Splitting trees with neutral Poissonian mutations I: Small families

Nicolas Champagnat, Amaury Lambert

Splitting trees with neutral Poissonian mutations I: Small families.

Nicolas Champagnat[1], Amaury Lambert[2]

Abstract

We consider a neutral dynamical model of biological diversity, where individuals live and reproduce independently. They have i.i.d. lifetime durations (which are not necessarily exponentially distributed) and give birth (singly) at constant rate b. Such a genealogical tree is usually called a splitting tree [9], and the population counting process $(N_t; t \geq 0)$ is a homogeneous, binary Crump–Mode–Jagers process.

We assume that individuals independently experience mutations at constant rate θ during their lifetimes, under the infinite-alleles assumption: each mutation instantaneously confers a brand new type, called allele, to its carrier. We are interested in the allele frequency spectrum at time t, i.e., the number $A(t)$ of distinct alleles represented in the population at time t, and more specifically, the numbers $A(k,t)$ of alleles represented by k individuals at time t, $k = 1, 2, \ldots, N_t$.

We mainly use two classes of tools: coalescent point processes, as defined in [15], and branching processes counted by random characteristics, as defined in [11, 12]. We provide explicit formulae for the expectation of $A(k,t)$ conditional on population size in a coalescent point process, which apply to the special case of splitting trees. We separately derive the a.s. limits of $A(k,t)/N_t$ and of $A(t)/N_t$ thanks to random characteristics, in the same vein as in [19].

Last, we separately compute the expected homozygosity by applying a method introduced in [14], characterizing the dynamics of the tree distribution as the origination time of the tree moves back in time, in the spirit of backward Kolmogorov equations.

MSC 2000 subject classifications: Primary 60J80; secondary 92D10, 60J85, 60G51, 60G55, 60J10, 60K15.

Key words and phrases. branching process – coalescent point process – splitting tree – Crump–Mode–Jagers process – linear birth–death process – allelic partition – infinite alleles model – Poisson point process – Lévy process – scale function – regenerative set – random characteristic.

1 Introduction

We consider a general branching population, where individuals reproduce independently of each other, have i.i.d. lifetime durations with arbitrary distribution, and give birth at constant rate during

[1]TOSCA project-team, INRIA Nancy – Grand Est, IECN – UMR 7502, Nancy-Université, Campus scientifique, B.P. 70239, 54506 Vandœuvre-lès-Nancy Cedex, France, E-mail: Nicolas.Champagnat@inria.fr

[2]Laboratoire de Probabilités et Modèles Aléatoires, UMR 7599 CNRS and UPMC Univ Paris 06, Case courrier 188, 4 Place Jussieu, F-75252 Paris Cedex 05, France, Email: amaury.lambert@upmc.fr

their lifetime. We also assume that each birth gives rise to a single newborn. The genealogical tree associated with this construction is known as a splitting tree [8, 9, 15]. The process $(N_t; t \geq 0)$ counting the population size is a non-Markovian birth–death process belonging to the class of general branching processes, or Crump–Mode–Jagers (CMJ) processes. Since births arrive singly and at constant rate, these processes are sometimes called homogeneous, binary CMJ processes.

Next, individuals are given a type, called allele or haplotype. They inherit their type at birth from their mother, and (their germ line) change type throughout their lifetime, at the points of independent Poisson point processes with rate θ, conditional on lifetimes (neutral mutations). The type conferred by a mutation is each time an entirely new type, an assumption known as the infinitely-many alleles model.

We are interested in the so-called allelic partition (partition into types) of the population alive at time t. A convenient way of describing this partition without labelling types is to define the number $A_\theta(k,t)$ of types carried by k individuals at time t. The sequence $(A_\theta(k,t); k \geq 1)$ is called the frequency spectrum of the allelic partition. We also denote by $A_\theta(t)$ the total number of distinct types at time t. The most celebrated mathematical result in this setting is Ewens' sampling formula, which yields the distribution of the frequency spectrum for the Kingman coalescent tree with neutral Poissonian mutations [7].

Credit is due to G. Yule [20] for the first study of a branching tree with mutations, but the interest for the infinitely-many alleles model applied to branching trees has started with the work of R.C. Griffiths and A.G. Pakes [10], where the tree under focus is a Galton–Watson tree and each individual, with a fixed probability, is independently declared mutant at birth. A fascinating monography dedicated to general branching processes (also undergoing mutations only at birth times) is due to Z. Taïb [19]. An extensive use is done there of a.s. limit theorems for branching processes counted by random characteristics, due to P. Jagers and O. Nerman [11, 12, 13, 16].

More recently, in a series of three companion papers, J. Bertoin [2, 3, 4] has set up a very general framework for Galton–Watson processes with mutations, where he has considered the allelic partition of the whole population from origination to extinction, and studied various scaling limits for large initial population sizes and low mutation probabilities. Branching processes have also been used in the study of multistage carcinogenesis. In this setting, the emphasis is put on the waiting time until a target mutation occurs, see [6, 18] and the references therein.

In this paper, we study the part of the frequency spectrum corresponding to families with a fixed number k of carriers, $k \geq 1$, that we call *small families*. We use three techniques: coalescent point processes, branching processes counted by random characteristics, and Kolomogorov-type equations as a function of the origination time of the tree. In a companion paper [5], we will discuss the part of the frequency spectrum corresponding to the largest or/and oldest families (the age of a family being that of their original mutation).

2 Model and statement of main results

2.1 Model

In this work, we consider genealogical trees satisfying the branching property and called *splitting trees* [8, 9]. Splitting trees are those random trees where individuals' lifetime durations are i.i.d. with an arbitrary distribution, but where birth events occur at Poisson times during each individual's lifetime. We call b this constant birth rate and we denote by V a r.v. distributed as the lifetime duration. Then set $\Lambda(dr) := b\mathbb{P}(V \in dr)$ a finite measure on $(0, \infty]$ with total mass b called the *lifespan measure*. We will always assume that a splitting tree is started with one unique progenitor born at time 0.

The process $(N_t; t \geq 0)$ counting the number of alive individuals at time t is a homogeneous, binary *Crump–Mode–Jagers process*, which is not Markovian unless Λ has an exponential density or is the Dirac mass at $\{+\infty\}$.

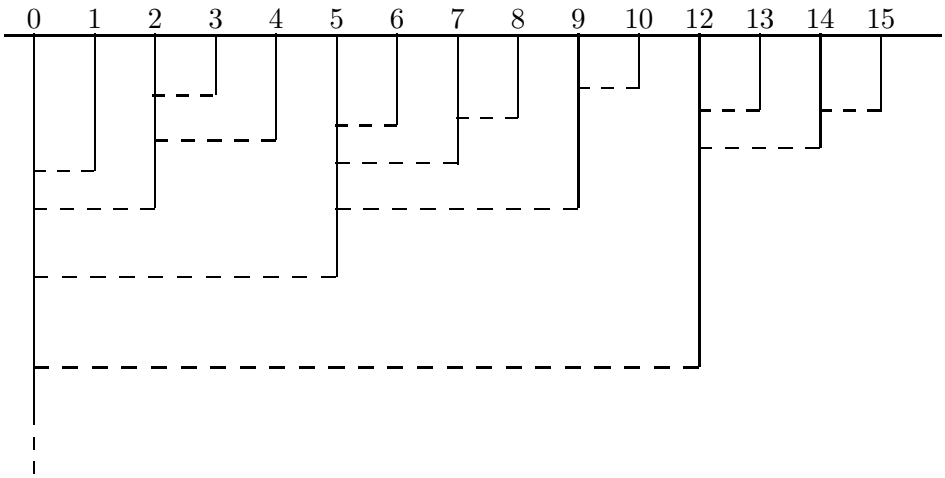

Figure 1: A coalescent point process for 16 individuals, hence 15 branches.

In [15], it is shown that the genealogy of a splitting tree conditioned to be extant at a fixed time t is given by a *coalescent point process*, that is, a sequence of i.i.d. random variables $H_i \stackrel{d}{=} H$, $i \geq 1$, killed at its first value greater than t. In particular, conditional on $N_t \neq 0$, N_t follows a geometric ditribution with parameter $\mathbb{P}(H < t)$. More specifically, for any $0 \leq i \leq N_t - 1$, the *coalescence time* between the i-th individual alive at time t and the j-th individual alive at time t (i.e., the time elapsed since the common lineage to both individuals splits into two distinct lineages) is the maximum of $H_{i+1}, \ldots, H_j$. The graphical representation on Figure 1 is straightforward. The common law of these so-called *branch lengths* is given by

$$\mathbb{P}(H > s) = \frac{1}{W(s)}, \tag{2.1}$$

where the nondecreasing function W is such that $W(0) = 1$ and is characterized by its Laplace transform. More specifically, these branch lengths are the depths of the excursions of the jump

contour process, say $Y^{(t)}$, of the splitting tree truncated below level t. They are i.i.d. because $Y^{(t)}$ is a Markov process. Indeed, it is shown in [15] that $Y^{(t)}$ has the law of a Lévy process, say Y, without negative jumps, reflected below t and killed upon hitting 0. The function W is called the scale function of Y, and is defined from the Laplace exponent ψ of Y:

$$\psi(x) = x - \int_{(0,+\infty]} \left(1 - e^{-rx}\right) \Lambda(dr) \qquad x \in \mathbb{R}_+. \tag{2.2}$$

Let α denote the largest root of ψ. In the supercritical case (i.e. $\int_{(0,\infty]} r\Lambda(dr) > 1$), and in this case only, α is positive and called the *Malthusian parameter*, because the population size grows exponentially at rate α on the survival event. Then the function W is characterized by

$$\int_0^\infty e^{-xr} W(r)\, dr = \frac{1}{\psi(x)} \qquad x > \alpha.$$

Actually, it is possible to show by path decompositions of the process Y that

$$W(x) = \exp\left(b \int_0^x dt\, \mathbb{P}(J > t)\right),$$

where J is the maximum of the path of Y killed upon hitting 0 and started from a random initial value, distributed as V. Note that since Y is also the contour process of a splitting tree, J has the law of the extinction time of the CMJ process $N = (N_t; t \geq 0)$ started from one individual.

In the next section, we consider coalescent point processes without reference to a splitting tree. The law of such a process is merely characterized by a random number N of i.i.d. r.v. (H_i) independent of N, both with arbitrary distributions. In this setting, (2.1) conversely serves as a *definition of* W, which is now an arbitrary nondecreasing function, whereas it was previously seen to be differentiable in the special case of splitting trees. The population size N can be fixed (possibly infinite) or truly random, e.g. following a geometric distribution. It will be written N_t when the law of H is supported by $[0, t]$. In this latter case, any result obtained under the assumption that N follows a geometric distribution can be applied to the case of splitting trees.

Throughout this work, we assume that individuals independently experience mutations at Poisson times during their lifetime, that each new mutation event confers a brand new type (called haplotype, or allele) to the individual, and that a newborn holds the same type as her mother at birth time. The mutation rate is denoted by θ.

2.2 Outline and statement of main results

The main technique we use relies on the previously described representation of the genealogy of a splitting tree by a sequence of i.i.d. r.v. $(H_i)_{i\geq 1}$, called the coalescent point process. This idea was first exploited by Aldous and Popovic [1] and Popovic [17] and then it was further developed by Lambert [15]. The common distribution of $H_1, H_2, \ldots$ is related to the scale function W. We will also use the scale function W_θ associated with the lifetime of clonal families (standard lifetime truncated at its first mutation event). Section 3 is dedicated to some fine computations in the general framework of coalescent point processes. For example, for a coalescent point process $(H_0, H_1, \ldots, H_X)$ of age t,

where X is an independent geometric r.v., Theorem 3.3 gives the expectation of $A_\theta(k,t)u^X$. Various corollaries are stated, giving the expectation, sometimes conditional on the population size, of specific quantities of biological interest at the fixed time t. Those statements extend results of [14] given under a doubly asymptotic regime ($t, n \to \infty$). For example, Corollary 3.4 gives the expectation of the number of distinct alleles and of homozygosity (probability of drawing two individuals carrying the same allele) and Corollary 3.10 gives the expectation of the number $Z_0(y;n)$ among the n first individuals who carry the ancestral type of lineage 0 y units of time in the past

$$\mathbb{E}\, Z_0(y;n) = e^{-\theta y} \sum_{k=0}^{n} \mathbb{P}(H \leq y)^k,$$

see Remark 3.11 for a simple interpretation of this formula.

In Section 4, some of the previous results are specified to the case of splitting trees. In particular, Proposition 4.1 yields the expectation of $A_\theta(k,t)u^{N_t}$, as well as of $Z_0(t)u^{N_t}$, where $Z_0(t)$ denotes the number of alive individuals at time t carrying the ancestral allele. The result for $A_\theta(k,t)$ can even be detailed to the case of haplotypes of a given age. As previously, various corollaries are provided for some quantities such as the homozygosity. Ruling out the information on the population size (i.e., taking $u = 1$) and on the age of the mutation, Corollary 4.3 reads

$$\mathbb{E}_t A_\theta(k,t) = W(t) \int_0^t dx\, \theta\, e^{-\theta x} \frac{1}{W_\theta(x)^2} \left(1 - \frac{1}{W_\theta(x)}\right)^{k-1},$$

and

$$\mathbb{P}_t\left(Z_0(t) = k\right) = W(t) \frac{e^{-\theta t}}{W_\theta(t)^2} \left(1 - \frac{1}{W_\theta(t)}\right)^{k-1},$$

where $\mathbb{P}_t$ is the conditional probability on survival up until time t. Note also that Subsection 4.2 provides the reader with a more explanatory proof of the previous formulae.

The theory of random characteristics [11, 12, 13, 16, 19], which is the second main technique we use, is displayed in Section 5. There, the random characteristic of individual i, say, can be for example the number $\chi_i^k(t)$ of mutations that i has experienced during her lifetime and which are carried by k alive individuals, t units of time after her birth ($\chi_i(t) = 0$ if $t < 0$). Then the total number of haplotypes carried by k individuals at time t (except possibly the ancestral type) is the sum over all individuals i (dead or alive) of $\chi_i(t - \sigma_i)$, where σ_i is the birth time of individual i. Now according to limit theorems by P. Jagers and O. Nerman [11, 12, 13, 16], these sums converge a.s. on the survival event in the supercritical case. Exploiting those limit theorems, we are able to deduce the following a.s. convergences in the supercritical case (see Proposition 5.1), where the limits are computed independently from the results obtained earlier. On the survival event,

$$\lim_{t\to\infty} \frac{A_\theta(k,t)}{A_\theta(t)} = \frac{U_k}{U} \qquad a.s.$$

and

$$\lim_{t\to\infty} \frac{A_\theta(t)}{N_t} = U \qquad a.s.,$$

where

$$U_k := \int_0^\infty dx\, \theta\, e^{-\theta x}\, \frac{1}{W_\theta(x)^2} \left(1 - \frac{1}{W_\theta(x)}\right)^{k-1},$$

and

$$U := \sum_{k\geq 1} U_k = \int_0^\infty dx\, \theta\, e^{-\theta x}\, \frac{1}{W_\theta(x)}.$$

In the final section (Section 6), we consider $G_\theta(t) := Z_0(t)(Z_0(t)-1)/2 + \sum_{k\geq 1} k(k-1)A_\theta(k,t)/2$, that we term absolute homozygosity, in reference to standard homozygosity, which is defined as $\bar{G}_\theta(t) = 2G_\theta(t)/N_t(N_t-1)$. Homozygosity is a well-known measure of diversity, that can be seen as the probability that two randomly sampled *distinct* individuals (or sequences) share the same allele. In the spirit of backward Kolmogorov equations, we derive the dynamics of the expectation of $G_\theta(t)u^{N_t}$ as the origination time of the tree moves back in time. Then the expected standard and absolute homozygosity can be computed. In passing, we recover formulae obtained in Section 4 by totally different methods. Specifically, we get $\mathbb{E}_t G_\theta(t) = W(t)(W_{2\theta}(t) - 1)$.

3 Expected haplotype frequencies for coalescent point processes

In this section, unless otherwise specified, we assume that the lineage of individual 0, sometimes called lineage 0, is infinite, and that all other branch lengths are i.i.d., distributed as some r.v. H. To each H_i corresponds an individual, that we call individual i. We also assume that mutations occur according to a Poisson point process on edge lengths with parameter θ.

3.1 The next branch with no extra mutation

We let $\mathcal{E}^\theta$ denote the set of individuals who *carry no more mutations* (but possibly less) than individual 0 (some of and at most exactly the mutations carried by 0, but no other mutation). We call such individuals $(0,\cdot)$-*type individuals* (same type as some point on lineage 0 at some time in the past).

Set $K_0^\theta := 0$ and for $i \geq 1$, define K_i^θ as the label of the i-th individual in $\mathcal{E}^\theta$. In addition, set

$$H_i^\theta := \max\{H_j : K_i^\theta < j \leq K_{i+1}^\theta\}$$

and

$$B_i^\theta := K_i^\theta - K_{i-1}^\theta.$$

See Figure 2 for a graphical representation of these quantities on a typical coalescent point process with mutations.

We write (B^θ, H^θ) in lieu of (B_1^θ, H_1^θ) and we define $W_\theta(x;\gamma)$ by

$$W_\theta(x;\gamma) := \frac{1}{1 - \mathbb{E}\left(\gamma^{B^\theta}, H^\theta \leq x\right)} \qquad x \geq 0, \gamma \in (0,1].$$

We will also need the following notation

$$W(x;\gamma) := \frac{1}{1 - \gamma\mathbb{P}(H \leq x)} \qquad x \geq 0, \gamma \in (0,1].$$

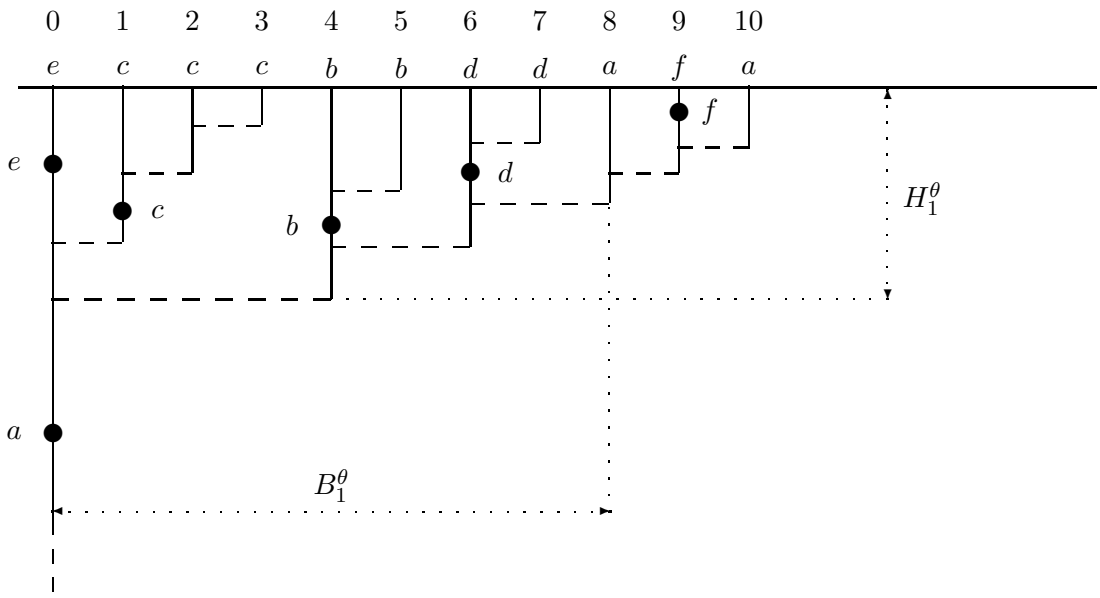

Figure 2: On this coalescent point process, the 8-th individual is the first one whose type is the same as some point on lineage 0 anywhere in the past, so that $8 \in \mathcal{E}^\theta$ and $B_1^\theta = 8$. The maximum H_1^θ of the first B_1^θ branch lengths is shown. Also note that $10 \in \mathcal{E}^\theta$ and $B_2^\theta = 2$.

Theorem 3.1 *The bivariate sequence* $((B_i^\theta, H_i^\theta); i \geq 1)$ *is a sequence of i.i.d. random pairs. In addition, the following formula holds for all* $x \geq 0$ *and* $\gamma \in (0, 1]$

$$W_\theta(x;\gamma) = e^{-\theta x} W(x;\gamma) + \theta \int_0^x W(y;\gamma)\, e^{-\theta y}\, dy.$$

Remark 3.2 *Differentiating both sides of the previous equation w.r.t. the first variable yields*

$$dW_\theta(x;\gamma) = e^{-\theta x}\, dW(x;\gamma).$$

Also, the formula in the previous statement was shown in [14] in the special case $\gamma = 1$.

Proof. First observe that the pair (K_1^θ, H_1^θ) does not depend on the haplotype of individual 0, and that the i-th $(0, \cdot)$-type individual is also the next individual after K_{i-1}^θ with no mutation other than those carried by individual K_{i-1}^θ. This ensures that $(K_i^\theta - K_{i-1}^\theta, H_i^\theta)$ has the same law as (K_1^θ, H_1^θ), and the independence between $(K_i^\theta - K_{i-1}^\theta, H_i^\theta)$ and previous pairs is due to the independence of branch lengths and the fact that new mutations can only occur on branches with labels strictly greater than K_{i-1}^θ.

As for the formula relating W^θ and W, we consider the renewal process S defined by $S_0 = 0$ and $S_n = \sum_{i=1}^n B_i^\theta$. Next, for any integer $k \geq 0$, let F_k denote the event

$$F_k := \{\exists n \geq 0 : S_n = k,\ M_n \leq x\},$$

where $M_n := \max\{H_i^\theta : 1 \leq i \leq n\}$. Let T_k denote the time elapsed since the lineages of individual 0 and individual k have split up, that is, $T_k = \max\{H_i : 1 \leq i \leq k\}$. Notice that by definition of H_i^θ,

$T_k = M_n$ on the event $\{S_n = k\}$, so that

$$F_k = \{\exists n \geq 0 : S_n = k,\ T_k \leq x\}.$$

So F_k is the event that the lineage of individual k has had no mutation between time $-T_k$ and present time 0 (i.e., no mutation on the part of its lineage not common with individual 0), and $T_k \leq x$. By standard properties of Poisson processes, we get

$$\begin{aligned}\mathbb{P}(F_k) &= \mathbb{E}\left(e^{-\theta T_k}, T_k \leq x\right) \\ &= \mathbb{P}(H \leq x)^k e^{-\theta x} + \theta \int_0^x \mathbb{P}(H \leq y)^k \, e^{-\theta y} \, dy. \end{aligned} \tag{3.1}$$

Note that the r.h.s. of this equation is obtained using the integration by parts formula for càdlàg functions (i.e. functions continuous on the right and admitting left limits at each points of the space, like $\mathbb{P}(H \leq x)$): if f is continuously differentiable and g is càdlàg with bounded variation,

$$f(x)g(x) = f(0)g(0) + \int_0^x f'(y)g(y)dy + \int_{(0,x]} f(y)dg(y). \tag{3.2}$$

Equation (3.1) yields

$$\sum_{k\geq 0} \gamma^k \mathbb{P}(F_k) = e^{-\theta x} W(x;\gamma) + \theta \int_0^x W(y;\gamma)\, e^{-\theta y}\, dy.$$

On the other hand,

$$\begin{aligned}\sum_{k\geq 0} \gamma^k \mathbb{P}(F_k) &= \sum_{k\geq 0} \gamma^k \sum_{n\geq 0} \mathbb{P}(S_n = k, M_n \leq x) \\ &= \sum_{n\geq 0} \mathbb{E}\left(\gamma^{S_n}, M_n \leq x\right) \\ &= \sum_{n\geq 0} \mathbb{E}\left(\gamma^{\sum_{i=1}^n B_i^\theta}, H_1^\theta \leq x, \ldots, H_n^\theta \leq x\right) \\ &= \sum_{n\geq 0} \left(\mathbb{E}\left(\gamma^{B^\theta}, H^\theta \leq x\right)\right)^n \\ &= \frac{1}{1 - \mathbb{E}\left(\gamma^{B^\theta}, H^\theta \leq x\right)}, \end{aligned}$$

which yields the desired result. □

3.2 Expected haplotype frequencies for geometrically distributed population sizes

Let X denote some independent geometric random variable with parameter γ, that is, $\mathbb{P}(X \geq n) = \gamma^n$ for any $n \geq 0$.

In the infinite-allele model, each haplotype is characterized by its most recent mutation. We denote by $A_\theta(k, y; \gamma)$ the number of haplotypes whose most recent mutation occurred between time $-y$ and present time 0 and which are carried by k individuals among $\{0, 1, \ldots, X\}$.

Theorem 3.3 *For all $k \geq 1$, $y > 0$, $\gamma \in (0,1]$, $u \in [0,1]$,*

$$\mathbb{E}\left(u^X A_\theta(k,y;\gamma)\right) = \frac{1-\gamma}{(1-u\gamma)^2}\int_0^y dx\,\theta\,e^{-\theta x}\,\frac{1}{W_\theta(x;u\gamma)^2}\left(1-\frac{1}{W_\theta(x;u\gamma)}\right)^{k-1}.$$

Let $I'_\theta(y;\gamma)$ (resp. $I'_\theta(y;n)$) denote the number of individuals among $\{0,1,\ldots,X\}$ (resp. $\{0,1,\ldots,n\}$) whose most recent mutation appeared between time $-y$ and present time 0.

Let $A_\theta(y;\gamma)$ (resp. $A_\theta(y;n)$) denote the number of distinct haplotypes represented in $\{0,1,\ldots,X\}$ (resp. $\{0,1,\ldots,n\}$) whose most recent mutation appeared between time $-y$ and present time 0.

Let $\bar{G}_\theta(y;n)$ denote the *conditional probability* that two *distinct* individuals randomly drawn ('conditional probability' here refers to this sampling, given the coalescent point process and the mutation times on branches) from $\{0,1,\cdots,n\}$ *share the same haplotype* and that the most recent mutation of this common haplotype appeared between time $-y$ and present time 0.

Corollary 3.4 *For any integer $n \geq 1$,*

$$\mathbb{E}\,I'_\theta(y;n-1) = n(1-\exp(-\theta y)),$$

$$\mathbb{E}\,A_\theta(y;n-1) = n\int_0^y dx\,\theta\,e^{-\theta x}\,\mathbb{P}(H^\theta > x) + \int_0^y dx\,\theta\,e^{-\theta x}\,\mathbb{E}\left(B^\theta \wedge n, H^\theta \leq x\right). \tag{3.3}$$

and in the case where the law of H has no atom

$$\mathbb{E}\,\bar{G}_\theta(y;n-1) = 2\sum_{k=1}^{n-1}\frac{k(n-k)}{n(n-1)}\int_0^y \mathbb{P}(H\in dx)\,\mathbb{P}(H\leq x)^{k-1}e^{-\theta x}\left(e^{-\theta x}-e^{-\theta y}\right).$$

Remark 3.5 *The first expectation can readily be deduced from some exchangeability argument, since each individual carries a mutation with age smaller than y with probability $1-\exp(-\theta y)$ (there is no edge effect since the ancestral lineage is infinite).*

Remark 3.6 *In [14], a pathwise result was shown for the number $A_\theta(\infty,n)$ of distinct haplotypes represented in $\{0,1,\cdots,n\}$, namely*

$$\lim_{n\to\infty} n^{-1}A_\theta(\infty,n) = \int_0^\infty dx\,\theta\,e^{-\theta x}\,\mathbb{P}(H^\theta > x) \qquad a.s.$$

Remark 3.7 *In the case where the law of H admits atoms, the computation of $\mathbb{E}\bar{G}_\theta(y;n-1)$ can be done following the same line as in the proof below, using the fact that $dW(x;\gamma)$ has an atomic part. The computation gives*

$$\mathbb{E}\,\bar{G}_\theta(y;n-1) = 2\sum_{k=1}^{n-1}\frac{(n-k)}{n(n-1)}\left\{k\int_0^y \mu_H^{n.a.}(dx)\,\mathbb{P}(H\leq x)^{k-1}e^{-\theta x}\left(e^{-\theta x}-e^{-\theta y}\right)\right.$$
$$\left. + \sum_{x\in[0,y]}\left(\mathbb{P}(H\leq x)^{k-1}-\mathbb{P}(H<x)^{k-1}\right)e^{-\theta x}\left(e^{-\theta x}-e^{-\theta y}\right)\right\},$$

where $\mu_H^{n.a.}$ is the non-atomic part of the law of H.

Proof of Corollary 3.4. For the first expectation, taking $u = 1$ in the theorem,

$$\mathbb{E}\, I'_\theta(y;\gamma) = \mathbb{E}\sum_{k\geq 1} kA_\theta(k,y;\gamma) = \frac{1-e^{-\theta y}}{1-\gamma},$$

using repeatedly Fubini–Tonelli theorem and $\sum_{k\geq 1} kx^{k-1} = (1-x)^{-2}$ for any $x \in [0,1)$. The result then follows from the inversion of the generating function using $(1-\gamma)^{-1} = \sum_{n\geq 0}(n+1)(1-\gamma)\gamma^n$.

For the second expectation,

$$\mathbb{E}\, A_\theta(y;\gamma) = \mathbb{E}\sum_{k\geq 1} A_\theta(k,y;\gamma) = \frac{1}{1-\gamma}\int_0^y dx\,\theta\, e^{-\theta x}\,\frac{1}{W_\theta(x;\gamma)} = \frac{1}{1-\gamma}\int_0^y dx\,\theta\, e^{-\theta x}\,\left(1-\mathbb{E}\left(\gamma^{B^\theta}, H^\theta \leq x\right)\right).$$

Next invert the generating function as follows

$$\begin{aligned}
\frac{1}{1-\gamma}\mathbb{E}\left(\gamma^{B^\theta}, H^\theta \leq x\right) &= \sum_{n\geq 0}(n+1)(1-\gamma)\gamma^n \sum_{j\geq 0}\mathbb{P}(B^\theta = j, H^\theta \leq x)\gamma^j \\
&= \sum_{n\geq 0}(1-\gamma)\gamma^n \sum_{k=0}^{n}(n+1-k)\mathbb{P}(B^\theta = k, H^\theta \leq x) \\
&= \sum_{n\geq 0}(1-\gamma)\gamma^n \mathbb{E}\left(n+1-B^\theta, B^\theta \leq n, H^\theta \leq x\right),
\end{aligned}$$

which entails

$$\begin{aligned}
\mathbb{E}\, A_\theta(y;n) &= \int_0^y dx\,\theta\, e^{-\theta x}\,\left(n+1-\mathbb{E}\left(n+1-B^\theta, B^\theta \leq n, H^\theta \leq x\right)\right) \\
&= \int_0^y dx\,\theta\, e^{-\theta x}\,\left((n+1)\mathbb{P}(H^\theta > x) + \mathbb{E}\left(n+1-(n+1-B^\theta)1_{\{B^\theta \leq n\}}, H^\theta \leq x\right)\right) \\
&= \int_0^y dx\,\theta\, e^{-\theta x}\,\left((n+1)\mathbb{P}(H^\theta > x) + \mathbb{E}\left((n+1)1_{\{B^\theta > n\}} + B^\theta 1_{\{B^\theta \leq n\}}, H^\theta \leq x\right)\right),
\end{aligned}$$

which yields the result.

For the third expectation, we use the fact that the expected number of (unordered) pairs of individuals sharing the same haplotype (younger than y) equals

$$\sum_{n\geq 0}(1-\gamma)\gamma^n \frac{n(n+1)}{2}\bar{G}_\theta(y;n) = \mathbb{E}[\bar{G}_\theta(y;\gamma)],$$

where

$$\bar{G}_\theta(y;\gamma) := \sum_{k\geq 2}\frac{k(k-1)}{2}A_\theta(k,y;\gamma).$$

Now since $\sum_{k\geq 2} k(k-1)x^{k-1} = 2x(1-x)^{-3}$, we get

$$\begin{aligned}
\mathbb{E}\bar{G}_\theta(y;\gamma) &= \frac{1}{1-\gamma}\int_0^y dx\,\theta\, e^{-\theta x}(W_\theta(x;\gamma)-1) \\
&= \frac{1}{1-\gamma}\int_0^y dx\,\theta\, e^{-\theta x}\int_0^x e^{-\theta z} dW(z;\gamma) \\
&= \frac{1}{1-\gamma}\int_0^y dW(z;\gamma)\, e^{-\theta z}\,\left(e^{-\theta z} - e^{-\theta y}\right),
\end{aligned}$$

where differentiation of W is understood w.r.t. the first variable. Then we use the fact, when the law of H has no atom,

$$dW(z;\gamma) = \frac{\gamma \mathbb{P}(H \in dz)}{(1-\gamma \mathbb{P}(H \le z))^2} = \mathbb{P}(H \in dz) \sum_{n \ge 0} n \gamma^n \mathbb{P}(H \le z)^{n-1}.$$

The proof ends writing the product series between the last entire series and $(1-\gamma)^{-2} = \sum_{n \ge 0}(n+1)\gamma^n$. □

Before proving the theorem, we insert a paragraph in which we state and prove a preliminary key result.

3.2.1 A key lemma

We denote by ℓ_i the time elapsed since the i-th most recent mutation on the lineage of individual 0, also called lineage 0. Let $N_i(y;\gamma)$ denote the number of $(0,\cdot)$-type individuals in $\{0, 1 \dots, X\}$ whose *most recent mutation time* in its haplotype is ℓ_i if $\ell_i \le y$, and $N_i(y;\gamma) = 0$ otherwise.

We also define $(0,y)$*-type individuals* as those individuals that have the same type as the point at time $-y$ on lineage 0. In other words, an individual is of $(0,y)$-type if the most recent mutation of its haplotype is ℓ_i for the unique i such that $\ell_{i-1} \le y < \ell_i$, with the convention that $\ell_0 := 0$. In the same vein, $(0,[0,y])$-type individuals are those individuals that have the same type as some point on lineage 0 at any time between time $-y$ and present time 0.

We denote by $Z_0(y;\gamma)$ the number of $(0,y)$-type individuals of $\{0,1,\dots,X\}$. Note that $Z_0(y;\gamma) = N_i(\infty;\gamma)$ where i is such that $\ell_{i-1} \le y < \ell_i$. Also set $I_0(y;\gamma)$ the number of $(0,[0,y])$-type individuals of $\{0,1,\dots,X\}$ and $I_0'(y;\gamma)$ the number of $(0,\cdot)$-type individuals of $\{0,1,\dots,X\}$ whose most recent mutation appeared between time $-y$ and present time 0. Otherwise said,

$$I_0(y;\gamma) = I_0'(y;\gamma) + Z_0(y;\gamma) \quad \text{and} \quad I_0'(y;\gamma) = \sum_{i \ge 1} N_i(y;\gamma)$$

Lemma 3.8 *For all $k \ge 1$, $y > 0$, $\gamma \in (0,1]$, $u \in [0,1]$,*

$$\sum_{i \ge 1} \mathbb{E}\left(u^X, N_i(y,\gamma) = k\right) = \frac{1-\gamma}{1-u\gamma} \int_0^y dz\, \theta\, e^{-\theta z}\, \frac{W(z;u\gamma)}{W_\theta(z;u\gamma)^2} \left(1 - \frac{1}{W_\theta(z;u\gamma)}\right)^{k-1}$$

and

$$\mathbb{E}\left(u^X, Z_0(y;\gamma) = k\right) = \frac{1-\gamma}{1-u\gamma} e^{-\theta y}\, \frac{W(y;u\gamma)}{W_\theta(y;u\gamma)^2} \left(1 - \frac{1}{W_\theta(y;u\gamma)}\right)^{k-1}.$$

Corollary 3.9 *For all $y > 0$, $\gamma \in (0,1]$, $u \in [0,1]$,*

$$\mathbb{E}\left(u^X I_0'(y;\gamma)\right) = \frac{1-\gamma}{1-u\gamma} \int_0^y dz\, \theta\, e^{-\theta z}\, W(z;u\gamma) \quad \textit{and} \quad \mathbb{E}\left(u^X Z_0(y;\gamma)\right) = \frac{1-\gamma}{1-u\gamma} e^{-\theta y}\, W(y;u\gamma).$$

Note that, in the case $\gamma = u = 1$, one must replace $(1-\gamma)/(1-u\gamma)$ by 1 in these results.

Proof. Use the formulae in Lemma 3.8 and Fubini–Tonelli theorem repeatedly, in particular to see that

$$\mathbb{E}\, I_0'(y;\gamma)\, u^X = \sum_{i\geq 1} \mathbb{E}\, u^X\, N_i(y;\gamma) = \sum_{i\geq 1}\sum_{k\geq 1} k\mathbb{E}\, u^X\, 1_{N_i(y;\gamma)=k} = \sum_{k\geq 1} k \sum_{i\geq 1} \mathbb{E}\, u^X\, 1_{N_i(y;\gamma)=k}.$$

The proof ends using $\sum_{k\geq 1} kx^{k-1} = (1-x)^{-2}$ for all $x \in [0,1)$. □

Let n be a non-negative integer. In the next corollary, $Z_0(y;n)$ denotes the number of $(0,y)$-type individuals of $\{0,1,\dots,n\}$ and $I_0'(y;n)$ the number of $(0,\cdot)$-type individuals of $\{0,1,\dots,n\}$ whose most recent mutation appeared between time $-y$ and present time 0.

Corollary 3.10 *For all $y>0$ and $n\geq 0$,*

$$\mathbb{E}\, I_0'(y;n) = \int_0^y dz\, \theta\, e^{-\theta z}\, \frac{1-\mathbb{P}(H\leq z)^{n+1}}{\mathbb{P}(H>z)} \qquad \text{and} \qquad \mathbb{E}\, Z_0(y;n) = e^{-\theta y}\, \frac{1-\mathbb{P}(H\leq y)^{n+1}}{\mathbb{P}(H>y)}\,.$$

Proof. We use $(1-\gamma)^{-1} = \sum_{k\geq 0}\gamma^k$ along with

$$W(z;\gamma) = \frac{1}{1-\gamma\mathbb{P}(H\leq z)} = \sum_{n\geq 0} \gamma^n \mathbb{P}(H\leq z)^n.$$

Plugging these equalities into the first formula of the first corollary evaluated at $u=1$ yields

$$\mathbb{E}\, I_0'(y;\gamma) = \int_0^y dz\, \theta\, e^{-\theta z} \frac{1}{1-\gamma}\, W(z;\gamma) = \int_0^y dz\, \theta\, e^{-\theta z} \sum_{n\geq 0} \gamma^n \sum_{k=0}^{n} \mathbb{P}(H\leq z)^k.$$

Inverting the generating function yields the expression proposed for $\mathbb{E}\, I_0'(y;n)$. The very same line of reasoning can be applied to get $\mathbb{E}\, Z_0(y;n)$. □

Remark 3.11 *Keeping the expression in the proof of the theorem under the shape of a sum is more informative. Indeed, differentiating each side of the equality, we then get*

$$\mathbb{E}\, I_0'(dy;n) = dy\, \theta\, e^{-\theta y} \sum_{k=0}^{n} \mathbb{P}(H\leq y)^k,$$

where $I_0'(dy;n)$ denotes the number of $(0,\cdot)$-type individuals of $\{0,1,\dots,n\}$ whose most recent mutation is of age in $(y,y+dy)$. The interpretation of this new expression goes as follows. The term $\theta\, dy$ is the probability that a mutation occurred on lineage 0 in the time interval $(y,y+dy)$ backwards in time; the term $\mathbb{P}(H\leq y)^k$ is the probability that the lineage of individual k split off lineage 0 more recently than y; the term $e^{-\theta y}$ is the probability that the lineage of individual k has undergone no mutation in the last y units of time.

Proof of Lemma 3.8. Set $D_1 := 1$ and for $i \geq 2$,

$$D_i := \min\{j \geq 1 : H^\theta_j > \ell_{i-1}\}.$$

Also recall the renewal process $S_n = \sum_{i=1}^n B^\theta_i$. Then we have for all $i \geq 1$

$$N_i(y;\gamma) = 1_{l_i \leq y} \left(1_{i=1} + \sum_{j=D_i}^{D_{i+1}-1} 1_{S_j \leq X} \right),$$

the indicator function of $i = 1$ being due to the count of individual 0 in that case. First, we work conditionally on the values v_i of the ages ℓ_i of mutations of lineage 0. Using repeatedly the lack-of-memory property of X, we get for all $i \geq 2$ and $k \geq 1$

$$\begin{aligned}\mathbb{E}\left(u^X, N_i(y;\gamma) = k \mid \ell_j = v_j, j \geq 1\right) &= \cdots \\ &\cdots 1_{v_i \leq y}\, \mathbb{E}\left(u^{S_{D_i-1}}, X \geq S_{D_i-1}\right) \mathbb{E}\left(u^{B^\theta}, B^\theta \leq X, H^\theta \leq v_i \mid H^\theta > v_{i-1}\right) \times \cdots \\ &\cdots \times \mathbb{E}\left(u^{B^\theta}, B^\theta \leq X, H^\theta \leq v_i\right)^{k-1} \left(\mathbb{E}\left(u^X, B^\theta > X\right) + \mathbb{E}\left(u^X, B^\theta \leq X, H^\theta > v_i\right)\right),\end{aligned}$$

where the last multiplicative term equals

$$\begin{aligned}\mathbb{E}\left(u^X, B^\theta > X\right) + \mathbb{E}\left(u^X, B^\theta \leq X, H^\theta > v_i\right) &= \mathbb{E}\left(u^X\right) - \mathbb{E}\left(u^X, B^\theta \leq X, H^\theta \leq v_i\right) \\ &= \mathbb{E}\left(u^X\right)\left(1 - \mathbb{E}\left(u^{B^\theta}, B^\theta \leq X, H^\theta \leq v_i\right)\right) \\ &= \frac{1-\gamma}{1-u\gamma}\left(1 - \mathbb{E}\left((u\gamma)^{B^\theta}, H^\theta \leq v_i\right)\right) \\ &= \frac{1-\gamma}{(1-u\gamma)W_\theta(v_i;u\gamma)}.\end{aligned}$$

Similarly for $i = 1$ and $k \geq 1$,

$$\begin{aligned}\mathbb{E}\left(u^X, N_1(y;\gamma) = k \mid \ell_j = v_j, j \geq 1\right) &= 1_{v_1 \leq y}\, \mathbb{E}\left(u^{B^\theta}, B^\theta \leq X, H^\theta \leq v_1\right)^{k-1} \mathbb{E}\left(u^X\right) \times \cdots \\ &\cdots \times \left(1 - \mathbb{E}\left(u^{B^\theta}, B^\theta \leq X, H^\theta \leq v_1\right)\right) \\ &= 1_{v_1 \leq y}\, \mathbb{E}\left((u\gamma)^{B^\theta}, H^\theta \leq v_1\right)^{k-1} \frac{1-\gamma}{(1-u\gamma)W_\theta(v_1;u\gamma)}.\end{aligned}$$

Now elementary probabilistic reasoning shows that for $i \geq 2$

$$\begin{aligned}&\mathbb{E}\left(u^{S_{D_i-1}}, X \geq S_{D_i-1} \mid \ell_j = v_j, j \geq 1\right) \\ &\qquad = \sum_{k \geq 1} \left(\mathbb{P}(H^\theta \leq v_{i-1})\right)^{k-1} \mathbb{P}(H^\theta > v_{i-1}) \mathbb{E}\left(u^{B^\theta}, B^\theta \leq X \mid H^\theta \leq v_{i-1}\right)^{k-1} \\ &\qquad = \frac{\mathbb{P}(H^\theta > v_{i-1})}{1 - \mathbb{E}\left((u\gamma)^{B^\theta}, H^\theta \leq v_{i-1}\right)} = \mathbb{P}(H^\theta > v_{i-1}) W_\theta(v_{i-1};u\gamma).\end{aligned}$$

As a consequence, for all $i \geq 2$,

$$\mathbb{E}\left(u^X, N_i(y;\gamma) = k \mid \ell_j = v_j, j \geq 1\right) = \cdots$$
$$\cdots 1_{v_i \leq y} \frac{1-\gamma}{1-u\gamma} \frac{W_\theta(v_{i-1}; u\gamma)}{W_\theta(v_i; u\gamma)} \left(\mathbb{E}\left((u\gamma)^{B^\theta}, H^\theta \leq v_i\right)\right)^{k-1} \mathbb{E}\left((u\gamma)^{B^\theta}, v_{i-1} < H^\theta \leq v_i\right),$$

whereas

$$\mathbb{E}\left(u^X, N_1(y;\gamma) = k \mid \ell_j = v_j, j \geq 1\right) = 1_{v_1 \leq y} \frac{1-\gamma}{1-u\gamma} \frac{1}{W_\theta(v_1; u\gamma)} \left(\mathbb{E}\left((u\gamma)^{B^\theta}, H^\theta \leq v_1\right)\right)^{k-1}.$$

It is well-known that for the Poisson point process of mutations,

$$\mathbb{P}(\ell_{i-1} \in dx, \ell_i \in dz) = \frac{\theta^i x^{i-2}}{(i-2)!} e^{-\theta z}\, dx\, dz \qquad 0 < x < z, i \geq 2,$$

so that

$$\sum_{i\geq 2} \mathbb{E}\left(u^X, N_i(y;\gamma) = k\right) = \frac{1-\gamma}{1-u\gamma} \sum_{i\geq 2} \int_0^y dz \int_0^z dx\, \frac{\theta^i x^{i-2}}{(i-2)!} e^{-\theta z} \frac{1}{W_\theta(z;u\gamma)} \left(1 - \frac{1}{W_\theta(z;u\gamma)}\right)^{k-1} F_\theta(x,z;u\gamma),$$

where

$$F_\theta(x,z;u\gamma) := W_\theta(x;u\gamma)\mathbb{E}\left((u\gamma)^{B^\theta}, x < H^\theta \leq z\right). \tag{3.4}$$

Since

$$\mathbb{E}\left(u^X, N_1(y;\gamma) = k\right) = \frac{1-\gamma}{1-u\gamma} \int_0^y dz\, \theta\, e^{-\theta z} \frac{1}{W_\theta(z;u\gamma)} \left(1 - \frac{1}{W_\theta(z;u\gamma)}\right)^{k-1},$$

we get

$$\sum_{i\geq 1} \mathbb{E}\left(u^X, N_i(y;\gamma) = k\right) = \frac{1-\gamma}{1-u\gamma} \int_0^y dz\, \theta\, e^{-\theta z} \frac{1}{W_\theta(z;u\gamma)} \left(1 - \frac{1}{W_\theta(z;u\gamma)}\right)^{k-1} \left[1 + \theta \int_0^z dx\, e^{\theta x} F_\theta(x,z;u\gamma)\right].$$

Now observe that

$$\begin{aligned} F_\theta(x,z;u\gamma) &= W_\theta(x;u\gamma)\left(\mathbb{E}\left((u\gamma)^{B^\theta}, H^\theta \leq z\right) - \mathbb{E}\left((u\gamma)^{B^\theta}, H^\theta \leq x\right)\right) \\ &= W_\theta(x;u\gamma)\left(\frac{1}{W_\theta(x;u\gamma)} - \frac{1}{W_\theta(z;u\gamma)}\right) \\ &= 1 - \frac{W_\theta(x;u\gamma)}{W_\theta(z;u\gamma)}, \end{aligned}$$

so that the integration by parts formula (3.2) yields

$$1 + \theta \int_0^z dx\, e^{\theta x} F_\theta(x,z;u\gamma) = 1 + \left[e^{\theta x}\left(1 - \frac{W_\theta(x;u\gamma)}{W_\theta(z;u\gamma)}\right)\right]_0^z + \int_0^z \frac{e^{\theta x}}{W_\theta(z;u\gamma)}\, dW_\theta(x;u\gamma),$$

where differentiation of W is understood w.r.t. the first variable. Since by Theorem 3.1, $dW_\theta(x;u\gamma) = e^{-\theta x} dW(x;u\gamma)$, we get

$$1 + \theta \int_0^z dx\, e^{\theta x} F_\theta(x,z;u\gamma) = \frac{W(z;u\gamma)}{W_\theta(z;u\gamma)}, \tag{3.5}$$

which ends the proof for the first formula. Let us turn to $Z_0(y;\gamma)$. The same kind of reasoning as previously shows that

$$\mathbb{E}\left(u^X, Z_0(y;\gamma) = k \mid \ell_j = v_j, j \geq 1\right) = \cdots$$
$$\cdots \sum_{i\geq 1} 1_{v_{i-1}<y<v_i} \mathbb{E}\left(u^{S_{D_i-1}}, X \geq S_{D_i-1}\right) \left(\mathbb{E}\left(u^{B^\theta}, B^\theta \leq X, H^\theta \leq y \mid H^\theta > v_{i-1}\right) 1_{i\geq 2} + 1_{i=1}\right) \times \cdots$$
$$\cdots \times \mathbb{E}\left(u^{B^\theta}, B^\theta \leq X, H^\theta \leq y\right)^{k-1} \left(\mathbb{E}\left(u^X, B^\theta > X\right) + \mathbb{E}\left(u^X, B^\theta \leq X, H^\theta > y\right)\right).$$

Referring to the calculations above, we easily get

$$\mathbb{E}\left(u^X, Z_0(y;\gamma) = k \mid \ell_j = v_j, j \geq 1\right) = \frac{1-\gamma}{1-u\gamma} \sum_{i\geq 1} 1_{v_{i-1}<y<v_i} \frac{1}{W_\theta(y;u\gamma)} \times \cdots$$
$$\cdots \times \left(1 - \frac{1}{W_\theta(y;u\gamma)}\right)^{k-1} \left[1_{i=1} + 1_{i\geq 2} W_\theta(v_{i-1};u\gamma) \mathbb{E}\left((u\gamma)^{B^\theta}, v_{i-1} < H^\theta \leq y\right)\right].$$

Integrating over the law of the Poisson point process of mutations yields

$$\mathbb{E}\left(u^X, Z_0(y;\gamma) = k\right) = \frac{1-\gamma}{1-u\gamma} e^{-\theta y} \frac{1}{W_\theta(y;u\gamma)} \left(1 - \frac{1}{W_\theta(y;u\gamma)}\right)^{k-1}$$
$$+ \frac{1-\gamma}{1-u\gamma} \sum_{i\geq 2} \int_y^\infty dz \int_0^y dx \, \frac{\theta^i x^{i-2}}{(i-2)!} e^{-\theta z} \frac{1}{W_\theta(y;u\gamma)} \left(1 - \frac{1}{W_\theta(y;u\gamma)}\right)^{k-1} F_\theta(x,y;u\gamma),$$

where F_θ was defined in (3.4). Thanks to equation (3.5), we get

$$\mathbb{E}\left(u^X, Z_0(y;\gamma) = k\right) = \frac{1-\gamma}{1-u\gamma} e^{-\theta y} \frac{1}{W_\theta(y;u\gamma)} \left(1 - \frac{1}{W_\theta(y;u\gamma)}\right)^{k-1} \left[1 + \theta \int_0^y dx \, e^{\theta x} F(x,y;u\gamma)\right]$$
$$= \frac{1-\gamma}{1-u\gamma} e^{-\theta y} \frac{W(y;u\gamma)}{W_\theta(y;u\gamma)^2} \left(1 - \frac{1}{W_\theta(y;u\gamma)}\right)^{k-1},$$

which is the desired formula. □

3.2.2 Proof of Theorem 3.3

Let $M_n(k,y;\gamma)$ denote the number of haplotypes whose most recent mutation occurred between time $-y$ and present time *on the n-th branch* (with i.i.d. lengths H_n, except $H_0 = +\infty$), and which are carried by k individuals among $\{0,1,\ldots,X\}$ (hence among $\{n,n+1,\ldots,X\}$). In particular,

$$A_\theta(k,y;\gamma) = \sum_{n\geq 0} M_n(k,y;\gamma).$$

First,

$$M_0(k,y;\gamma) = \sum_{i\geq 1} 1_{N_i(y,\gamma)=k},$$

so thanks to Lemma 3.8,

$$\mathbb{E}\left(u^X M_0(k,y;\gamma)\right) = \int_0^y dz \, F(k,z;u\gamma),$$

where we have used the following definition

$$F(k,z;u\gamma) := \frac{1-\gamma}{1-u\gamma}\,\theta\,e^{-\theta z}\,\frac{W(z;u\gamma)}{W_\theta(z;u\gamma)^2}\left(1-\frac{1}{W_\theta(z;u\gamma)}\right)^{k-1}.$$

Second, for all $n \geq 1$, by the lack-of-memory property of the geometric variable X,

$$\begin{aligned}\mathbb{E}\left(u^X M_n(k,y;\gamma)\right) &= u^n\mathbb{P}(X\geq n)\left[\int_0^y \mathbb{P}(H_n\in dx)\mathbb{E}\left(u^X M_0(k,x;\gamma)\right) + \mathbb{P}(H_n\geq y)\mathbb{E}\left(u^X M_0(k,y;\gamma)\right)\right]\\ &= (u\gamma)^n\left[\int_0^y \mathbb{P}(H\in dx)\int_0^x dz\,F(k,z;u\gamma) + \mathbb{P}(H\geq y)\int_0^y dz\,F(k,z;u\gamma)\right]\\ &= (u\gamma)^n\int_0^y dz\,F(k,z;u\gamma)\,\mathbb{P}(H\geq z).\end{aligned}$$

Now since $A_\theta(k,y;\gamma) = \sum_{n\geq 0} M_n(k,y;\gamma)$, we get

$$\begin{aligned}\mathbb{E}\left(u^X A_\theta(k,y;\gamma)\right) &= \int_0^y dz\,F(k,z;u\gamma) + \sum_{n\geq 1}(u\gamma)^n\int_0^y dz\,F(k,z;u\gamma)\,\mathbb{P}(H\geq z)\\ &= \int_0^y dz\,F(k,z;u\gamma)\left[1+\frac{u\gamma}{1-u\gamma}\,\mathbb{P}(H\geq z)\right]\\ &= \int_0^y dz\,F(k,z;u\gamma)\left[(1-u\gamma)W(z;u\gamma)\right]^{-1},\end{aligned}$$

hence the result, recalling the definition of F. □

4 Splitting trees: Expected haplotype frequencies at fixed time

4.1 Joint expected haplotype frequencies with population size distribution

In this subsection, we apply the results of the previous section to a splitting tree started at time $-t$ from one single individual and conditioned to be extant at present time 0. Then the population at present time is $\{0,1,\ldots,N_t-1\}$, where N_t is the population size and N_t-1 follows the geometric distribution with parameter

$$\gamma_t := \mathbb{P}(H\leq t) \qquad t>0,$$

that is, $\mathbb{P}_t(N_t-1\geq n) = \gamma_t^n$ for any integer $n\geq 0$, where $\mathbb{P}_t$ denotes the probability conditional on the population being extant at time 0, that is, t units of time after foundation. We recall that, in the case of splitting trees, the law of the branch lengths H is always absolutely continuous w.r.t. Lebesgue's measure.

The difference with the previous section is that the lengths of branches are (still i.i.d. but) *distributed as H conditional on $H\leq t$*. As a consequence, everything we have done in the previous section holds for the standing population of a splitting tree founded t units of time ago and conditioned upon survival up to t, replacing γ with γ_t and W with (from Theorem 3.1)

$$W^{(t)}(x;\alpha) := \frac{1}{1-\alpha\mathbb{P}(H\leq x\mid H\leq t)} \qquad x\in[0,t],\alpha\in(0,1].$$

In particular we now use $W_\theta^{(t)}$ instead of W_θ, with

$$W_\theta^{(t)}(x;\alpha) = e^{-\theta x} W^{(t)}(x;\alpha) + \theta \int_0^x dy\, W^{(t)}(y;\alpha)\, e^{-\theta y}.$$

Noticing that $W^{(t)}(x; u\gamma_t) = W(x;u)$, we also have $W_\theta^{(t)}(x; u\gamma_t) = W_\theta(x;u)$, where we stick to the notation from the previous section, namely,

$$W(x;u) = \frac{1}{1 - u\mathbb{P}(H \le x)} \qquad x \ge 0, u \in (0,1],$$

and

$$W_\theta(x;u) = e^{-\theta x} W(x;u) + \theta \int_0^x dy\, W(y;u)\, e^{-\theta y}.$$

We call a *derived haplotype* a haplotype which is different from the ancestral haplotype. Then the following statement stems readily from Theorem 3.3 and Lemma 3.8. Recall that $W(x) = W(x;1)$ and that $W_\theta(x) = W_\theta(x;1)$.

Proposition 4.1 *Let $A_\theta(k,t)$ denote the number of derived haplotypes represented by k individuals in the standing population of a splitting tree founded t units of time ago and $Z_0(t)$ the number of individuals in the standing population carrying the ancestral haplotype. Then for all $t \ge 0$ and $u \in (0,1]$,*

$$\mathbb{E}_t\left(u^{N_t - 1} A_\theta(k,t)\right) = \frac{W(t;u)^2}{W(t)} \int_0^t dx\, \theta\, e^{-\theta x} \frac{1}{W_\theta(x;u)^2} \left(1 - \frac{1}{W_\theta(x;u)}\right)^{k-1}.$$

and

$$\mathbb{E}_t\left(u^{N_t - 1}, Z_0(t) = k\right) = \frac{W(t;u)^2}{W(t)} \frac{e^{-\theta t}}{W_\theta(t;u)^2} \left(1 - \frac{1}{W_\theta(t;u)}\right)^{k-1}.$$

Remark 4.2 *Not to overload with notation, we have not considered the alleles of age less than y. If $A_\theta(k,y,t)$ denotes the number of derived haplotypes of age less than y, represented by k individuals in the standing population of a splitting tree founded t units of time ago, then we get the same formula as in the previous statement, but where the upper bound of the integral has changed*

$$\mathbb{E}_t\left(u^{N_t - 1} A_\theta(k,y,t)\right) = \frac{W(t;u)^2}{W(t)} \int_0^{y \wedge t} dx\, \theta\, e^{-\theta x} \frac{1}{W_\theta(x;u)^2} \left(1 - \frac{1}{W_\theta(x;u)}\right)^{k-1}.$$

The following corollary is obtained by taking $u = 1$ in the last statement. A more explanatory proof is given in the next subsection.

Corollary 4.3 *We have*

$$\mathbb{E}_t A_\theta(k,t) = W(t) \int_0^t dx\, \theta\, e^{-\theta x} \frac{1}{W_\theta(x)^2} \left(1 - \frac{1}{W_\theta(x)}\right)^{k-1}$$

and

$$\mathbb{P}_t\left(Z_0(t) = k\right) = W(t) \frac{e^{-\theta t}}{W_\theta(t)^2} \left(1 - \frac{1}{W_\theta(t)}\right)^{k-1}.$$

The same kinds of calculations as those done for the corollaries of the previous section yield the following statement, where the first equation could readily be deduced by exchangeability arguments.

Corollary 4.4 *Recall that $Z_0(t)$ is the number of individuals in the standing population carrying the ancestral type and set $A_\theta(t)$ the number of derived haplotypes represented in the standing population. Then for any positive real number t and positive integer n,*

$$\mathbb{E}(Z_0(t) \mid N_t = n) = n \exp(-\theta t)$$

and

$$\mathbb{E}(A_\theta(t) \mid N_t = n) = n \int_0^t dx\, \theta\, e^{-\theta x} \mathbb{E}\left(1 - \mathbb{P}(H \le t)^{-B^\theta} 1_{\{H^\theta \le x\}}\right) + \int_0^y dx\, \theta\, e^{-\theta x} \mathbb{E}\left(\left(B^\theta \wedge n\right) \mathbb{P}(H \le t)^{-B^\theta}, H^\theta \le x\right).$$

Proof. The first result is clear letting y go to $+\infty$ in Corollary 3.10. In view of (3.3) in Corollary 3.4, in order to prove the second result, we only need to check that

$$\tilde{\mathbb{P}}(H^\theta > x) = \mathbb{E}\left(1 - \mathbb{P}(H \le t)^{-B^\theta} 1_{\{H^\theta \le x\}}\right)$$

and

$$\tilde{\mathbb{E}}(B^\theta \wedge n, H^\theta \le x) = \mathbb{E}\left(\left(B^\theta \wedge n\right) \mathbb{P}(H \le t)^{-B^\theta}, H^\theta \le x\right),$$

where $\tilde{\mathbb{P}}$ is the law of the coalescent point process when the r.v. (H_i) are i.i.d. with common law $\mathbb{P}(H \in \cdot \mid H \le t)$. Now,

$$\begin{aligned}\tilde{\mathbb{P}}(H^\theta \le x) &= \mathbb{P}(H^\theta \le x \mid \forall i \le B^\theta,\ H_i \le t) \\ &= \sum_{k \ge 1} \mathbb{P}(B^\theta = k,\ H^\theta \le x) \mathbb{P}(H \le t)^{-k} \\ &= \mathbb{E}\left(1 - \mathbb{P}(H \le t)^{-B^\theta} 1_{\{H^\theta \le x\}}\right).\end{aligned}$$

The second equality, very similar, is left to the reader. □

Recall that $G_\theta(t)$ denotes the (absolute) homozygosity in the standing population, that is,

$$G_\theta(t) = \frac{Z_0(t)(Z_0(t)-1)}{2} + \sum_{k \ge 2} \frac{k(k-1)}{2} A_\theta(k,t),$$

then we easily get

Proposition 4.5 *For all $t \ge 0$ and $u \in (0,1]$,*

$$\mathbb{E}_t\left(u^{N_t - 1} G_\theta(t)\right) = \frac{W(t;u)^2}{W(t)} (W_{2\theta}(t;u) - 1).$$

Note that explicit formulas can also be obtained for the expectation of the standard homozygosity $\bar{G}_\theta(t) = 2G_\theta(t)/N_t(N_t-1)$, which is the probability that two randomly sampled individuals in the population at time t have the same haplotype. Formulas are given in Section 6, where they are obtained thanks to an alternative proof based on moment generating function computations.

Proof. We use Proposition 4.1 and the fact that $\sum_{k\geq 2} k(k-1)x^{k-2} = 2/(1-x)^3$. An integration by parts yields

$$\begin{aligned}\mathbb{E}_t\left(u^{N_t-1}G_\theta(t)\right) &= \frac{W(t;u)^2}{W(t)}\,e^{-\theta t}(W_\theta(t;u)-1) + \frac{W(t;u)^2}{W(t)}\int_0^t dx\,\theta\,e^{-\theta x}\,(W_\theta(x;u)-1)\\ &= \frac{W(t;u)^2}{W(t)}\,e^{-\theta t}(W_\theta(t;u)-1) + \frac{W(t;u)^2}{W(t)}\left(\Big[-e^{-\theta x}(W_\theta(x;u)-1)\Big]_0^t + \int_0^t dx\,e^{-\theta x}\,W'_\theta(x;u)\right),\end{aligned}$$

where differentiation is understood w.r.t. the first variable. Recalling that $W'_\theta(x;u) = e^{-\theta x}\,W'(x;u)$ provides the announced formula. □

4.2 An explanatory proof of Corollary 4.3

Consider the standing population at time t conditioned on being nonempty (probability measure $\mathbb{P}_t$). For any real number $y \in (0,t)$, for any non-negative integer i, let $C_i(y;dy)$, $D_i(y)$ and $E_i(y)$ denote the following events

$$C_i(y;dy) := \{i \leq N_t - 1,\ \text{the } i\text{-th branch length has size } H_i \geq y \text{ and carries a mutation with age in } (y, y+dy)\}$$

$$D_i(y) := \{\text{the type carried by the lineage of the } i\text{-th individual at time } t-y \text{ has at least one alive representative}\}$$

$$E_i(k,y) := \{\text{the type carried by the lineage of the } i\text{-th individual at time } t-y \text{ has } k \text{ alive representatives}\}$$

Then define $A_\theta(k,t,y;dy)$ as the number of haplotypes of age in the interval $(y, y+dy)$ represented by exactly k alive individuals at time t. Hereafter, we compute the expectation under $\mathbb{P}_t$ of $A_\theta(k,t,y;dy)$. The result will follow from the equality

$$A_\theta(k,t) = \int_0^t A_\theta(k,t,y;dy).$$

Now it is readily seen that

$$A_\theta(k,t,y;dy) = \sum_{i\geq 0} 1_{C_i(y;dy)\cap E_i(k,y)}$$

so that

$$\mathbb{E}_t A_\theta(k,t,y;dy) = \sum_{i\geq 0} \mathbb{P}_t(C_i(y;dy)\cap E_i(k,y)).$$

Next observe that $E_i(k,y) \subseteq D_i(y)$, so that

$$\begin{aligned}\mathbb{P}_t(C_i(y;dy)\cap E_i(k,y)) &= \mathbb{P}_t(C_i(y;dy))\mathbb{P}_t(D_i(y)\mid C_i(y;dy))\mathbb{P}_t(E_i(k,y)\mid D_i(y)\cap C_i(y;dy))\\ &= \mathbb{P}_t(C_i(y;dy))\mathbb{P}_t(D_0(y))\mathbb{P}_t(E_0(k,y)\mid D_0(y)).\end{aligned}$$

Thus, we record that

$$\mathbb{E}_t A_\theta(k,t,y;dy) = \mathbb{P}_t(D_0(y))\mathbb{P}_t(E_0(y) \mid D_0(y)) \sum_{i\geq 0} \mathbb{P}_t(C_i(y;dy)). \tag{4.1}$$

We will now prove the three following equalities

$$\sum_{i\geq 0} \mathbb{P}_t(C_i(y;dy)) = \theta\, dy\, \frac{W(t)}{W(y)}, \tag{4.2}$$

$$\mathbb{P}_t(D_0(y)) = \frac{W(y)\, e^{-\theta y}}{W_\theta(y)}, \tag{4.3}$$

$$\mathbb{P}_t(E_0(k,y) \mid D_0(y)) = \frac{1}{W_\theta(y)} \left(1 - \frac{1}{W_\theta(y)}\right)^{k-1}. \tag{4.4}$$

These three equalities, along with (4.1), yield the expected expression

$$\mathbb{E}_t A_\theta(k,t,y;dy) = \theta\, dy\, W(t) \frac{e^{-\theta y}}{W_\theta(y)^2} \left(1 - \frac{1}{W_\theta(y)}\right)^{k-1}, \tag{4.5}$$

which now sheds light on the meaning of each of the terms in the formula given in Corollary 4.3. Let us now prove equations (4.2), (4.3) and (4.4). First,

$$\begin{aligned}
\mathbb{P}_t(C_i(y;dy)) &= \mathbb{P}_t(N_t - 1 \geq i)\ \theta\, dy\ (1_{i=0} + 1_{i\geq 1}\mathbb{P}(H \geq y \mid H < t)) \\
&= \left(1 - \frac{1}{W(t)}\right)^i \theta\, dy\ \left(1_{i=0} + 1_{i\geq 1} \frac{\frac{1}{W(y)} - \frac{1}{W(t)}}{1 - \frac{1}{W(t)}}\right) \\
&= \theta\, dy\ \left[1_{i=0} + 1_{i\geq 1} \left(1 - \frac{1}{W(t)}\right)^{i-1} \left(\frac{1}{W(y)} - \frac{1}{W(t)}\right)\right],
\end{aligned}$$

so we get (4.2).

Second, let L denote an independent exponential r.v. with parameter θ, so that $(y-L)^+$ is the age of the oldest mutation on lineage 0 with age smaller than y, with the convention that this age is zero when there is no such mutation. Then either $L \geq y$, and $D_0(y)$ is realized because lineage 0 has carried the same type since time $t-y$, or $L < y$ and $D_0(y)$ is realized iff the next branch with no extra mutation than 0 for which the maximum of past branch lengths exceeds $t-L$ satisfies that this maximum does not exceed y (see Subsection 3.1). Conditional on $L = x$, this last event occurs with probability $\mathbb{P}(H_\theta \leq y \mid H_\theta > y - x)$. As a consequence, we get

$$\begin{aligned}
\mathbb{P}_t(D_0(y)) &= e^{-\theta y} + \int_0^y dx\, \theta\, e^{-\theta x} \left(1 - \frac{W_\theta(y-x)}{W_\theta(y)}\right) \\
&= 1 - \frac{1}{W_\theta(y)} \int_0^y dx\, \theta\, e^{-\theta x}\, W_\theta(y-x) \\
&= 1 - \frac{e^{-\theta y}}{W_\theta(y)} \int_0^y du\, \theta\, e^{\theta u}\, W_\theta(u),
\end{aligned}$$

and an integration by parts using the relationship between W and W_θ (see Remark 3.2) yields (4.3).

Finally, (4.4) stems from the definition of W_θ (see again Subsection 3.1).

5 Splitting trees: A.s. convergence of haplotype frequencies

In this section, we rely on the theory of random characteristics introduced in the seminal papers [11, 16] and further developed in [12, 13] and especially in [19], where the emphasis, as here, is on branching populations experiencing mutations (but there the mutation scheme is different, since mutation events occur simultaneously with births).

We will assume that the splitting tree starts at time 0 with one individual. Then recall from the last subsection that N_t denotes the number of individuals alive at time t, $A_\theta(t)$ denotes the number of derived haplotypes carried by alive individuals at time t, $A_\theta(k,t)$ denotes the number of derived haplotypes carried by k alive individuals at time t, and $Z_0(t)$ denotes the number of alive individuals at time t carrying the ancestral haplotype.

For any individual i, in the population, we let $\chi_i(t)$ (resp. $\chi_i^k(t)$) be the number of mutations that i has experienced during her lifetime that are carried by alive individuals (resp. by k alive individuals) t units of time after her birth ($\chi_i(t) = 0$ if $t < 0$). Then χ and the χ^k are random characteristics, in the sense given in the previously cited papers. In particular,

$$A_\theta(t) = \sum_i \chi_i(t - \sigma_i),$$

and

$$A_\theta(k,t) = \sum_i \chi_i^k(t - \sigma_i),$$

where σ_i denotes the birth time of i and the sum is taken over all individuals, dead or alive at time t, in the population. This allows us to make use of limit theorems for individuals counted by random characteristics proved in [11, 12, 13, 16], using the formulation of [19, Appendix A]. Different limit theorems hold depending whether the random characteristic is *individual*, in the sense that it only depends on the life history of the focal individual, or *general*, in the sense that it may also depend on the life history of the whole descendance of the focal individual, which is for example the case of χ and χ^k.

Recall that b is the birth rate of our homogeneous Crump–Mode–Jagers process, that V denotes a random lifetime duration, and that α denotes the Malthusian parameter, which satisfies $\psi(\alpha) = 0$, where ψ is defined in (2.2).

Let us restate the results in [19, Appendix A] in our setting. Set

$$\beta := \int_{(0,\infty]} u\, e^{-\alpha u} d\mu(u),$$

where the last integral is a Stieltjes integral w.r.t. the nondecreasing function

$$\mu(t) = \mathbb{E}(\#\text{ offspring born on } (0,t]) = b\mathbb{E}(t \wedge V) = \int_{(0,+\infty]} (r \wedge t)\Lambda(dr).$$

Also for any random characteristic, say χ, define $\widehat{\chi}(\alpha)$ as its Laplace transform at α

$$\widehat{\chi}(\alpha) := \int_{(0,+\infty)} dt\, e^{-\alpha t}\chi(t),$$

where it is implicit that χ is the characteristic of the progenitor (born at time 0). Hereafter, we apply Theorems 1 and 5 of [19, Appendix A], which apply to general random characteristics. These theorems need some technical assumptions to hold, which we verify at the end of the proof of the next statement. These theorems ensure first that

$$\lim_{t\to\infty} e^{-\alpha t}\mathbb{E}A_\theta(k,t) = \frac{\mathbb{E}\widehat{\chi^k}(\alpha)}{\beta}$$

and second that, on the survival event,

$$\lim_{t\to\infty} \frac{A_\theta(k,t)}{A_\theta(t)} = \frac{\mathbb{E}\widehat{\chi^k}(\alpha)}{\mathbb{E}\widehat{\chi}(\alpha)} \qquad a.s.$$

In addition to verifying the validity of the aforementioned technical assumptions, it remains to compute the quantities β, $\mathbb{E}\widehat{\chi}(\alpha)$ and $\mathbb{E}\widehat{\chi^k}(\alpha)$. With the following definitions, [**Les formules ci-dessous ne sont semble-t-il pas correctes. Voir les changements dans la preuve ci-dessous.**]

$$U_k := \int_0^\infty dx\,\theta\, e^{-\theta x}\,\frac{1}{W_\theta(x)^2}\left(1-\frac{1}{W_\theta(x)}\right)^{k-1},$$

and

$$U := \sum_{k\geq 1} U_k = \int_0^\infty dx\,\theta\, e^{-\theta x}\,\frac{1}{W_\theta(x)},$$

we have $\beta = \psi'(\alpha)/\alpha$, $\mathbb{E}\widehat{\chi^k}(\alpha) = U_k/b$ and of course $\mathbb{E}\widehat{\chi}(\alpha) = U/b$. This can be recorded in the following proposition.

Proposition 5.1 *In the supercritical case,*

$$\lim_{t\to\infty} e^{-\alpha t}\mathbb{E}A_\theta(k,t) = \frac{\alpha U_k}{b\psi'(\alpha)} \tag{5.1}$$

and

$$\lim_{t\to\infty} e^{-\alpha t}\mathbb{E}A_\theta(t) = \frac{\alpha U}{b\psi'(\alpha)}. \tag{5.2}$$

And on the survival event,

$$\lim_{t\to\infty} \frac{A_\theta(k,t)}{A_\theta(t)} = \frac{U_k}{U} \qquad a.s.$$

Remark 5.2 *Note that it can be shown similarly that*

$$\lim_{t\to\infty} e^{-\alpha t}\mathbb{E}N_t = \frac{\alpha}{b\psi'(\alpha)},$$

and that, for example,

$$\lim_{t\to\infty} \frac{A_\theta(t)}{N_t} = U \quad a.s.$$

This is reminiscent of Theorem 3.2 in [14] where the same limit is obtained after conditioning on the population size to equal n and letting $n\to\infty$. This a.s. convergence is made possible by embedding all populations of fixed size on the same space thanks to an infinite coalescent point process: the population of size n is that generated by the first n values of the coalescent point process.

Remark 5.3 *In [15], it is proved in the supercritical case ($\alpha > 0$) that the survival probability is α/b and that the scale function W has the following asymptotic behaviour*

$$\lim_{t\to\infty} W(t)e^{-\alpha t} = \frac{1}{\psi'(\alpha)}.$$

One could have used these two facts and the monotone convergence theorem to recover (5.1) and (5.2) from Corollary 4.3. In the following proof, we prefer to show the agreement with Corollary 4.3 by computing directly β, $\mathbb{E}\widehat{\chi}(\alpha)$ and $\mathbb{E}\widehat{\chi^k}(\alpha)$.

Proof. Let us first prove that $\beta = \psi'(\alpha)/\alpha$. Recalling the definition of β, we get

$$\begin{aligned}
\beta &= b\mathbb{E}\int_0^\infty du\, ue^{-\alpha u}1_{\{u<V\}} \\
&= \int_{(0,+\infty]} \Lambda(dr)\int_0^r du\, ue^{-\alpha u} \\
&= \frac{1}{\alpha^2}\int_{(0,+\infty]} \Lambda(dr)\left(1-e^{-\alpha r}-\alpha r e^{-\alpha r}\right) \\
&= \frac{1}{\alpha^2}(\alpha-\psi(\alpha)) - \frac{1}{\alpha}(1-\psi'(\alpha)) \\
&= \frac{\psi'(\alpha)}{\alpha}.
\end{aligned}$$

Next let us compute $\mathbb{E}\widehat{\chi^k}(\alpha)$. Denote by $R_t^{(a,b)}$ the number of individuals alive at time t descending clonally from the time interval (a,b). More specifically, for a progenitor individual alive on the time interval (a,b) and experiencing no mutation between times a and b, $R_t^{(a,b)}$ is the number of individuals alive at t (including possibly this progenitor) descending from those daughters of the progenitor who were born during the time interval (a,b), and that still carry the same type that the progenitor carried at time a. Conditionally on the event that this progenitor has alive descendants at time t, the genealogy of the clonal descendants alive at time t is given by the coalescent point process associated with the r.v. $(H_i^\theta)_{i\geq 1}$ constructed in Section 3.1. Hence, writing wis a slight abuse of notation $\mathbb{P}_t$ for the conditional probability on the survival of the (not necessarily clonal) descendance of the time interval (a,b),

$$\begin{aligned}
\mathbb{P}_t\left(R_t^{(a,b)}=k\right) &= \mathbb{P}_{t-a}(N_{t-a}^\theta = k \mid \zeta = b-a) \\
&= \mathbb{P}_{t-a}(N_{t-a}^\theta \neq 0 \mid \zeta = b-a)\mathbb{P}_{t-a}(N_{t-a}^\theta = k \mid N_{t-a}^\theta \neq 0) \\
&= \left(1-1_{t>b}\frac{W_\theta(t-b)}{W_\theta(t-a)}\right)\left(1-\frac{1}{W_\theta(t-a)}\right)^{k-1}\frac{1}{W_\theta(t-a)}, \qquad (5.3)
\end{aligned}$$

where N^θ is the population size process of a clonal splitting tree and ζ is the lifetime of the progenitor. In addition, since the jump contour process of the splitting tree is a Lévy process without negative jumps with scale function W,

$$\mathbb{P}\Big(\text{the time interval } (a,b) \text{ has alive descendants at time } t\Big) = \frac{W((b\wedge t)-a)}{W(t-a)}. \qquad (5.4)$$

Now let us start with a progenitor with lifetime distributed as V and denote by ℓ_i the time of the i-th point of a Poisson point process with intensity θ (the i-th mutation of the progenitor). Then

$$\begin{aligned}
\mathbb{E}\widehat{\chi^k}(\alpha) &= \mathbb{E}\int_0^\infty dt\, e^{-\alpha t} \sum_{i\geq 1} 1_{\{\ell_i < V\wedge t\}}\, 1\left(R_t^{(\ell_i, V\wedge \ell_{i+1})} = k\right) \\
&= \mathbb{E}\int_0^\infty dt\, e^{-\alpha t} \sum_{i\geq 1} \int_0^\infty dz\, e^{-\theta z} \int_0^z dy\, \frac{\theta^{i+1} y^{i-1}}{(i-1)!}\, 1_{\{y<V\wedge t\}}\, 1\left(R_t^{(y,V\wedge z)} = k\right) \\
&= \mathbb{E}\int_0^\infty dt\, e^{-\alpha t} \int_0^\infty dz\theta\, e^{-\theta z} \int_0^{z\wedge V\wedge t} dy\, \theta\, e^{\theta y}\, 1\left(R_t^{(y,V\wedge z)} = k\right) \\
&= \mathbb{E}\int_0^\infty dt\, e^{-\alpha t} \int_0^{V_\theta\wedge t} dy\, \theta\, e^{\theta y}\, 1\left(R_t^{(y,V_\theta)} = k\right),
\end{aligned}$$

where V_θ denotes the minimum of V and of an independent exponential r.v. with parameter θ. Then

$$\begin{aligned}
\mathbb{E}\widehat{\chi^k}(\alpha) &= \int_0^\infty dt\, e^{-\alpha t} \int_{(0,\infty)} \mathbb{P}(V_\theta \in du) \int_0^t dy\, 1_{\{y<u\}}\, \theta\, e^{\theta y}\, \mathbb{P}\left(R_t^{(y,u)} = k\right) \\
&= \int_0^\infty dt\, e^{-\alpha t} \int_{(0,\infty)} \mathbb{P}(V_\theta \in du) \int_0^t dx\, 1_{\{t-x<u\}} \theta\, e^{\theta(t-x)}\, \mathbb{P}\left(R_t^{(t-x,u)} = k\right) \\
&= \int_0^\infty dx\, \theta\, e^{-\theta x} \int_{(0,\infty)} \mathbb{P}(V_\theta \in du) \int_x^{u+x} dt\, e^{(\theta-\alpha)t}\, \mathbb{P}\left(R_t^{(t-x,u)} = k\right),
\end{aligned}$$

which, thanks to (5.3) and (5.4), yields **[j'ai corrigé jusqu'à la seconde ligne ci-dessous. Après, je sèche pour le calcul...]**

$$\begin{aligned}
\mathbb{E}\widehat{\chi^k}(\alpha) &= \int_0^\infty dx\, \frac{\theta\, e^{-\theta x}}{W(x)W_\theta(x)} \left(1 - \frac{1}{W_\theta(x)}\right)^{k-1} \int_{(0,\infty)} \mathbb{P}(V_\theta \in du) \\
&\qquad \int_x^{u+x} dt\, e^{(\theta-\alpha)t}\, W(u\wedge t + x - t) \left(1 - 1_{t>u} \frac{W_\theta(t-u)}{W_\theta(x)}\right) \\
&= \int_0^\infty dx\, \frac{\theta\, e^{-\theta x}}{W_\theta(x)} \left(1 - \frac{1}{W_\theta(x)}\right)^{k-1} \left(F_1(x) - \frac{F_2(x)}{W_\theta(x)}\right),
\end{aligned}$$

where

$$F_1(x) := \int_{(0,\infty)} \mathbb{P}(V_\theta \in du) \int_x^{u+x} dt\, e^{(\theta-\alpha)t}$$

and

$$F_2(x) := \int_{(0,\infty)} \mathbb{P}(V_\theta \in du) \int_x^{u+x} dt\, e^{(\theta-\alpha)t} 1_{t>u} W_\theta(t-u).$$

Let us compute F_1 and F_2. Set

$$\psi_\theta(x) := x - \int_{(0,\infty)} \left(1 - e^{-rx}\right)\, b\, \mathbb{P}(V_\theta \in dr) \qquad x \geq 0.$$

Then [14] $\psi_\theta(x) = x\psi(x+\theta)/(x+\theta)$, and $1/\psi_\theta$ is the Laplace transform of W_θ. Also recall that $\psi(\alpha) = 0$, so that $\psi_\theta(\alpha-\theta) = 0$. First, if $\theta = \alpha$, then $F_1(x) = \int_{(0,\infty)} u\, \mathbb{P}(V_\theta \in du) = (1-\psi'_\alpha(0+))/b =$

$1/b$. Second, if $\theta \neq \alpha$, then

$$F_1(x) = \frac{e^{(\theta-\alpha)x}}{\alpha-\theta} \int_{(0,\infty)} \mathbb{P}(V_\theta \in du) \left(1 - e^{-(\alpha-\theta)u}\right) = \frac{e^{(\theta-\alpha)x}}{b(\alpha-\theta)}(\alpha - \theta - \psi_\theta(\alpha-\theta)),$$

so that whatever the respective values of α and θ,

$$F_1(x) = \frac{1}{b} e^{(\theta-\alpha)x}.$$

We use Laplace transforms to compute F_2. For any $\kappa > 0$,

$$\begin{aligned}
\int_0^\infty dx\, \kappa\, e^{-\kappa x} F_2(x) &= \int_{(0,\infty)} \mathbb{P}(V_\theta \in du) \int_u^\infty dt\, e^{(\theta-\alpha)t} W_\theta(t-u) \int_{t-u}^t dx\, \kappa\, e^{-\kappa x} \\
&= \int_{(0,\infty)} \mathbb{P}(V_\theta \in du)\, (e^{\kappa u} - 1) \int_u^\infty dt\, e^{(\theta-\alpha-\kappa)t} W_\theta(t-u) \\
&= \int_{(0,\infty)} \mathbb{P}(V_\theta \in du)\, (e^{\kappa u} - 1)\, e^{(\theta-\alpha-\kappa)u} \int_0^\infty ds\, e^{(\theta-\alpha-\kappa)s} W_\theta(s) \\
&= \frac{1}{b}(\kappa + \alpha - \theta - \psi_\theta(\kappa+\alpha-\theta) - (\alpha - \theta - \psi_\theta(\alpha-\theta))) \frac{1}{\psi_\theta(\kappa+\alpha-\theta)} \\
&= \frac{\kappa}{b\psi_\theta(\kappa+\alpha-\theta)} - \frac{1}{b},
\end{aligned}$$

so that

$$F_2(x) = \frac{1}{b} e^{(\theta-\alpha)x} W_\theta(x) - \frac{1}{b},$$

and

$$F_1(x) - \frac{F_2(x)}{W_\theta(x)} = \frac{1}{bW_\theta(x)}.$$

As a consequence, we get

$$\mathbb{E}\widehat{\chi^k}(\alpha) = \int_0^\infty dx\, \frac{\theta\, e^{-\theta x}}{bW_\theta(x)^2} \left(1 - \frac{1}{W_\theta(x)}\right)^{k-1},$$

which is the announced U_k/b.

Last, let us check the technical assumptions required for Theorems 1 and 5 in [19, Appendix A] to hold. For the first theorem, we have to check the following two requirements

$$\sum_{n\geq 0} \sup_{[n,n+1]} e^{-\alpha u} \mathbb{E}\chi(u) < \infty \tag{5.5}$$

$$t \mapsto \mathbb{E}\chi(t) \text{ is a.e. continuous.} \tag{5.6}$$

For the second theorem, we have to check the following two requirements

$$\exists\ 0 < \eta < \alpha,\ \mathbb{E} \sup_{t\geq 0} e^{-\eta t} \chi(t) < \infty \tag{5.7}$$

$$\exists\ 0 < \eta < \alpha,\ \hat{\mu}(\eta) < \infty. \tag{5.8}$$

The following equality in distribution is easily seen

$$\chi(t) = \sum_{i\geq 1} 1_{\{T_i \leq t\wedge V\}} 1_{\{\sum_{j\geq 1} N_j(t-S_j) 1_{\{T_i<S_j<T_{i+1}\wedge t\wedge V\}} \in A\}},$$

where V is distributed as a lifetime, the (T_i) are the ranked atoms of an independent Poisson point process with rate θ (mutation times), the (S_i) are the ranked atoms of an independent Poisson point process with rate b (birth times), the (N_i) form an independent sequence of i.i.d. homogeneous, binary CMJ processes (descendances of daughters), and A is taken equal to $\mathbb{N}$, but can be taken equal to $\{k\}$ in the case of the random characteristic χ^k. In any case, χ is dominated by a Poisson point process with rate θ, so that $\mathbb{E}\chi(t) \leq \theta t$. This ensures that (5.5) holds. As for (5.6), notice from the last displayed equation that $\mathbb{E}\chi(t) = \sum_{i\geq 1} F_i(t)$, where

$$F_i(t) := \int_0^t \int_u^\infty \mathbb{P}(T_i \in du, T_{i+1} \in ds) \int_{[u,\infty)} \mathbb{P}(V \in dr) \mathbb{P}\left(\sum_{j\geq 1} N_j(t-S_j) 1_{\{u<S_j<s\wedge t\wedge r\}} \in A \right).$$

Because T_i has a density w.r.t. Lebesgue measure, each F_i is everywhere continuous on, say, $[0, t_0]$. In addition, for any $t \in [0, t_0]$, $F_i(t) \leq \mathbb{P}(T_i \leq t) \leq \mathbb{P}(T_i \leq t_0)$ and $\sum_{i\geq 1} \mathbb{P}(T_i \leq t_0) = \theta t_0 < \infty$, so we get continuity of $t \mapsto \mathbb{E}\chi(t)$ on $[0, t_0]$ by dominated convergence. Because t_0 is arbitrary, $t \mapsto \mathbb{E}\chi(t)$ is continuous everywhere.

Let us treat the last two requirements. The last requirement (5.8) merely stems from the obvious inequality $\mu(t) \leq bt$. To prove (5.7), because χ is dominated by a Poisson point process, it suffices to show that for any Poisson point process Y with rate 1, say, and for any $\eta > 0$, $\mathbb{E}\sup_{t\geq 0} e^{-\eta t} Y_t < \infty$. In fact, setting $M_c(t) := e^{-\eta t}(Y_t + c)$, we claim that for large enough c, M_c^2 is a supermartingale. Then using the inequality $\mathbb{P}(\sup_t M_c^2(t) \geq z) \leq c/z$, we get

$$\mathbb{P}(\sup_t\ Y_t\, e^{-\eta t} \geq y) \leq \mathbb{P}(\sup_t\ (Y_t + c)\, e^{-\eta t} \geq y) = \mathbb{P}(\sup_t\ M_c^2(t) \geq y^2) \leq \frac{c}{y^2},$$

so that $\mathbb{E}(\sup_t\ Y_t\, e^{-\eta t}) < \infty$. The only thing left to show is that M_c^2 is a supermartingale. Writing $(\mathcal{F}_t)$ for the natural filtration of Y and P_s for a Poisson random variable with parameter s independent of Y_t, we get

$$\mathbb{E}(M_c(t+s)^2 \mid \mathcal{F}_t) = e^{-2\eta(t+s)} \mathbb{E}\left((Y_t + c + P_s)^2\right) = e^{-2\eta(t+s)} \left((Y_t + c + s)^2 + s\right) \leq M_c(t)^2,$$

where the last inequality holds for any $s, t \geq 0$ if there is some positive c (depending only on η) such that

$$e^{-2\eta s}\left((x+s)^2 + s\right) \leq x^2 \qquad x \geq c, s \geq 0.$$

Then we study the function $f : s \mapsto x^2 e^{2\eta s} - (x+s)^2 - s$. Since $f''(s) = 4\eta^2 x^2 e^{2\eta s} - 2$, f' is nondecreasing on $[0, +\infty)$ as soon as $x^2 \geq 1/2\eta^2$. On the other hand, $f'(0) = 2\eta x^2 - 1 - 2x$. Let $x^\star$ be the largest root of $x \mapsto 2\eta x^2 - 1 - 2x$. As soon as $x \geq x^\star$, $f'(0) \geq 0$. Setting $c := \max(1/\eta\sqrt{2}, x^\star)$, as soon as $x \geq c$, $f'(0) \geq 0$ and f' is nondecreasing on $[0, \infty)$, so that f is nondecreasing on $[0, \infty)$. Since $f(0) = 0$, we conclude that f is non-negative on $[0, \infty)$, so that M_c^2 indeed is a supermartingale. □

6 Expected homozygosities through moment generating functions

We consider again the coalescent point process of Section 3, constructed from $H_0 = +\infty$ and the i.i.d. sequence of r.v. $(H_i)_{i\geq 1}$, with common law $\mathbb{P}(H \in \cdot)$. Let us recall that, in the case of splitting trees, the law of H has a density w.r.t. Lebesgue's measure. We introduce the derivative of $\log W(t)$:

$$p(t)dt = \mathbb{P}(H \leq t + dt \mid H > t) = W(t)\mathbb{P}(H \in dt). \tag{6.1}$$

For any time t, we consider the splitting tree obtained from $H_0, \ldots, H_{N_t-1}$, where $N_t := \inf\{i \geq 1 : H_i > t\}$. We then define the (standard) homozygosity $\bar{G}_\theta(t)$ as the probability that two *distinct* randomly sampled individuals in the population at time t share the same haplotype, and the *absolute* homozygosity $G_\theta(t)$ as the number of pairs of *distinct* individuals in the population at time t that share the same haplotype. Note that both of these quantities are 0 on the event $\{N_t = 1\}$, and on the complement event,

$$\bar{G}_\theta(t) = \frac{2G_\theta(t)}{N_t(N_t - 1)}. \tag{6.2}$$

The notation $G_\theta(t)$ coincides with that of Subsection 4.1. We also recall that $Z_0(t)$ denotes the number of individuals sharing the ancestral haplotype, defined here as the haplotype of individual 0 at time $-t$.

Our goal in this section is to compute $\mathbb{E}_t(G_\theta(t))$ and $\mathbb{E}_t(\bar{G}_\theta(t))$ using another method than in Section 3. As in [14], we characterize the joint law of $(G_\theta(t), N_t, Z_0(t))$ as time increases in a similar fashion as for branching processes, in order to obtain backward Kolmogorov equations for moment generating functions involving these random variables. The result will then follow by solving these equations.

Proposition 6.1 *For all $t \geq 0$, the expected* absolute *homozygosity is given by*

$$\mathbb{E}_t\left(G_\theta(t)\right) = W(t)(W_{2\theta}(t) - 1),$$

whereas the expected standard *homozygosity is given by*

$$\mathbb{E}_t(\bar{G}_\theta(t)) = \frac{e^{-2\theta t}(W(t) - 1)}{2W(t)} + 2\theta \int_0^t e^{-2\theta s} \frac{W(s) - 1}{W(t) - W(s)} \left[\frac{\log W(t) - \log W(s)}{W(t) - W(s)} - \frac{1}{W(t)}\right] ds.$$

6.1 Joint dynamics of $G_\theta(t)$, N_t and $Z_0(t)$

Consider two splitting trees of age t, with respective absolute homozygosity, population size, number of ancestral individuals and height processes $G_\theta(t)$, N_t, $Z_0(t)$, $(H_i)_{i\geq 0}$ and $G'_\theta(t)$, N'_t, $Z'_0(t)$, $(H'_i)_{i\geq 0}$. We call *merger* of these two splitting trees the splitting tree obtained from the sequence of heights $H_0 = +\infty, H_1, \ldots, H_{N_t-1}, H''_0, H'_1, \ldots, H'_{N'_t-1}$, where H''_0 is obtained from the infinite branch H'_0 by cutting the part below $-t$. In addition, all the mutation times are kept unchanged on each branch of the tree.

After this merger event, the new splitting tree has population size $N_t + N'_t$, the new number of ancestral individuals is $Z_0(t) + Z'_0(t)$ and the new absolute homozygosity is, counting first the pairs

of ancestral individuals

$$\frac{(Z_0(t)+Z_0'(t))(Z_0(t)+Z_0'(t)-1)}{2}+G_\theta(t)-\frac{Z_0(t)(Z_0(t)-1)}{2}+G_\theta'(t)-\frac{Z_0'(t)(Z_0'(t)-1)}{2}$$
$$=G_\theta(t)+G_\theta'(t)+Z_0(t)Z_0'(t).$$

Now, we have $(G_\theta(0),N_0,Z_0(0))=(0,1,1)$ and, if the law of $(G_\theta(t),N_t,Z_0(t))$ is known for some $t\geq 0$, then, on the time interval $[t,t+dt]$,

- either a mutation occurs on the ancestral branch, with probability $\theta\,dt$, and
$$(G_\theta(t+dt),N_{t+dt},Z_0(t+dt))=(G_\theta(t),N_t,0),$$
- or $H_{N_t}\in[t,t+dt]$, with probability $p(t)dt$ defined in (6.1), and
$$(G_\theta(t+dt),N_{t+dt},Z_0(t+dt))=(G_\theta(t)+G_\theta'(t)+Z_0(t)Z_0'(t),N_t+N_t',Z_0(t)+Z_0'(t)),$$
where $(G_\theta'(t),N_t',Z_0'(t))$ is an i.i.d. copy of $(G_\theta(t),N_t,Z_0(t))$,
- or nothing happens (the probability that both previous events occurs is $o(dt)$).

In other words, when the ancestral time t increases, the process $(G_\theta(t),N_t,Z_0(t))$ jumps to $(G_\theta(t),N_t,0)$ with rate θ and to $(G_\theta(t)+G_\theta'(t)+Z_0(t)Z_0'(t),N_t+N_t',Z_0(t)+Z_0'(t))$ with instantaneous rate $p(t)$.

Of course, the previous argument is quite informal, but it could easily be made rigorous by considering all the possible events that could occur in the time interval $[t,t+s]$, and letting $s\to 0$. In particular, the Kolmogorov equations of the following subsection can easily be justified this way.

6.2 Moment generating functions computations

We define the moment generating functions

$$L(t,u)=\mathbb{E}_t(G_\theta(t)u^{N_t-2}) \tag{6.3}$$
$$M(t,u,v)=\mathbb{E}_t(u^{N_t-1}v^{Z_0(t)}), \tag{6.4}$$

for all $u,v\in[-1,1]$ and $t\geq 0$. Since $G_\theta(t)=0$ if $N_t\leq 1$ and the quantities inside the expectations are bounded by N_t^2, these functions have finite values. Our goal here is to compute explicit expressions for these quantities.

Note that, for any i.i.d. triples of nonnegative r.v. (G_θ,N,Z_0) and (G_θ',N',Z_0'),

$$\mathbb{E}((G_\theta+G_\theta'+Z_0Z_0')u^{N+N'-2})=2\mathbb{E}(G_\theta u^{N-2})\mathbb{E}(u^N)+\left(\mathbb{E}(Z_0u^{N-1})\right)^2.$$

Using this equation and the previous construction of the process, we can write the forward Kolmogorov equation for the moment generating functions L and M: for all $u,v\in[-1,1]$ and $t\geq 0$,

$$\begin{cases}\partial_tL(t,u)=-(\theta+p(t))L(t,u)+\theta\,L(t,u)+p(t)\Big[2\,u\,L(t,u)\,M(t,u,1)+(\partial_vM(t,u,1))^2\Big]\\ L(0,u)=0,\end{cases} \tag{6.5}$$

and

$$\begin{cases} \partial_t M(t,u,v) = -(\theta + p(t))M(t,u,v) + \theta\, M(t,u,1) + p(t)\, u\, (M(t,u,v))^2 \\ M(0,u,v) = v. \end{cases} \tag{6.6}$$

The explicit computation of the solutions of these equations requires several steps. First, for fixed u and v, the function $M(t,u,v)$ is solution to an ODE known as Riccati's equation. In the case where $v = 1$, the function $f(t) = M(t,u,1)$ is solution to

$$\dot{f} = pf(uf - 1),$$

which is known as Bernoulli's equation. It can be solved by making the change of unknown function $\tilde{f} = 1/f$, which makes the ODE linear. This yields

$$f(t) = M(t,u,1) = \left(u + (1-u)\exp\int_0^t p(s)ds\right)^{-1} = \frac{W(t;u)}{W(t)}, \tag{6.7}$$

where we used that p is the derivative of the function $\log W(t)$.

Second, for all $u, v \in [-1,1]$, the function $M(t,u,1)$ is a particular solution of (6.6) (with different initial condition). Hence, the function $g(t) = M(t,u,v) - M(t,u,1) = M(t,u,v) - f(t)$ solves the Bernoulli ODE

$$\dot{g} = -(\theta + p - 2upf)g + upg^2,$$

for which the previous trick again works. This yields

$$M(t,u,v) = f(t) + \frac{\exp\left(-\int_0^t(\theta + p(s) - 2up(s)f(s))ds\right)}{(v-1)^{-1} - u\int_0^t p(s)\exp\Big(-\int_0^s(\theta + p(\tau) - 2up(\tau)f(\tau))d\tau\Big)ds}.$$

Since $uW(s;u)\mathbb{P}(H \in ds)$ is the derivative of $\log W(\cdot;u)$, it follows from (6.7) that

$$\int_0^t p(s)(1 - 2uf(s))ds = \log W(t) - 2\log W(t;u). \tag{6.8}$$

Hence, we obtain

$$M(t,u,v) = \frac{W(t;u)}{W(t)}\left(1 + \frac{e^{-\theta t}\, W(t;u)}{(v-1)^{-1} - u\int_0^t e^{-\theta s}W(s;u)^2\mathbb{P}(H \in ds)}\right).$$

Observing that $uW(s;u)^2\mathbb{P}(H \in ds)$ is the derivative of $W(\cdot;u)$, an integration by parts and Theorem 3.1 finally yield

$$M(t,u,v) = \frac{W(t;u)}{W(t)}\left(1 - \frac{e^{-\theta t}\, W(t;u)}{\frac{v}{1-v} + W_\theta(t;u)}\right).$$

We then compute

$$M(t,u,1) = \frac{W(t;u)}{W(t)} = f(t) \quad \text{and} \quad \partial_v M(t,u,1) = \frac{W(t;u)^2\, e^{-\theta t}}{W(t)} =: q(t).$$

Third, the linear equation (6.5) can be explicitly solved:

$$L(t,u) = \exp\left(-\int_0^t p(s)(1-2uf(s))ds\right)\int_0^t p(s)q^2(s)\exp\left(\int_0^s p(\tau)(1-2uf(\tau))d\tau\right)ds.$$

Using (6.8) again, we obtain

$$L(t,u) = \frac{W(t;u)^2}{W(t)}\int_0^t e^{-2\theta s}\,W(s;u)^2\,\mathbb{P}(H\in ds).$$

Using integration by parts as above finally yields

$$L(t,u) = \frac{W(t;u)^2}{W(t)}\,\frac{W_{2\theta}(t;u)-1}{u}, \tag{6.9}$$

which is consistent with Proposition 4.5.

Fourth, using Theorem 3.1, we have

$$\frac{W_{2\theta}(t;u)-1}{u} = e^{-2\theta t}\,\frac{W(t;u)-1}{u} + 2\theta\int_0^t e^{-2\theta s}\,\frac{W(s;u)-1}{u}du.$$

This yields

$$L(t,u) = \frac{W(t;u)^2}{W(t)}\left[e^{-2\theta t}\,\mathbb{P}(H\le t)\,W(t;u) + 2\theta\int_0^t e^{-2\theta s}\,\mathbb{P}(H\le s)\,W(s;u)ds\right].$$

Writing the product series of $(1-v)^{-1} = \sum_{n\ge 0} v^n$ and $(1-v)^{-2} = \sum_{n\ge 0}(n+1)v^n$ and observing that

$$\sum_{k=0}^n (k+1)a^k b^{n-k} = \frac{d}{da}\left(a\sum_{k=0}^n a^k b^{n-k}\right) = \frac{(n+1)a^{n+2} - (n+2)a^{n+1}b + b^{n+2}}{(a-b)^2},$$

we get

$$L(t,u) = \frac{e^{-2\theta t}\mathbb{P}(H\le t)}{2W(t)}\sum_{n\ge 2} n(n-1)(\mathbb{P}(H\le t)u)^{n-2} + \frac{2\theta}{W(t)}\int_0^t ds e^{-2\theta s}\mathbb{P}(H\le s)\times$$
$$\sum_{n\ge 0 u^2}\frac{(n+1)\mathbb{P}(H\le t)^{n+2} - (n+2)\mathbb{P}(H\le t)^{n+1}\mathbb{P}(H\le s) + \mathbb{P}(H\le s)^{n+2}}{\mathbb{P}(s<H\le t)^2}\,u^n. \tag{6.10}$$

Finally, we compute the expected standard homozygosity as follows: by (6.2),

$$\partial_u^2\left(\mathbb{E}\left(\bar{G}_\theta(t)u^{N_t}\right)\right) = L(t,u), \quad \text{or} \quad \mathbb{E}(\bar{G}_\theta(t)) = \int_0^1 du\int_0^u dv\,L(t,v).$$

Integrating (6.10) twice and using the equation

$$(1-x)\log(1-x) + x = \sum_{n\ge 2}\frac{x^n}{n(n-1)}$$

yields

$$\mathbb{E}_t[\bar{G}_\theta(t)] = \frac{e^{-2\theta t}(W(t)-1)}{2W(t)} + 2\theta\int_0^t ds\,e^{-2\theta s}\,\frac{W(s)-1}{W(t)-W(s)}\left[\frac{\log\frac{W(t)}{W(s)}}{W(t)-W(s)} - \frac{1}{W(t)}\right],$$

which ends the proof of Proposition 6.1.

Acknowledgments. This work was funded by projects MAEV 'Modèles Aléatoires de l'Évolution du Vivant' 06-BLAN-3 146282 and MANEGE 'Modèles Aléatoires en Écologie, Génétique et Évolution' 09-BLAN-0215 of ANR (French national research agency)

References

[1] Aldous, D., Popovic, L. (2005)
A critical branching process model for biodiversity. *Adv. in Appl. Probab.* **37(4)** 1094–1115.

[2] Bertoin, J. (2009)
The structure of the allelic partition of the total population for Galton-Watson processes with neutral mutations. *Ann. Probab.* **37** 1502–1523.

[3] Bertoin, J. (2010)
A limit theorem for trees of alleles in branching processes with rare neutral mutations. *Stoch. Proc. Appl.* **120** 678–697.

[4] Bertoin, J. (2010)
Asymptotic regimes for the partition into colonies of a branching process with emigration. *Ann. Appl. Probab.* **20** 1967–1988.

[5] Champagnat, N., Lambert, A. (2010)
Splitting trees with neutral Poissonian mutations II: Large or old families. In preparation.

[6] Durrett, R., Moseley, S. (2010)
Evolution of resistance and progression to disease during clonal expansion of cancer. *Theoret. Popul. Biol.* **77** 42–48.

[7] Ewens, W.J. (2005)
Mathematical Population Genetics. 2nd edition, Springer–Verlag, Berlin.

[8] Geiger, J. (1996)
Size-biased and conditioned random splitting trees. *Stoch. Proc. Appl.* **65** 187–207.

[9] Geiger, J., Kersting, G. (1997)
Depth-first search of random trees, and Poisson point processes, in *Classical and modern branching processes* (Minneapolis, 1994) IMA Math. Appl. Vol. 84. Springer-Verlag, New York.

[10] Griffiths, R.C., Pakes, A.G. (1988)
An infinite-alleles version of the simple branching process. *Adv. Appl. Prob.* **20** 489–524.

[11] Jagers, P. (1974)
Convergence of general branching processes and functionals thereof. *J. Appl. Prob.* **11** 471–478.

[12] Jagers, P., Nerman, O. (1984)
The growth and composition of branching populations. *Adv. Appl. Prob.* **16** 221–259.

[13] Jagers, P., Nerman, O. (1984)
Limit theorems for sums determined by branching processes and other exponentially growing processes. *Stoch. Proc. Appl.* **17** 47–71.

[14] Lambert, A. (2009)
The allelic partition for coalescent point processes. *Markov Proc. Relat. Fields* **15** 359–386.

[15] Lambert, A. (2010)
The contour of splitting trees is a Lévy process. *Ann. Probab.* **38** 348—395.

[16] Nerman, O. (1981)
On the convergence of supercritical general (CMJ) branching processes. *Z. Wahrscheinlichkeitstheorie verw. Gebiete* **57** 365—395.

[17] Popovic, L. (2004)
Asymptotic genealogy of a critical branching process. *Ann. Appl. Prob.* **14** 2120–2148.

[18] Sagitov, S., Serra, M.C. (2009)
Multitype Bienaymé–Galton–Watson processes escaping extinction. *Adv. Appl. Prob.* **41** 225–246.

[19] Taïb, Z. (1992)
Branching processes and neutral evolution. Lecture Notes in Biomathematics Vol. 93. Springer-Verlag, Berlin.

[20] Yule, G. (1924)
A mathematical theory of evolution based on the conclusions of Dr. J. C. Willis, F.R.S. *Phil. Transac. Roy. Soc. London* **213** 21–87.