define the left subtree $T'$ of $T$ as $T' = (G_{<\gamma}, m|_{G_{<\gamma}})$ and analogously the right subtree $T''$ as $T'' = (G_{>\gamma}, m|_{G_{>\gamma}})$, both obviously with fewer nodes than $T$. For convenience, we will denote $L_{<\gamma}$ ($L_{>\gamma}$) by $L'$ ($L''$) and the respective nodes by $G' = G_{<\gamma}$ ($G'' = G_{>\gamma}$). If $\gamma$ has no or only one offspring node, we take the empty tree (where nothing happens) as left and/or right subtree. In any case, the offspring nodes of $\gamma$ are specified through the preimage of $\gamma$ under the function $m$, i.e. $m^{-1}(\gamma) \in \{\{\varnothing\}, \{\gamma'\}, \{\gamma''\}, \{\gamma', \gamma''\}\}$. $T$ is then obtained by joining these subtrees together at $\gamma$. In terms of the segmentation process, this corresponds to the very first cut (of $L$, at link $\gamma$). Since this may happen at any time $j \in \{1, \ldots, \tau\}$, (i.e. the root branch lasts for $i = j-1 \in \{0, \ldots, \tau-1\}$ times while $T'$ and $T''$ apply for the remaining $\tau - 1 - i$ time steps), it is clear that

$$\mathbb{P}_\tau(T) = \sum_{i=0}^{\tau-1} \lambda^i_\varnothing \varrho_\gamma \mathbb{P}_{\tau-1-i}(T') \mathbb{P}_{\tau-1-i}(T''). \tag{38}$$

Before we can evaluate (38), we need three preparatory results.

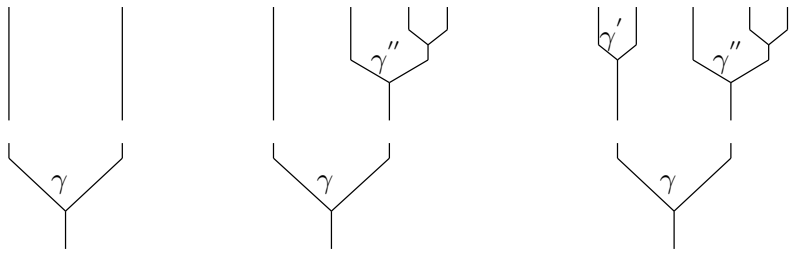


**Fig. 8** Joining together the left and the right subtrees ($T'$ and $T''$) at the initial branching point $\gamma$. Three different cases arise, depending on whether one or both of the subtrees are empty trees.

(A) *Product rule for the $\lambda$'s*

For all $L = L' \cup L'' \cup \{\gamma\}$ and all $G' \subseteq L'$, $G'' \subseteq L''$, we have

$$\lambda_{G'}^{(L')} \cdot \lambda_{G''}^{(L'')} = \lambda_{G' \cup G'' \cup \{\gamma\}}^{(L)} , \tag{39}$$

which follows immediately from (32).

(B) *Product rule for $f$*

For all $H \subseteq G$ with $H' := G' \cap H$ and $H'' := G'' \cap H$, one has

$$\begin{aligned} G_\alpha(H) &= G_\alpha(H') = G'_\alpha(H') \quad \text{for all } \alpha \in G' \text{ and} \\ G_\alpha(H) &= G_\alpha(H'') = G''_\alpha(H'') \quad \text{for all } \alpha \in G'' , \end{aligned} \tag{40}$$

so that consequently

$$\begin{aligned} T_\alpha(H) &= T'_\alpha(H') \quad \text{for all } \alpha \in G' \text{ and} \\ T_\alpha(H) &= T''_\alpha(H'') \quad \text{for all } \alpha \in G'' . \end{aligned} \tag{41}$$

This then leads to the following product rule for the function $f$ from (36):

$$\begin{aligned} f(T,H) &= g(T_\gamma(H)) \cdot \prod_{\alpha \in G'} g(T_\alpha(H)) \prod_{\beta \in G''} g(T_\beta(H)) \\ &= g(T_\gamma(H)) \cdot \prod_{\alpha \in G'} g(T'_\alpha(H')) \prod_{\beta \in G''} g(T''_\beta(H'')) = g(T_\gamma(H)) f(T', H') f(T'', H'') . \end{aligned} \tag{42}$$

Note that in case $G' = \varnothing$ or $G'' = \varnothing$, the corresponding empty product is 1 as usual.

(C) *Assembly of initial branching point and subtrees*

As an immediate consequence of Definition 3, one obtains for all $C \subseteq m^{-1}(\gamma) \subseteq H \subseteq G$:

$$\{\gamma\} \cup \bigcup_{\eta \in C} G_\eta(H) = G_\gamma(H \setminus C) = G_\gamma((H \cup \{\gamma\}) \setminus C) . \tag{43}$$

(In fact, this relationship is not restricted to $\gamma$ but also holds for arbitrary $\alpha \in G$, in an analogous way; but this will not be used in what follows.)

(D) *Subtree summation*

Recall the classic identity

$$\sum_{i=0}^{n} a^i b^{n-i} = \begin{cases} \frac{b^{n+1}-a^{n+1}}{b-a}\,, & a \neq b, \\ (n+1)a^n\,, & a = b, \end{cases} \tag{44}$$

where the second case is also the l'Hôpital limit of the first. Together with (37), this implies

$$\begin{aligned} \varrho_\alpha \sum_{i=0}^{\tau-1} (\lambda_{\varnothing}^{(I_\alpha)})^i (\lambda_{G_\alpha(H)}^{(I_\alpha)})^{\tau-i-1} &= \frac{\varrho_\alpha}{\lambda_{G_\alpha(H)}^{(I_\alpha)} - \lambda_{\varnothing}^{(I_\alpha)}} \cdot \big( (\lambda_{G_\alpha(H)}^{(I_\alpha)})^\tau - (\lambda_{\varnothing}^{(I_\alpha)})^\tau \big) \\ &= g\big(T_\alpha(H)\big) \big( (\lambda_{G_\alpha(H)}^{(I_\alpha)})^\tau - (\lambda_{\varnothing}^{(I_\alpha)})^\tau \big) \end{aligned} \tag{45}$$

for all $T = (G, m)$, $\varnothing \neq G \subseteq L$, $\alpha \in G$ and $H \subseteq L$.

We now continue the proof of the theorem and proceed via induction over $|G|$. For $G = \varnothing$, the claim holds trivially, that is, $\mathbb{P}_\tau(T) = \lambda_{\varnothing}^\tau$ for all $\tau \geqslant 0$. For $G \subseteq L$ with $|G| = 1$, i.e. $G = \{\gamma\}$, both the left and the right subtrees are empty trees (see the left case in Figure 8). Using first (38), then the result for $G' = G'' = \varnothing$ on $L'$ and $L''$, respectively (see Remark 2 (5)), then (39), and finally (45), we obtain

$$\begin{aligned} \mathbb{P}_\tau(T) &= \varrho_\gamma \sum_{i=0}^{\tau-1} \lambda_{\varnothing}^i \mathbb{P}_{\tau-1-i}(T') \mathbb{P}_{\tau-1-i}(T'') = \varrho_\gamma \sum_{i=0}^{\tau-1} \lambda_{\varnothing}^i (\lambda_{\varnothing}^{(L')})^{\tau-1-i} (\lambda_{\varnothing}^{(L'')})^{\tau-1-i} \\ &= \varrho_\gamma \sum_{i=0}^{\tau-1} \lambda_{\varnothing}^i \lambda_\gamma^{\tau-1-i} = g\big(T_\gamma(\{\gamma\})\big)(\lambda_\gamma^\tau - \lambda_{\varnothing}^\tau)\,. \end{aligned}$$

We now assume the claim to hold for all tree topologies $T = (G, m)$ for all $G \subseteq L$ with $|G| \leqslant k$ for some $k \geqslant 1$; by Remark 2, it then holds likewise with $L$ replaced by a contiguous $\tilde{L} \subseteq L$ as long as $G \subseteq \tilde{L}$. We next turn to $G = \{\alpha_1, \ldots, \alpha_{k+1}\} \subseteq L$ and fixed $T = (G, m)$; recall our convention $\alpha_1 < \alpha_2 < \ldots < \alpha_{k+1}$. In the following, we will always write $A \cup \alpha := A \cup \{\alpha\}$ and $A \setminus \alpha := A \setminus \{\alpha\}$ for $A \subseteq L$ and $\alpha \in L$. We have to distinguish the two remaining cases in Figure 8 (middle and right):

*Case* $\gamma = \alpha_1$

Then $T^{'}$ is an empty tree while $T^{''}$ has initial branching point $\gamma^{''} \in G^{''}$, i.e. $m^{-1}(\gamma) = \{\gamma^{''}\}$. We then find

$$
\begin{aligned}
\mathbb{P}_\tau(T) &= \varrho_\gamma \sum_{i=0}^{\tau-1} \lambda_\varnothing^i \mathbb{P}_{\tau-1-i}(T^{'}) \mathbb{P}_{\tau-1-i}(T^{''}) \\
&= \varrho_\gamma \sum_{i=0}^{\tau-1} \lambda_\varnothing^i (\lambda_\varnothing^{(L^{'})})^{\tau-1-i} \sum_{\gamma'' \in H'' \subseteq G''} (-1)^{|H^{''}|-1} \Big( (\lambda^{(L^{''})}_{G''_{\gamma''}(H'')})^{\tau-1-i} - (\lambda_\varnothing^{(L^{''})})^{\tau-1-i} \Big) f(T^{''}, H^{''}) \\
&= \varrho_\gamma \sum_{i=0}^{\tau-1} \lambda_\varnothing^i \sum_{\gamma'' \in H'' \subseteq G''} (-1)^{|H^{''}|-1} \big( \lambda^{\tau-1-i}_{\gamma \cup G''_{\gamma''}(H'')} - \lambda_\gamma^{\tau-1-i} \big) f(T^{''}, H^{''}) \\
&= \varrho_\gamma \sum_{i=0}^{\tau-1} \lambda_\varnothing^i \sum_{\gamma'' \in H'' \subseteq G''} (-1)^{|H^{''}|-1} \big( \lambda^{\tau-1-i}_{G_\gamma((H'' \cup \gamma) \setminus \gamma'')} - \lambda^{\tau-1-i}_{G_\gamma(H'' \cup \gamma)} \big) f(T^{''}, H^{''}) \\
&= \sum_{\gamma'' \in H'' \subseteq G''} (-1)^{|H^{''}|-1} g(T_\gamma((H^{''} \cup \gamma) \setminus \gamma^{''})) \big( \lambda^\tau_{G_\gamma((H'' \cup \gamma) \setminus \gamma'')} - \lambda^\tau_\varnothing \big) f(T^{''}, H^{''}) \\
&\quad - \sum_{\gamma'' \in H'' \subseteq G''} (-1)^{|H^{''}|-1} g(T_\gamma(H^{''} \cup \gamma)) \big( \lambda^\tau_{G_\gamma(H'' \cup \gamma)} - \lambda^\tau_\varnothing \big) f(T^{''}, H^{''}) \\
&= \sum_{\gamma'' \in H'' \subseteq G''} (-1)^{|H^{''}|-1} \big( \lambda^\tau_{G_\gamma((H'' \cup \gamma) \setminus \gamma'')} - \lambda^\tau_\varnothing \big) f(T, (H^{''} \cup \gamma) \setminus \gamma^{''}) \\
&\quad - \sum_{\gamma'' \in H'' \subseteq G''} (-1)^{|H^{''}|-1} \big( \lambda^\tau_{G_\gamma(H'' \cup \gamma)} - \lambda^\tau_\varnothing \big) f(T, H^{''} \cup \gamma) \\
&= \sum_{\gamma \in H \subseteq G \setminus \gamma''} (-1)^{|H|-1} \big( \lambda^\tau_{G_\gamma(H)} - \lambda^\tau_\varnothing \big) f(T, H) - \sum_{\{\gamma, \gamma''\} \subseteq H \subseteq G} (-1)^{|H|} \big( \lambda^\tau_{G_\gamma(H)} - \lambda^\tau_\varnothing \big) f(T, H) \\
&= \sum_{\gamma \in H \subseteq G} (-1)^{|H|-1} \big( \lambda^\tau_{G_\gamma(H)} - \lambda^\tau_\varnothing \big) f(T, H) \,.
\end{aligned}
$$

In the first step, we have used (38), in the second the induction hypothesis (applied to $T''$ on $L''$), in the third the product structure of the $\lambda$'s (39), in the fourth (40) (read in the backward direction) and (43) (applied to $H = H''$ with $C = \{\gamma''\}$ and $C = \varnothing$, respectively). In the fifth step, we have invoked (45) (separately on each term in parentheses), in the sixth step we have used (42) with $H = (H^{''} \cup \gamma) \setminus \gamma^{''}$ and $H = H^{''} \cup \gamma$, respectively, and finally we have changed the summation variable (where we set $G = \gamma \cup G^{''}$).

*Case* $\gamma \in \{\alpha_2, \ldots, \alpha_k\}$

Now we have to consider the left subtree $T^{'}$ with initial branching point $\gamma^{'}$ and the right subtree $T^{''}$ with initial branching point $\gamma^{''}$, i.e. $m^{-1}(\gamma) = \{\gamma^{'}, \gamma^{''}\}$. Proceeding in analogy with the previous

case, we obtain

$$
\begin{aligned}
\mathbb{P}_\tau(T) &= \sum_{i=0}^{\tau-1} \lambda_\varnothing^i \varrho_\gamma \mathbb{P}_{\tau-1-i}(T^{'}) \mathbb{P}_{\tau-1-i}(T^{''}) \\
&= \sum_{i=0}^{\tau-1} \lambda_\varnothing^i \varrho_\gamma \sum_{\gamma^{'} \in H^{'} \subseteq G^{'}} (-1)^{|H^{'}|-1} \Big( (\lambda^{(L^{'})}_{G^{'}_{\gamma^{'}}(H^{'})})^{\tau-1-i} - (\lambda_\varnothing^{(L^{'})})^{\tau-1-i} \Big) f(T^{'}, H^{'}) \\
&\quad \times \sum_{\gamma^{''} \in H^{''} \subseteq G^{''}} (-1)^{|H^{''}|-1} \Big( (\lambda^{(L^{''})}_{G^{''}_{\gamma^{''}}(H^{''})})^{\tau-1-i} - (\lambda_\varnothing^{(L^{''})})^{\tau-1-i} \Big) f(T^{''}, H^{''}) \\
&= \sum_{i=0}^{\tau-1} \lambda_\varnothing^i \varrho_\gamma \sum_{\gamma^{'} \in H^{'} \subseteq G^{'}} \sum_{\gamma^{''} \in H^{''} \subseteq G^{''}} (-1)^{|H^{'}|+|H^{''}|} \Big( \lambda^{\tau-1-i}_{G_\gamma((H^{'} \cup H^{''} \cup \gamma) \setminus \{\gamma^{'}, \gamma^{''}\})} \\
&\quad - \lambda^{\tau-1-i}_{G_\gamma((H^{'} \cup H^{''} \cup \gamma) \setminus \gamma^{''})} - \lambda^{\tau-1-i}_{G_\gamma((H^{'} \cup H^{''} \cup \gamma) \setminus \gamma^{'})} + \lambda^{\tau-1-i}_{G_\gamma(H^{'} \cup H^{''} \cup \gamma)} \Big) f(T^{'}, H^{'}) f(T^{''}, H^{''}) \\
&= \sum_{\gamma^{'} \in H^{'} \subseteq G^{'}} \sum_{\gamma^{''} \in H^{''} \subseteq G^{''}} (-1)^{|H^{'}|+|H^{''}|} \Big( g(T_\gamma((H^{'} \cup H^{''} \cup \gamma) \setminus \{\gamma^{'}, \gamma^{''}\})) (\lambda^\tau_{G_\gamma((H^{'} \cup H^{''} \cup \gamma) \setminus \{\gamma^{'}, \gamma^{''}\})} - \lambda^\tau_\varnothing) \\
&\quad - g(T_\gamma((H^{'} \cup H^{''} \cup \gamma) \setminus \{\gamma^{''}\})) (\lambda^\tau_{G_\gamma((H^{'} \cup H^{''} \cup \gamma) \setminus \{\gamma^{''}\})} - \lambda^\tau_\varnothing) \\
&\quad - g(T_\gamma((H^{'} \cup H^{''} \cup \gamma) \setminus \{\gamma^{'}\})) (\lambda^\tau_{G_\gamma((H^{'} \cup H^{''} \cup \gamma) \setminus \{\gamma^{'}\})} - \lambda^\tau_\varnothing) \\
&\quad + g(T_\gamma(H^{'} \cup H^{''} \cup \gamma)) (\lambda^\tau_{G_\gamma(H^{'} \cup H^{''} \cup \gamma)} - \lambda^\tau_\varnothing) \Big) f(T^{'}, H^{'}) f(T^{''}, H^{''}) \\
&= \sum_{\gamma \in H \subseteq G \setminus \{\gamma^{'}, \gamma^{''}\}} (-1)^{|H|-1} (\lambda^\tau_{G_\gamma(H)} - \lambda^\tau_\varnothing) f(T, H) - \sum_{\{\gamma, \gamma^{'}\} \subseteq H \subseteq G \setminus \{\gamma^{''}\}} (-1)^{|H|} (\lambda^\tau_{G_\gamma(H)} - \lambda^\tau_\varnothing) f(T, H) \\
&\quad - \sum_{\{\gamma, \gamma^{''}\} \subseteq H \subseteq G \setminus \{\gamma^{'}\}} (-1)^{|H|} (\lambda_{G_\gamma(H)})^\tau - \lambda^\tau_\varnothing) f(T, H) + \sum_{\{\gamma, \gamma^{'}, \gamma^{''}\} \subseteq H \subseteq G} (-1)^{|H|-1} (\lambda^\tau_{G_\gamma(H)} - \lambda^\tau_\varnothing) f(T, H) \\
&= \sum_{\gamma \in H \subseteq G} (-1)^{|H|-1} \big( \lambda^\tau_{G_\gamma(H)} - \lambda^\tau_\varnothing \big) f(T, H) \,.
\end{aligned}
$$

The remaining case $\gamma = \alpha_{k+1}$ is analogous to $\gamma = \alpha_1$. □

Now that we have an explicit formula for the probability of our tree topologies, let us further comment on its structure.

*Remark 3* Using the subtree decomposition of Def. 3, we can state (36) alternatively as

$$
f(T, H) = \prod_{\alpha \in H} \prod_{\beta \in G_\alpha(H)} g(T_\beta(H)) \,, \tag{46}
$$

which implies some kind of independence across subtrees.

Returning to the segmentation process, we may conclude that

$$
\mathbb{P}(F_\tau = G) = \sum_m \mathbb{P}(\tilde{F}_\tau = (G, m)),
$$

where the sum is over all mappings consistent with a full binary tree. The final result then follows directly from Theorem 3.

**Corollary 1 (Solution of recombination equation via ARTs)** *The discrete-time recombination equation* (5) *has the solution*

$$
p_t = \sum_{G \subseteq L} a_G(t) R_G(p_0),
$$

*where*

$$a_G(t) = \sum_m \mathbb{P}(\tilde{F}_t = (G, m)) \tag{47}$$

*for all $G \subseteq L$, where the sum is over all tree-consistent $m$, with $\mathbb{P}(\tilde{F}_t = (G, m)) = \mathbb{P}_t(T)$ as given in Theorem 4.* □

## 4 Discussion

The piece of research presented here has solved the long-standing deterministic single-crossover dynamics forward in time by considering the corresponding stochastic process (the Wright-Fisher model with recombination) backward in time, via looking at single individuals and tracing back their ancestries. We will first compare the result with the previous recursive approaches and then turn to the connection with the ancestral recombination graph (ARG; the usual approach to recombination in *finite* populations, see Wakeley 2008, Chap. 7.2 and Durrett 2008, Chap. 3.4).

Evaluating the coefficients $a_G(t)$ via the traditional recursive approaches is an algebraic strategy, which relies on linearisation and diagonalisation of the underlying (forward, deterministic) dynamical system. In contrast, the ART approach presented here starts from a summation over all paths of the (backward, stochastic) process that give rise to a given set of segments after $t$ generations; this is reflected in the sum over all ultrametric trees, as in (33). This formula contains sums over all tree toplogies *and* over all combinations of branch lengths, which is not useful in itself, in particular for large $t$. The simplification obtained here consists in carrying out the summation over the branch lengths, so that one is left with the tree topologies only. This is particularly useful for large $t$ and small recombination probabilities, since long branches (where nothing happens) are 'contracted' in this way.

Both the recursive and the ART solution are of similar computational complexity. Wheter $a_G(t)$ is evaluated via ARTs (Corollary 1 and Theorem 4) or by solving the recursions (e.g., von Wangenheim et al. 2010, Eqs. (43)–(45) or the related ones of Dawson 2000, 2002), the effort grows exponentially with $n := |G|$. In the recursions, the complexity comes from multiple sums over nested sets of subsets of $G$. In the case of the ARTs, it is due to the summation over all possible tree topologies with internal node set $G$; their number grows exponentially with $n$. To be precise, there are $C(n)$ such topologies (Gross and Yellen 1999, Chap. 3.4; Stanley 1999, Ex. 6.19.d), where $C(n)$ is the $n$'th Catalan number. After all, clever algorithms are available for the generation and enumeration of these trees (Gupta 1992; Proskurowski 1980; Zaks 1980).

The ART formula is therefore not superior in computational terms. However, it bears the great advantage to relate to objects with an immediate meaning in terms of the underlying process. After all, the probabilities for the tree topologies may lend themselves to future use if, for example, one is interested in the distribution of tree shape(s), or if mutation is included in the model (that is, superimposed on the trees). This is in contrast with the manifestly non-intuitive recursions, for which we are not aware of an interpretation in terms of the underlying process.

The backward approach that gives rise to the ARTs differs from the ARG in two ways. First, we let $N$ tend to infinity without rescaling any parameters; that is, recombination probabilities remain constant when $N \to \infty$. This corresponds to the assumption that recombination is so strong (that is, loci are so far apart) that the majority of the recombination events takes place before coalescence sets in. In contrast, the ARG assumes weak recombination (in that recombination parameters scale inversely with population size), so that recombination and coalescence take place on the same time scale. Second, we focus on the ancestry of single individuals rather than of samples, which further simplifies matters. As a reward, one obtains semi-explicit answers for all quantities of interest here.

The differences between the scopes of the two approaches can be illustrated nicely in the context of the paper by Wiuf and Hein (1997), who analyse the ancestry of the genetic material of an entire chromosome from a single individual. More precisely, they investigate the partitioning process (a close relative of our $\{\Sigma_\tau\}_{\tau\geqslant 0}$) *at stationarity*, i.e., for $\tau \to \infty$, in the diffusion limit. They employ approximations and simulations and focus on the concrete example of the human chromosome 1, with realistic estimates of the recombination parameters and the effective population size. They do find large contributions from unordered partitions, in the sense of the frequent occurrence of segments of ancestral material interspersed with nonancestral ('trapped') material between them. This is a consequence of the large number of genetic ancestors of the chromosome (estimated at 6800) relative to the effective population size ($N_e = 20000$). (For loci at opposite ends of the chromosome, the diffusion limit is, however, not expected to yield a good approximation, see below.)

In contrast, our approach does not aim at any stationary situation. Rather, it should provide a faithful picture in situations where recombination rates and population size are large enough so that there is a time horizon governed by recombination alone. The detailed time course of individual ancestries over this time horizon will then be described by the segmentation process. Let us note that a well-known method of simulating the ARG (the so-called sequential Markov coalescent by McVean and Cardin 2005) also uses an approximation of the partitioning process by the segmentation process in that coalescence of sequences that both carry ancestral material are neglected; the authors demonstrate that this yields a good approximation to the full ARG over a wide range of parameters.

To be more precise, we stipulate that there are, in fact, three scaling regimes to be considered in the context of recombination: weak recombination, strong recombination, and free recombination. We can certainly not delineate them precisely here (and they are also expected to overlap). However, with the usual assumption of a recombination probability of the order of $10^{-8}$ per generation and base pair of a DNA sequence (Kauppi et al. 2004), it is clear that sites at a distance of the order of $10^3$–$10^4$ base pairs will fall into the weak recombination regime: The recombination probability between them, of $10^{-5}$–$10^{-4}$, is of the same order as the coalescence probability of $1/N$ (between any pair of branches per generation). In contrast, sites at a distance of $10^8$ base pairs (like the opposing ends of a chromosome) will be essentially independent since there will be, on average, one crossover between them in every generation; this is the case of free recombination. Between the extremes, at a distance of $10^6$ base pairs, say, the recombination probability of $10^{-2}$ between a pair of sites is well-separated from the coalescence probability of $1/N$ per pair of branches for all but the smallest populations; this should be a case for strong recombination over a time horizon where the number of branches is not too large. Note that the 'number of branches' that counts here is the number of genetic ancestors of an individual (that is, those that carry ancestral material), not the number of genealogical ancestors (that is, all parents, grandparents and so on). The number of genetic ancestors only increases roughly linearly with (backward) time $\tau$, whereas an individual has up to $2^\tau$ genealogical ancestors $\tau$ generations back; see the discussions by Donelly (1983) and Ralph and Coop (2013). Note also that, in this parameter regime, the assumption of at most one crossover per generation is well justified, while recombination is still strong relative to coalescence. Last not least, it is worth mentioning that this is the parameter regime where the deterministic dynamics yields a valid description.

The results presented here should also pave the way towards a solution of the multiple crossover model. Biologically, this is relevant when more distant loci are considered. Recombination in a given generation will again be described by a partition of $S$ into two parts (corresponding to the two parents), but, this time, these partitions may be arbitrary (as opposed to the ordered partitions that arise due to single crossovers). But we expect that the corresponding ancestral recombination trees will continue to be binary trees whose subtrees are conditionally independent, so that the methods developed here may be generalised to this case involving arbitrary partitions.

**Acknowledgements** It is our pleasure to thank the 'recombination seminar' (M. Baake, C. Huck, T. Hustedt, P. Zeiner) for thorough discussion of all stages of this research, and A. Wakolbinger for an enlightening conversation. We are grateful to M. Baake, K. Schneider, M. Esser, and S. Probst for critically reading the manuscript, and to an anonymous referee for valuable suggestions. This work was supported by Deutsche Forschungsgemeinschaft (DFG) in the framework of the Research Training Group 'Bioinformatics' at Bielefeld University (GK 635), and of the Priority Programme 'Probabilistic Structures in Evolution', SPP 1590 (BA1070/12-1 and BA2469/7-1).

## References

Aigner, M., 1979. *Combinatorial Theory*, Springer, Berlin.

Baake, E., 2001. Mutation and recombination with tight linkage. *J. Math. Biol.* 42: 455–488.

Baake, M. 2005. Recombination semigroups on measure spaces. *Monatsh. Math.* 146: 267–278 and 150: 83–84 (2007) (Addendum).

Baake, E. 2010. Deterministic and stochastic aspects of single–crossover recombination. *Proc. of the International Congress of Mathematicians,* Hyderabad, India, 2010, Vol. IV, ed. R. Bhatia, Hindustan Book Agency, 3037-3053.

Baake, M., Baake, E. 2003. An exactly solved model for mutation, recombination and selection. *Can. J. Math.* 55: 3–41 and (2008) 60: 264–265 (Erratum).

Baake, E., Herms, I. 2008. Single-crossover dynamics: finite versus infinite populations. *Bull. Math. Biol.* 70: 603–624.

Baake, E., Hustedt, T. 2011. Moment closure in a Moran model with recombination. *Markov Proc. Relat. Fields* 17: 429–446.

Bennett, J. H. 1954. On the theory of random mating. *Ann. Human Genetics* 18: 311–317.

Bürger, R. 2000. *The Mathematical Theory of Selection, Recombination and Mutation*. Wiley, Chichester.

Christiansen, F. B. 1999. *Population Genetics of Multiple Loci*. Wiley, Chichester.

Cohn, D. L. 1980. *Measure Theory*. Birkhäuser, Boston.

Dawson, K. J. 2000. The decay of linkage disequilibria under random union of gametes: How to calculate Bennett's principal components. *Theor. Popul. Biol.* 58: 1–20.

Dawson, K. J. 2002. The evolution of a population under recombination: How to linearise the dynamics. *Lin. Alg. Appl.* 348: 115–137.

Donnelly, K.P. 1983. The probability that related individuals share some section of genome identical by descent. *Theor. Popul. Biol.* 23: 34–63.

Durrett, R. 2008. *Probability Models for DNA Sequence Evolution*. 2nd ed., Springer, New York.

Ethier, S. N., Kurtz, T. G. 1986; Reprint 2005. *Markov Processes – Characterization and Convergence*. Wiley, New York.

Ewens, W. 2004. *Mathematical Population Genetics*. 2nd ed., Springer, Berlin.

Geiringer, H. 1944. On the probability theory of linkage in Mendelian heredity. *Ann. Math. Stat.* 15: 25–57.

Gill, J., Linusson, S., Moulton, V., Steel, M. 2008. A regular decomposition of the edge-product space of phylogenetic trees. *Adv. Appl. Math.* 41: 158–176.

Griffiths, R. C., Marjoram, P. 1996. Ancestral inference from samples of DNA sequences with recombination. *J. Comput. Biol.* 3: 479–502.

Grimmett, G., Stirzaker, D. 2001 *Probability and Random Processes*. 3rd ed., Oxford University Press, Oxford.

Gross, J., Yellen, J. 1999. *Graph Theory and its Applications*. CRC Press, Boca Raton.

Gupta, D. K. 1992. Generation of binary trees from $(0-1)$ codes. *Intern. J. Computer Math.* 42: 157–162.

Hein, J., Schierup, M. H., Wiuf, C. 2005; corr. reprint 2006. *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory*. Oxford University Press, Oxford.

Kauppi, L., Alec J. Jeffreys, A.L., Keeney, S. 2004. Where the crossovers are: recombination distributions in mammals. *Nature Rev. Genet.* 5: 413–424.

Lyubich, Y. I. 1992. *Mathematical Structures in Population Genetics*. Springer, Berlin.

McHale, D., Ringwood, G. A. 1983. Haldane linearisation of baric algebras. *J. London Math. Soc.* (2) 28: 17–26.

McVean, G.A.T., Cardin, N.J. 2005. Approximating the coalescent with recombination. *Phil. Trans. Roy. Soc. B* 360: 1387–1393.

Proskurowski, A. 1980. On the generation of binary trees. *J. ACM* 27: 1–2.

Ralph, P., Coop, G. 2013. The geography of recent genetic ancestry across Europe. Submitted; arXiv:1207.3815.

Robbins, R. B. 1918. Some applications of mathematics to breeding problems III. *Genetics* 3: 375–389.

Semple, C., Steel, M. 2003. *Phylogenetics*, Oxford University Press, Oxford.

Stanley, R. P. 1999. *Enumerative Combinatorics*, Vol. 2, Cambridge University Press, Cambridge.

Wakeley, J. 2008. *Coalescent Theory: An Introduction*. Roberts & Company Publishers, Greenwood Village, CO.

Watterson, G.A. 1975. On the number of segregating sites in genetical models without recombination. *Theor. Pop. Biol.* 7: 256–276.

Wiuf, C., Hein, J., 1997. On the number of ancestors to a DNA sequence. *Genetics* 147: 1459–1468.

von Wangenheim, U., Baake, E., Baake, M. 2010. Single-crossover recombination in discrete time. *J. Math. Biol.* 60: 727–760.

Zaks, S. 1980. Lexicographic generation of ordered trees. *Theor. Comp. Sci.* 10: 63–82.