HAL open science

The allelic partition for coalescent point processes

Amaury Lambert

▶ To cite this version:

Amaury Lambert. The allelic partition for coalescent point processes. 2009. hal-00273808v2

HAL Id: hal-00273808
https://hal.science/hal-00273808v2

Preprint submitted on 9 Feb 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

The allelic partition for coalescent point processes

By Amaury Lambert

February 9, 2009

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
E-mail: amaury.lambert@upmc.fr
URL: http://ecologie.snv.jussieu.fr/amaury/

Abstract

Assume that individuals alive at time t in some population can be ranked in such a way that the coalescence times between consecutive individuals are i.i.d. The ranked sequence of these branches is called a coalescent point process. We have shown in a previous work [14] that splitting trees are important instances of such populations.

Here, individuals are given DNA sequences, and for a sample of n DNA sequences belonging to distinct individuals, we consider the number S_n of polymorphic sites (sites at which at least two sequences differ), and the number A_n of distinct haplotypes (sequences differing at one site at least).

It is standard to assume that mutations arrive at constant rate (on germ lines), and never hit the same site on the DNA sequence. We study the mutation pattern associated with coalescent point processes under this assumption. Here, S_n and A_n grow linearly as n grows, with explicit rate. However, when the branch lengths have infinite expectation, S_n grows more rapidly, e.g. as $n \ln(n)$ for critical birth–death processes.

Then, we study the frequency spectrum of the sample, that is, the numbers of polymorphic sites/haplotypes carried by k individuals in the sample. These numbers are shown to grow also linearly with sample size, and we provide simple explicit formulae for mutation frequencies and haplotype frequencies. For critical birth–death processes, mutation frequencies are given by the harmonic series and haplotype frequencies by Fisher's logarithmic series.

Running head. The allelic partition for coalescent point processes.
MSC Subject Classification (2000). Primary 92D10; secondary 60-06, 60G10, 60G51, 60G55, 60G70, 60J10, 60J80, 60J85.
Key words and phrases. coalescent point process – splitting tree – Crump–Mode–Jagers process – linear birth–death process – Yule process – allelic partition – infinite site model – infinite allele model – Poisson point process – Lévy process – scale function – law of large numbers – Kingman coalescent – Fisher logarithmic series.

1 Introduction

1.1 The coalescent point process

Splitting trees are those random trees where individuals give birth at constant rate b during a lifetime with general distribution $\Lambda(\cdot)/b$, to i.i.d. copies of themselves (see [12]), where Λ is a positive measure on $(0, \infty]$ with total mass b called the *lifespan measure*. In [14], we have shown that if the splitting tree is started from one individual with known birth time, say 0, and known death time, then individuals alive at time t can be ranked in such a way that the coalescence times between consecutive individuals are i.i.d.

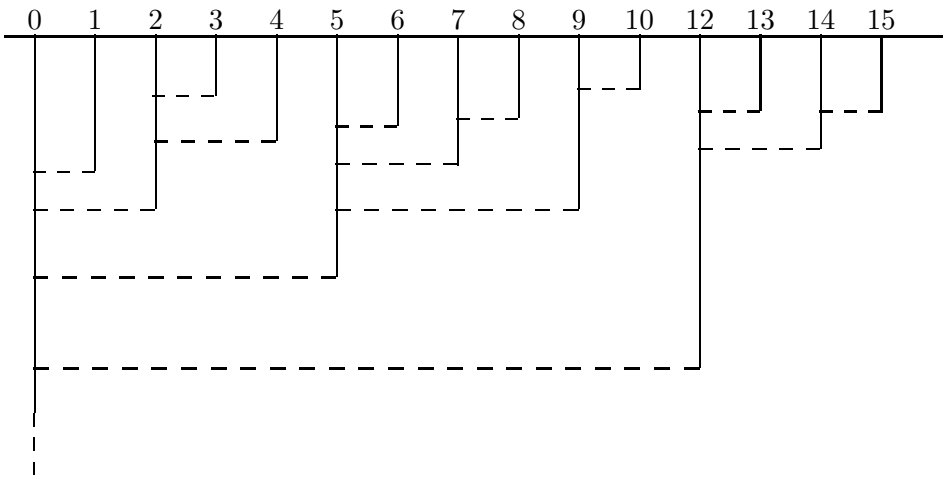

Figure 1: A coalescent point process for $n = 16$ individuals.

Specifically, let N_t be the number of individuals alive at time t. The process $(N_t; t \geq 0)$ is a (homogeneous, binary) Crump–Mode–Jagers process, and is not Markovian unless Λ has an exponential density or is a point mass at ∞. To these N_t individuals, give labels $0, 1, \ldots, N_t - 1$ according to the (unique) order complying with the following rule : 'any individual comes before her own descendants, but after her younger siblings and their descendants'. For any integers i, k such that $0 \leq i < i+k < N_t$, we let $C_{i,i+k}$ be the *coalescence time* (or *divergence time*) between individual i and individual $i+k$, that is, the time elapsed since the lineages of individual i and $i+k$ have diverged. Also define $H_{i+1} := C_{i,i+1}$. Then recall from [14] that for a splitting tree,

$$C_{i,i+k} = \max\{H_{i+1}, \ldots, H_{i+k}\} \tag{1}$$

and conditional on $\{N_t \neq 0\}$, the sequence $(H_i; 1 \leq i \leq N_t - 1)$ has the same law as a sequence of i.i.d. r.v. killed at its first value $\geq t$. As a by-product, we get that the law of N_t conditional on $\{N_t \neq 0\}$ is geometric.

The aforementioned property comes from the fact that the jumping contour process of the splitting tree is a Lévy process $X = (X_s; s \geq 0)$ with Lévy measure Λ and drift coefficient -1. Then the excursions of the contour process between consecutive visits of points at height t are i.i.d. excursions of X. As a consequence, the (H_i) are also i.i.d., and their common distribution is that of $H' := t - \inf_s X_s$, where X is started at t and killed upon hitting $\{0\} \cup (t, +\infty)$. Note that all branch lengths but the last one are distributed as some r.v. H which is H' conditioned

to be smaller than t. The distribution of H' can be expressed in terms of a nonnegative, nondecreasing, differentiable function W, called the *scale function* of X, such that $W(0) = 1$

$$\mathbb{P}(H' > x) = \frac{1}{W(x)} \qquad x \geq 0. \tag{2}$$

The scale function W is characterised by its Laplace transform (see e.g. [6])

$$\int_0^\infty dx\, e^{-\lambda x}\, W(x) = \left(\lambda - \int_0^\infty \Lambda(dx)(1 - e^{-\lambda x})\right)^{-1}. \tag{3}$$

From now on, with no need to refer to the framework of splitting trees, we will consider the genealogy of what we call a *coalescent point process* (originating from [17] where $\Lambda(dx) = b^2 \exp(-bx)dx$) :

1. let $H_1, H_2, \ldots$ be a sequence of independent random variables called *branch lengths* all distributed as some positive r.v. H, and set H_0 to equal $+\infty$.
2. the genealogy of the population $\{0, 1, 2, \ldots\}$ is given by (1).

We will stick to the notation

$$W(x) := \frac{1}{\mathbb{P}(H > x)} \qquad x \geq 0.$$

It will always be implicit that a *sample* of n individuals refers to the *first* n individuals $\{0, 1, \ldots, n-1\}$.

Remark 1 *In the case of splitting trees, conditional on $\{N_t \neq 0\}$, N_t is geometric with success probability $\mathbb{P}(H' > t)$, and conditional on $\{N_t = n\}$, the branch lengths $(H_i; 1 \leq i \leq n-1)$ are i.i.d. with distribution $\mathbb{P}(H' \in \cdot \mid H' < t)$. In what follows, we will repeatedly refer to the genealogy of a splitting tree with n leaves by setting the r.v. H to equal H', without the conditioning (i.e. $t \to \infty$). In the subcritical case, this amounts to considering quasi-stationary populations, which are those populations conditioned to be still alive at time t, as $t \to \infty$ (see e.g. [15]). Another possibility would be, as in [2], to give a prior distribution to the time t of origin, and condition the whole tree on $\{N_t = n\}$. Then as $n \to \infty$, the posterior distribution of t goes to ∞, and we would be left with a (possibly different) distribution of H charging the whole half-line.*

Remark 2 *No distribution of edge lengths can make the coalescent point process coincide with the Kingman coalescent [13]. Indeed, here, the smallest branch length in a sample of n individuals is the minimum of $n-1$ i.i.d. random variables, whereas in the Kingman coalescent, it is the minimum of $n(n-1)/2$ i.i.d. random variables (with exponential distribution).*

Our goal is to characterise the mutation pattern for samples of n individuals, mainly as n gets large. We specify the mutation scheme in the next subsection.

Works studying mutation patterns arising from random genealogies are numerous. Mutation patterns related to populations with fixed size (Wright–Fisher model, Kingman coalescent) are well-known and culminate in *Ewens' sampling formula* (see [9] for a comprehensive account on that subject). More recent works concern mutation patterns related to more general coalescents [4, 16], to branching populations [1, 7], or to both [5].

1.2 Mutation scheme

We adopt two classical assumptions on mutation schemes from population genetics (see e.g. [10])

1. mutations occur at *constant rate* θ on germ lines,
2. mutations are *neutral*, that is, they have no effect on birth rates and lifetimes.

As is usual, we assume that mutations are point substitutions occurring at a single site on the DNA sequence, and that each site can be hit at most once by a mutation. This last assumption is known as the *infinitely-many sites model* (ISM). Instances of DNA sequence are called *alleles* or *haplotypes*, so that under the ISM, each mutation yields a new allele. Without reference to DNA sequences, this last assumption by itself is known as the *infinitely-many alleles model* (IAM).

Specifically, we let $(\mathcal{P}_i; i = 0, 1, 2 \ldots)$ be independent Poisson measures on $(0, \infty)$ with intensity θ (cf. assumption 1). For each i we denote the atoms of $\mathcal{P}_i$ by $\ell_{i1} < \ell_{i2} < \cdots$ and call them *mutations*. Now let $H_1, H_2, \ldots$ be an independent coalescent point process (cf. assumption 2). In agreement with the genealogical structure of a coalescent point process explained in the beginning of this section, we will say that individual $i + k$ *carries* (or *bears*) mutation ℓ_{ij} if $k \geq 0$ and

$$\max\{H_{i+1}, \ldots H_{i+k}\} < \ell_{ij} < H_i,$$

where we agree that $\max \emptyset = 0$ and $H_0 = +\infty$. The second inequality is trivially due to the fact that we throw away all atoms ℓ_{ij} such that $H_i \leq \ell_{ij}$. The set of mutations that an individual bears is her *allele* or her *haplotype*, or merely her type.

For a sample of n individuals, we call S_n the number of *polymorphic sites*, that is, the number of mutations $(\ell_{ij}; 0 \leq i \leq n-1, j \geq 1)$ that are carried by at least one individual and at most $n - 1$. Formally, this yields

$$S_n = \mathrm{Card}\{\ell_{ij} < H_i, 1 \leq i \leq n-1, j \geq 1\} + \mathrm{Card}\{\ell_{0j} < \max\{H_1, \ldots, H_{n-1}\}, j \geq 1\}.$$

Further, we define $S_n(k)$ as the number of mutations carried by k individuals in the sample. In particular,

$$S_n = \sum_{k=1}^{n-1} S_n(k).$$

The sequence $(S_n(1), \ldots, S_n(n-1))$ is called the *site frequency spectrum* of the sample.

Similarly, we call A_n the number of *distinct haplotypes* in a sample of n individuals, that is, the number of alleles that are carried by at least one individual, and $A_n(k)$ as the number of alleles carried by k individuals. In particular, we have

$$A_n = \sum_{k=1}^{n} A_n(k) \text{ and } \sum_{k=1}^{n} kA_n(k) = n.$$

The sequence $(A_n(1), \ldots, A_n(n))$ is called the *allele frequency spectrum* of the sample.

Remark 3 *One always has the inequality $S_n \geq A_n - 1$. Indeed, apart from the ancestral haplotype, each new haplotype independent of at least one new mutation.*

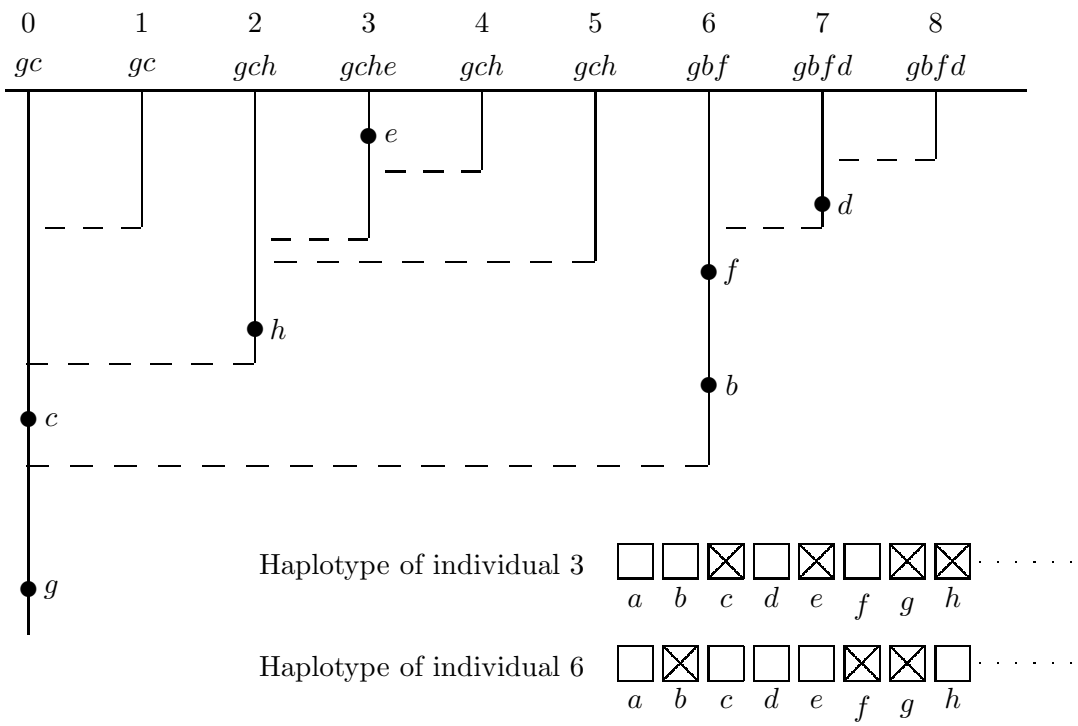

Figure 2: A coalescent point process with mutations for a sample of $n = 9$ individuals. Site a is *not* polymorphic because *no* individual in the sample carries a mutation at that site; site g is *not* polymorphic because *all* individuals in the sample carry the mutation at that site. The number of polymorphic sites is $S_n = 6$. The number of distinct haplotypes is $A_n = 5$.

1.3 Examples of coalescent point processes

Before going into the main part of this work, we provide a few simple examples of coalescent point processes derived from splitting trees, in part for application purposes.

Yule tree. When Λ is a point mass at ∞, the splitting tree is a Yule tree, and $(N_t; t \geq 0)$ is a pure-birth binary process with birth rate, say a. Then $W(x) = e^{ax}$, and H has an exponential distribution with parameter a (see [17]).

Birth–death process. When Λ has an exponential density, $(N_t; t \geq 0)$ is a Markovian birth–death process with (birth rate b and) death rate, say d. Then it is known (see [14] for example) that if $b \neq d$, then

$$W(x) = \frac{d - be^{(b-d)x}}{d - b} \qquad x \geq 0,$$

whereas if $b = d =: a$,

$$W(x) = 1 + ax \qquad x \geq 0.$$

Notice that in the subcritical case ($b < d$), H can take the value ∞ with probability $1 - (b/d)$, which is due to the constrained size of quasi-stationary populations (see Remark 1). Elementary

calculations show that H conditioned to be finite has the same law as the branch length of a *supercritical* birth–death process with birth rate d and death rate b.

Consistency and sampling. The genealogy associated with a coalescent point process is *consistent* in the sense that the genealogy of a sample of n individuals has the same law as that of a sample of $n+1$ individuals from which the *last* individual has been withdrawn (in the splitting tree framework, the last individual is the individual who has no descendants in the sample, and whose ancestors have no elder sibling with descendants in the sample). This property would not hold any longer if the withdrawn individual was chosen at random.

On the other hand, if all individuals in the population are censused *independently* with probability c, then the genealogy of the census is still that of a coalescent point process. Indeed, the typical branch length is H'', where

$$H'' \stackrel{\mathcal{L}}{=} \max\{H_1, \ldots, H_K\},$$

and K is an independent (modified) geometric r.v., that is, $\mathbb{P}(K = j) = c(1-c)^{j-1}$. As a consequence,

$$\frac{1}{W_c(x)} := \mathbb{P}(H'' > x) = 1 - \sum_{j \geq 1} c(1-c)^{j-1} \mathbb{P}(H \leq x)^j \qquad x \geq 0.$$

This last equation also reads

$$W_c = 1 - c + cW.$$

Applying this Bernoulli sampling procedure with intensity c to the previous examples yields the following elementary results.

- the census of a Yule population has the genealogy of a birth–death process population, with birth rate ac and death rate $a(1-c)$
- the census of a birth–death process population has the genealogy of another birth–death process population with birth rate bc and death rate $d - b(1-c)$. In particular, censusing a critical birth–death process population with rate $b = d =: a$ amounts to replacing a with ac.

Infinite lifespan measure. Actually, everything that was stated about splitting trees still holds if the lifespan measure is infinite, provided the lifespans of children remain summable, that is $\int_0^\infty (1 \wedge r)\Lambda(dr) < \infty$. In particular, one still has $W(0) = 1$, and the number of individuals alive at a fixed time remains a.s. finite.

On the contrary, it is a completely different task to define the real tree whose jumping contour process is a Lévy process with no negative jumps but *infinite variation* (see [6]). However, in our setting, this only requires replacing the coalescent point process $H_1, H_2, \ldots$ with a true Poisson point process with intensity measure $ds\,\nu(dx)$, where ν is a σ-finite positive measure defined as the push forward of the excursion measure of X away from $\{t\}$ by the function which maps an excursion ϵ into $t - \inf_s \epsilon_s$. Similarly as in the finite variation case,

$$W(x) := \frac{1}{\nu((x, \infty))} \qquad x \geq 0.$$

In the Brownian case, for example $\nu(dx) = x^{-2}dx$ (again, see [17]), that is, $W(x) = x$.

Here, the analogue of Bernoulli sampling with intensity c consists in taking the maximum H'' of the point process on an interval with exponential length of parameter c (instead of a geometric length). Now c can take any positive value. Standard calculations then yield

$$\frac{1}{W_c(x)} := \mathbb{P}(H'' > x) = 1 - \frac{c}{c + \nu((x, \infty))} \qquad x \geq 0,$$

so that

$$W_c = 1 + cW.$$

As far as splitting trees with infinite variation are concerned, we will only focus on the stable case, where $W(x) = x^{\alpha-1}$ for some $\alpha \in (1, 2]$, the Brownian case corresponding to $\alpha = 2$. In particular, we see that the Brownian coalescent point process censused with intensity c has the same law as the coalescent point process associated with a critical birth–death process with rate c.

1.4 Statements, outline, examples

Our results regarding polymorphic sites are stated in Section 2.

In the first two subsections of Section 2, we assume that $\mathbb{E}(H)$ is finite. Theorem 2.1 provides a law of large numbers and a central limit theorem (if H has a second moment) on the number S_n of polymorphic sites. In particular,

$$\lim_{n\to\infty} \frac{S_n}{n} = \theta\, \mathbb{E}(H) \qquad \text{a.s.} \tag{4}$$

We also give exact explicit formulae for the *expectation* of the number $S_n(k)$ of mutations carried by k individuals in a sample of n.

In the third subsection, we make the less stringent assumption that $\mathbb{E}(\min(H_1, H_2))$ is finite. Theorem 2.3 then gives the asymptotic behaviour of the site frequency spectrum of large samples via the following a.s. convergence

$$\lim_{n\to\infty} \frac{S_n(k)}{n} = \theta \int_0^\infty \frac{dx}{W(x)^2} \left(1 - \frac{1}{W(x)}\right)^{k-1}. \tag{5}$$

In the fourth subsection, we treat the case of stable laws with parameter α, that is, W is given by $W(x) = 1 + cx^{\alpha-1}$, where $\alpha \in (1, 2]$ and c is some positive parameter that can be interpreted as a sampling intensity. Since here $\mathbb{E}(H) = \infty$, the only result holding in the stable case is (5), and only for $\alpha > 3/2$. Theorems 2.4 and 2.5 give the asymptotic behaviour of S_n. When $\alpha = 2$, $S_n/n\ln(n)$ converges in probability (to θ/c), and when $\alpha \neq 2$, S_n/n^β converges in distribution, with $\beta = 1/(\alpha - 1)$.

Section 3 displays our results regarding distinct haplotypes. The trick is to characterise the law of the branch length H^θ of the next individual bearing no mutation other than those carried by, say, individual 0. Proposition 3.1 does this as follows

$$\frac{1}{\mathbb{P}(H_\theta > x)} =: W_\theta(x) = 1 + \int_0^x W'(u)e^{-\theta u}\, du \qquad x \geq 0.$$

Theorem 3.2 states a.s. convergences without moment existence assumptions. Specifically,

$$\lim_{n\to\infty} \frac{A_n}{n} = \mathbb{E}\left(1 - e^{-\theta H^\theta}\right) \qquad \text{a.s.,} \tag{6}$$

and the allele frequency spectrum for large samples is given by the following a.s. convergence

$$\lim_{n\to\infty} \frac{A_n(k)}{n} = \int_0^\infty dx\, \theta\, e^{-\theta x} \frac{1}{W_\theta(x)^2} \left(1 - \frac{1}{W_\theta(x)}\right)^{k-1} . \tag{7}$$

Before ending this last subsection, we want to point out that in some cases, more explicit formulae can be computed. First, for the *Yule process* with birth rate 1, (or with parameter a, but after replacing θ with $a\theta$), that is, when $W(x) = e^x$, one gets easily

$$\lim_{n\to\infty} \frac{S_n}{n} = \theta \quad \text{and} \quad \lim_{n\to\infty} \frac{S_n(k)}{n} = \frac{\theta}{k(k+1)} .$$

Computations are not as straightforward for the number of haplotypes. Second, for the *critical birth–death process* with birth rate 1 (or with parameter a, but after replacing θ with $a\theta$), that is, when $W(x) = 1 + x$, one gets

$$\lim_{n\to\infty} \frac{S_n}{n\ln(n)} = \theta \quad \text{and} \quad \lim_{n\to\infty} \frac{S_n(k)}{n} = \frac{\theta}{k} .$$

In addition,

$$\lim_{n\to\infty} \frac{A_n}{n} = \theta \ln\left(1 + \theta^{-1}\right) \quad \text{and} \quad \lim_{n\to\infty} \frac{A_n(k)}{n} = \frac{\theta}{k}\,(1+\theta)^{-k} .$$

Remark 4 *It is amusing to notice that the rescaled number $A_n(k)$ of haplotypes with k representatives is also the probability that a species has k representatives in Fisher's* log-series of species abundance *[11]. In Fisher's model, a given species has an unknown density which is assumed to be drawn from a Gamma distribution with parameter a. As a result of Bernoulli sampling in a large population, it is then assumed that given the value d of this density, the number X of individuals spotted from this species is Poisson with parameter ρd, where ρ is the sampling intensity. It can then be shown that as $a \downarrow 0$, conditional on $\{X \geq 1\}$ (since at least one individual must be spotted for the species to be recorded), $\mathbb{P}(X = k)$ goes to $C(1+1/\rho)^{-k}/k$, for some normalising constant C.*

Remark 5 *In a coalescent point process, divergence times are on average deeper than in the Kingman coalescent (our trees are more 'star-like'). This forbids convergence of our statistics without rescaling (by the sample size n or by $n\ln(n)$). In particular, notice that the asymptotic proportion of individuals in a cluster of size greater than K, i.e. $\lim_n n^{-1}\sum_{k\geq K} A_n(k)$, vanishes as K grows to ∞. This shows that the largest cluster in a sample of n has neglectable size w.r.t. n, which contrasts with the Kingman coalescent, where the allele frequency spectrum is given by Ewens' sampling formula (see [9, 10]). As $n \to \infty$, the numbers of haplotypes $A_n(k)$ carried by k individuals [3] converge to independent Poisson r.v. with parameter θ/k, and the i-th eldest haplotype [8] is carried by approximately $P_i n$ individuals, where $(P_i; i \geq 1)$ is a Poisson–Dirichlet r.v.*

2 Number of polymorphic sites

Results for polymorphic sites depend on integrability assumptions on H. Of course these are always fulfilled if the time t when the population was founded is known, since then $H \leq t$ a.s. We will see that the critical assumptions are either $\mathbb{E}(\min(H_1, H_2)) < \infty$, or the more stringent $\mathbb{E}(H) < \infty$. Notice that the first assumption is equivalent to the integrability of $1/W^2$, and the second one to the integrability of $1/W$.

2.1 Law of large numbers and central limit theorem

Recall that S_n is the number of polymorphic sites in the sample of n individuals.

Theorem 2.1 *If $\mathbb{E}(H) < \infty$, then*

$$\lim_{n\to\infty} n^{-1} S_n = \theta\, \mathbb{E}(H) \qquad \textit{a.s. and in } L^1.$$

If in addition $\mathbb{E}(H^2) < \infty$, then

$$\sqrt{n}\left(n^{-1} S_n - \theta\, \mathbb{E}(H)\right)$$

converges in distribution to a centered normal variable with variance $\theta\, \mathbb{E}(H) + \theta^2\, Var(H)$.

Proof. Set $Y_n := \max\{H_1, \ldots, H_{n-1}\}$. Recall from the Introduction that

$$S_n = \sum_{i=1}^{n-1} Q_i + R_n,$$

where Q_i is the number of points of the Poisson point process $\mathcal{P}_i$ in $(0, H_i)$, and R_n is the number of points of the Poisson point process $\mathcal{P}_0$ in $(0, Y_n)$. By the strong law of large numbers, we know that

$$\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n-1} Q_i = \theta\, \mathbb{E}(H) \qquad \text{a.s. and in } L^1,$$

so we need to prove that

$$\lim_{n\to\infty} n^{-1} R_n = 0 \qquad \text{a.s. and in } L^1.$$

Now because R_n/Y_n converges to θ a.s. and in L^1, it is sufficient to prove that

$$\lim_{n\to\infty} n^{-1} Y_n = 0 \qquad \text{a.s. and in } L^1.$$

Because $Y_n < \sum_{i=1}^{n-1} H_i$,

$$\limsup_{n\to\infty} n^{-1} Y_n =: Y < \infty \text{ a.s.}$$

By the 0-1 law, Y is not random. To prove that $Y = 0$, we let $Y_n^{(1)}$ (resp. $Y_n^{(2)}$) be the maximum of the H_i's indexed by odd (resp. even) numbers. Then it is clear that $Y_n = \max(Y_n^{(1)}, Y_n^{(2)})$, and that $n^{-1} Y_n^{(1)}$ as well as $n^{-1} Y_n^{(2)}$ both converge to $Y/2$. This shows that $Y = Y/2$, so that $Y = 0$.

For convergence in L^1, pick any $x > 0$, and notice that

$$\begin{aligned} n^{-1}\mathbb{E}(Y_n) &= n^{-1}\mathbb{E}(Y_n, Y_n \leq x) + n^{-1}\mathbb{E}(Y_n, Y_n > x) \\ &\leq n^{-1}x + n^{-1}\mathbb{E}\left(\sum_{i=1}^{n-1} H_i \mathbf{1}_{\{H_i > x\}}\right) \\ &\leq n^{-1}x + \mathbb{E}(H, H > x). \end{aligned}$$

Since $\mathbb{E}(H) < \infty$, this last inequality shows that $n^{-1}\mathbb{E}(Y_n)$ vanishes as $n \to \infty$.

Now we prove the central limit theorem for S_n. It is elementary to compute $\mathrm{Var}(Q_1)$ as $\theta\,\mathbb{E}(H) + \theta^2\mathrm{Var}(H)$, so by the classical central limit theorem applied to the sum of Q_i's, we only have to prove that $R_n/\sqrt{n}$ converges to 0 in probability. For any $\lambda > 0$,

$$\mathbb{E}\left(\exp\left(-\lambda R_n/\sqrt{n}\right)\right) = \mathbb{E}\left(\exp\left(-\theta Y_n\left(1 - e^{-\lambda/\sqrt{n}}\right)\right)\right),$$

which shows it is sufficient to prove that $Y_n/\sqrt{n}$ converges to 0 in probability. As previously, we write

$$\begin{aligned} n^{-1}\mathbb{E}\left(Y_n^2\right) &= n^{-1}\mathbb{E}\left(Y_n^2, Y_n \leq x\right) + n^{-1}\mathbb{E}\left(Y_n^2, Y_n > x\right) \\ &\leq n^{-1}x^2 + n^{-1}\mathbb{E}\left(\sum_{i=1}^{n-1} H_i^2 \mathbf{1}_{\{H_i > x\}}\right) \\ &\leq n^{-1}x^2 + \mathbb{E}\left(H^2, H > x\right). \end{aligned}$$

Thus, convergence of $Y_n/\sqrt{n}$ to 0 holds in L^2, and subsequently, it holds in probability. □

2.2 Explicit formulae for the expected frequency spectrum

Recall that $S_n(k)$ denotes the number of mutant sites that are carried by exactly k individuals in the sample of n individuals (and since we only count polymorphic sites, $S_n(n) = 0$).

Theorem 2.2 *For all $1 \leq k \leq n-1$,*

$$\mathbb{E}(S_n(k)) = \theta \int_0^\infty dx \left(1 - \frac{1}{W(x)}\right)^{k-1} \left(\frac{n-k-1}{W(x)^2} + \frac{2}{W(x)}\right),$$

which is finite if and only if $\mathbb{E}(H) < \infty$. Then in particular,

$$\lim_{n\to\infty} n^{-1}\mathbb{E}(S_n(k)) = \theta \int_0^\infty \frac{dx}{W(x)^2}\left(1 - \frac{1}{W(x)}\right)^{k-1}.$$

Remark 6 *Taking the sum over k in the r.h.s. of the last equality of the theorem, one gets $\theta\,\mathbb{E}(H)$, so that, thanks to the L^1 convergence in Theorem 2.1,*

$$\lim_{n\to\infty} n^{-1}\sum_{k=1}^{n-1}\mathbb{E}(S_n(k)) = \theta\,\mathbb{E}(H) = \sum_{k\geq 1}\lim_{n\to\infty} n^{-1}\mathbb{E}(S_n(k)).$$

Before giving a proof of the previous theorem, we want to make a point that will also be useful in the next subsection. For any tree with point mutations, a mutation is carried by k individuals if and only of it is in the part of the tree subtending k leaves. Then in any given tree with edge lengths and Poisson point process of mutations (with rate θ) independent of the genealogy (as in our situation), the expectation of the number of mutations carried by k individuals is θL_k, where L_k is the Lebesgue measure of the part of the tree subtending k leaves (i.e., tips). In our setting, we will call $L_k(n)$, for $k \leq n-1$, the Lebesgue measure of the part of the tree subtending k tips among individuals $\{0, 1, \ldots, n-1\}$, so that

$$\mathbb{E}(S_n(k)) = \theta\, \mathbb{E}(L_k(n)) \qquad 1 \leq k \leq n-1.$$

Remark 7 *The last equality along with more specific considerations given in the next subsection provide a less analytic and more transparent proof than the proof we give hereafter. However, we stick to it for the interest of the method itself.*

Proof of Theorem 2.2 We set $N(x)$ to be the smallest $i \geq 1$ such that $H_i > x$. The proof relies on the fact that

$$\mathbb{E}(S_n(k)) = \lim_{x\to\infty} \theta\, \mathbb{E}(L_k(N(x)) \mid N(x) = n) \qquad 1 \leq k \leq n-1.$$

On the event $\{N(x) = n\}$, we will need to extend the definition of $L_k(N(x))$ to $k = n$, as being the Lebesgue measure of the part of the tree *up to time* $-x$ subtending all tips $\{0, 1, \ldots, n-1\}$, that is, $L_n(N(x)) = x - \max_{i=1,\ldots,n-1} H_i$.

For editing reasons, we will prefer to write $F(x) = \mathbb{P}(H > x)$, instead of $1/W(x)$. Since F is a.e. differentiable and our goal is to let $x \to \infty$, we can set $f(x) := -F'(x)$ without loss of generality. We let $\tilde{H}$ denote the branch length $H_{N(x)}$, and we set

$$\tilde{N} := \min\{k \geq 1 : H_{N(x)+k} > x + dx\},$$

as well as $\tilde{L}_k$ the Lebesgue measure of the part of the tree subtending k leaves among individuals $\{N(x), N(x)+1, \ldots, N(x+dx)-1\}$. Note that $(\tilde{N}, \tilde{L}_k, \tilde{H})$ are independent of $(N(x), L_k(N(x)))$; that $\tilde{H}$ is distributed as H conditional on $\{H > x\}$; and that $(\tilde{N}, \tilde{L}_k)$ is independent of $\tilde{H}$ and distributed as $(N(x+dx), L_k(N(x+dx))$. Next observe that if $\tilde{H} > x + dx$, then $N(x+dx) = N(x)$ and $L_k(N(x+dx)) = L_k(N(x))$, except if $k = n$, where by definition $L_n(N(x+dx)) = L_n(N(x)) + dx$. On the other hand, if $\tilde{H} \in dx$, $L_k(N(x+dx))$ is the sum of measures of edges subtending k tips in $\{0, 1, \ldots, N(x)-1\}$ with measures of edges subtending k tips in $\{N(x), \ldots, N(x) + \tilde{N} - 1\}$. This reads

$$\begin{aligned} L_k(N(x+dx))\mathbf{1}_{\{N(x+dx)=n\}} &= \mathbf{1}_{\{\tilde{H}>x+dx\}}\mathbf{1}_{\{N(x)=n\}}\left(L_k(N(x)) + dx\mathbf{1}_{k=n}\right) \\ &\quad + \mathbf{1}_{\{\tilde{H}\leq x+dx\}} \sum_{j=1}^{n-1} \mathbf{1}_{\{N(x)=j\}}\mathbf{1}_{\{\tilde{N}=n-j\}}\left(L_k(N(x)) + \tilde{L}_k(\tilde{N})\right), \end{aligned}$$

where we have used the extension of the definition of L_k specified earlier (cases when $k = j$ or $k = n - j$ in the sum). Now set

$$U_{k,n}(x) := \mathbb{E}(L_k(N(x)), N(x) = n).$$

By the independences stated previously, taking expectations, we get

$$U'_{k,n}(x+) = -U_{k,n}(x)\frac{f}{F}(x) + \mathbf{1}_{k=n}\mathbb{P}(N(x)=k) + 2\sum_{j=1}^{n-1} U_{k,j}(x)\mathbb{P}(N(x)=n-j)\frac{f}{F}(x).$$

Setting

$$V_k(x;s) := \sum_{n\geq k} U_{k,n}(x)s^n \qquad s\in[0,1),$$

and observing that $|U_{k,n}(x)| \leq nx$, and (so) that $|U'_{k,n}(x)| \leq c(x)n^2$ for some positive $c(x)$ independent of k and n, we get

$$\frac{\partial V_k}{\partial x}(x;s) = -\frac{f}{F}(x)V_k(x;s) + \mathbb{P}(N(x)=k)s^k + 2\frac{f}{F}(x)\sum_{n\geq k} s^n \sum_{j=1}^{n-1} U_{k,j}(x)\mathbb{P}(N(x)=n-j).$$

Since $U_{k,j}(x)=0$ when $j\leq k-1$, the last term equals

$$\begin{aligned} 2\frac{f}{F}(x)\sum_{n\geq k+1} s^n \sum_{j=k}^{n-1} U_{k,j}(x)\mathbb{P}(N(x)=n-j) &= 2\frac{f}{F}(x)\sum_{j\geq k} U_{k,j}(x)s^j \sum_{n\geq j+1} s^{n-j}\mathbb{P}(N(x)=n-j) \\ &= 2\frac{f}{F}(x)V_k(x;s)\sum_{n\geq 1} s^n\mathbb{P}(N(x)=n). \end{aligned}$$

As a consequence, we get the following differential equation

$$\frac{\partial V_k}{\partial x}(x;s) = G_k(x;s)V_k(x;s) + \mathbb{P}(N(x)=k)s^k,$$

where we have put

$$G_k(x;s) := \left(2\mathbb{E}\left(s^{N(x)}\right) - 1\right)\frac{f}{F}(x).$$

Now since $\mathbb{P}(N(x)=k) = F(x)(1-F(x))^{k-1}$, we easily get

$$\int_0^x G_k(y;s)\,dy = \ln\left[\frac{F(x)}{(1-s+sF(x))^2}\right].$$

This allows us to integrate the differential equation in $V_k(.;s)$ to finally arrive at

$$V_k(x;s) = \frac{s^kF(x)}{(1-s+sF(x))^2}\int_0^x (1-F(y))^{k-1}(1-s+sF(y))^2\,dy.$$

With the shortcuts $u := 1-F(x)$ and $v := 1-F(y)$, and using the series expansion of $(1-us)^{-2}$, we get

$$V_k(x;s) = s^k(1-u)\int_0^x v^{k-1}(1-vs)^2\sum_{j\geq 1} ju^{j-1}s^{j-1}\,dy.$$

It is elementary algebra to compute the following equality

$$(1-vs)^2\sum_{j\geq 1} ju^{j-1}s^{j-1} = 1 + \sum_{j\geq 1} s^ju^{j-2}\left(j(u-v)^2 + u^2 - v^2\right),$$

which yields

$$V_k(x;s) = s^k(1-u)\int_0^x v^{k-1}\,dy + (1-u)\sum_{j\geq 1}\int_0^x v^{k-1}s^{k+j}u^{j-2}\left(j(u-v)^2+u^2-v^2\right)\,dy.$$

Identifying this entire series with the definition of V_k, we get for all $1 \leq k \leq n-1$,

$$U_{k,n}(x) = F(x)(1-F(x))^{n-k-2}\int_0^x (1-F(y))^{k-1}\times \\ \times\left((n-k)(F(y)-F(x))^2+(1-F(x))^2-(1-F(y))^2\right)\,dy.$$

As a consequence,

$$\mathbb{E}(L_k(N(x)) \mid N(x)=n) = (1-F(x))^{-k-1}\int_0^x (1-F(y))^{k-1}\times \\ \times\left((n-k)(F(y)-F(x))^2+(1-F(x))^2-(1-F(y))^2\right)\,dy.$$

which, by Beppo Levi's theorem, converges, as $x \to \infty$, to

$$\theta^{-1}\mathbb{E}(S_n(k)) = \int_0^\infty (1-F(y))^{k-1}\left((n-k)F(y)^2+1-(1-F(y))^2\right)\,dy,$$

and this finishes the proof. □

2.3 Site frequency spectrum of large samples

Here, we assume that $\mathbb{E}(\min(H_1,H_2)) < \infty$, that is, $1/W^2$ is integrable.

Theorem 2.3 *For all $1 \leq k \leq n-1$, the following convergence holds a.s. (and in L^1 as well if $\mathbb{E}(H) < \infty$)*

$$\begin{aligned}\lim_{n\to\infty} n^{-1}S_n(k) &= \theta\,\mathbb{E}\left((\min\{H_1,H_{k+1}\}-\max\{H_2,\ldots,H_k\})^+\right)\\ &= \theta\int_0^\infty \frac{dx}{W(x)^2}\left(1-\frac{1}{W(x)}\right)^{k-1}.\end{aligned}$$

Proof. Reasoning similarly as in the previous subsection, we see that a point mutation occurring on branch i is carried by k individuals if and only if it is carried by individuals $i, i+1, \ldots, i+k-1$, and by no one else. This happens if and only if this mutation, corresponding to the atom ℓ_{ij}, say, of $\mathcal{P}_i$, has

$$\max\{H_{i+1},\ldots,H_{i+k-1}\} < \ell_{ij} < H_i,$$

for the mutation to be carried by individuals $i, i+1, \ldots, i+k-1$, along with

$$\ell_{ij} < H_{i+k},$$

for the mutation not to be carried by others. More formally, we set $\mathcal{F}$ the space of point processes on $(0,\infty)$, and F_k the set of $(k+1)$-dimensional arrays with values in $\mathcal{F}\times(0,\infty)$. Next, for any $\Xi \in F_k$, written as $\Xi = ((p_0,x_0),\ldots,(p_k,x_k))$ we define

$$G(\Xi) := \mathrm{Card}\left(p_0 \cap (\max\{x_1,\ldots,x_{k-1}\},\min\{x_0,x_k\})\right),$$

where it is understood that the interval (a, b) is empty if $a \geq b$. Then the number of mutations carried by k individuals among the first n can be written as

$$S_n(k) = \sum_{i=0}^{n-k} G(\Xi_i),$$

where

$$\Xi_i := ((\mathcal{P}_i, H_i), \ldots, (\mathcal{P}_{i+k}, H_{i+k}))$$

and, for the last term of the sum to be correctly written, H_n is set to $+\infty$ (as H_0). Next, observe that

$$\mathbb{E}(G(\Xi_1)) = \theta\, \mathbb{E}\left((\min\{H_1, H_{k+1}\} - \max\{H_2, \ldots, H_k\})^+\right),$$

so that $G(\Xi_1)$ is integrable (assumption stated before the theorem). Now for any $0 \leq r \leq k$, the random values $G(\Xi_i)$, for i such that $i = r\ [k+1]$ (standing for mod $(k+1)$), are i.i.d. and integrable, so by the strong law of large numbers, we have the following a.s. convergence

$$\lim_{n\to\infty} n^{-1} \sum_{0 \leq i = r[k+1] \leq n-k} G(\Xi_i) = \frac{1}{k+1}\, \mathbb{E}(G(\Xi_1)).$$

Actually, the convergence would also hold in L^1 if we had discarded mutations carried by individual 0 and individual $n-k$, which involve terms that are not integrable if $\mathbb{E}(H) = \infty$. If $\mathbb{E}(H) < \infty$, then convergence holds in L^1. Summing over r these $k+1$ equalities, we get the convergence of $n^{-1}S_n(k)$ to $\mathbb{E}(G(\Xi_1))$, and

$$\begin{aligned}
\mathbb{E}(G(\Xi_1)) &= \theta\, \mathbb{E}\left((\min\{H_1, H_{k+1}\} - \max\{H_2, \ldots, H_k\})^+\right) \\
&= \theta\, \mathbb{E} \int_0^\infty dx\, \mathbf{1}_{x < \min\{H_1, H_{k+1}\}}\, \mathbf{1}_{x > \max\{H_2, \ldots, H_k\}} \\
&= \theta \int_0^\infty dx\, \mathbb{P}(H > x)^2\, \mathbb{P}(H < x)^{k-1},
\end{aligned}$$

which ends the proof. □

2.4 Stable laws

Here, we tackle the case when H is in the domain of attraction of a stable law, which happens in particular for a splitting tree whose contour process is a stable Lévy process with no negative jumps with index $\alpha \in (1, 2]$. If such a population is censused with intensity $c > 0$ then the corresponding function W (see Introduction) is

$$W(x) = 1 + cx^{\alpha-1} \qquad x \geq 0.$$

From now on, we will assume that W has the form given in the foregoing display. Recall that $1/W(x)$ is the probability that a branch has length greater than x. Observe that here H is not integrable, so that Theorems 2.1 and 2.2 do not apply. However, asymptotic results for the site frequency spectrum of large samples given in Theorem 2.3 apply for $\alpha > 3/2$.

2.4.1 Brownian case

Here, we assume that $\alpha = 2$, which corresponds both to a (censused) Brownian population and to the (censused or not) population of a critical birth–death process.

Theorem 2.4 *When $W(x) = 1 + cx$, we have the following convergence in probability*

$$\lim_{n\to\infty} \frac{S_n}{n\ln(n)} = \theta/c.$$

Proof. Recall that S_n is to be written as

$$S_n = \sum_{i=1}^{n-1} Q_i + R_n,$$

where Q_i is the number of points of the Poisson point process $\mathcal{P}_i$ in $(0, H_i)$, and R_n is the number of points of the Poisson point process $\mathcal{P}_0$ in $(0, Y_n)$, where $Y_n = \max\{H_1, \ldots, H_{n-1}\}$. Now observe that

$$\mathbb{P}(Y_n > \varepsilon n\ln(n)) = 1 - \left(1 - \frac{1}{1 + c\varepsilon n\ln(n)}\right)^{n-1},$$

which vanishes as $n \to \infty$, so that $Y_n/n\ln(n)$ converges to 0 in probability. This implies in turn that $R_n/n\ln(n)$ also converges to 0 in probability. As a consequence, we can focus on the sum of Q_i's. Pick any $\lambda > 0$ and check that

$$\mathbb{E}\left(\exp -\frac{\lambda}{n\ln(n)}\sum_{i=1}^{n-1} Q_i\right) = \left(\mathbb{E}\left(\exp -\theta H\left(1 - e^{-\lambda/n\ln(n)}\right)\right)\right)^{n-1},$$

We are bound to study the behaviour of $\mathbb{E}(\exp -yH)$ as $y \to 0$.

$$\begin{aligned}
\mathbb{E}(\exp -yH) &= 1 - y\int_0^\infty \frac{e^{-yx}}{W(x)}\,dx \\
&= 1 - y\int_0^\infty \frac{e^{-u}}{y + cu}\,du \\
&= 1 - y\int_1^\infty \frac{e^{-u}}{y + cu}\,du + y\int_0^1 \frac{1 - e^{-u}}{y + cu}\,du - yc^{-1}\ln((y + c)/y) \\
&= 1 + c^{-1}y\ln(y) + O(y),
\end{aligned}$$

where $O(y)/y$ is bounded near 0. Setting $u_n := \theta\left(1 - e^{-\lambda/n\ln(n)}\right)$, there is a vanishing sequence v_n such that

$$\begin{aligned}
\mathbb{E}\left(\exp -\lambda\frac{S_n}{n\ln(n)}\right) &= \left(1 + c^{-1}u_n\ln(u_n) + O(u_n)\right)^n (1 + v_n) \\
&= \exp\left(c^{-1}nu_n\ln(u_n) + O(nu_n)\right)(1 + v_n),
\end{aligned}$$

which converges to $\exp(-\lambda\theta/c)$. □

2.4.2 Stable case $\alpha \neq 2$

Here, we assume that $W(x) = 1 + cx^{\alpha-1}$, for some $\alpha \in (1,2)$.

Theorem 2.5 *When $W(x) = 1 + cx^{\alpha-1}$, we have the following convergence in distribution*

$$\lim_{n\to\infty} \frac{S_n}{n^{1/(\alpha-1)}} = Z_{\varphi(\mathbf{e})},$$

where $(Z_t; t \geq 0)$ is the stable subordinator with Laplace exponent $\lambda \mapsto c^{-1}\theta^{\alpha-1}\lambda^{\alpha-1}$, $\mathbf{e}$ is an independent exponential r.v. with parameter 1, and φ is defined by

$$\varphi(x) = x^{1-\alpha}\, e^{-x} + \int_0^x ds\, s^{1-\alpha}\, e^{-s} \qquad x > 0.$$

Remark 8 *Observe that φ decreases on $(0,\infty)$ from $+\infty$ to a positive limit, equal to $\Gamma(2-\alpha)$. Also, recall that $S_n = \sum_{i=1}^{n-1} Q_i + R_n$, where R_n is the extra contribution from the maximum branch length. Then it is possible to see by the same kind of proof as that of the theorem, that $\sum_{i=1}^{n-1} Q_i$ converges in distribution to $Z_{\Gamma(2-\alpha)}$. This indicates that, opposite to the Brownian case, the (double) contribution of the maximum branch length is not negligible here.*

Proof. Let us compute the limiting distribution of $n^{-1/(\alpha-1)}(Y_n + \sum_{i=1}^{n-1} H_i)$, where $Y_n = \max\{H_1, \ldots, H_{n-1}\}$. Set $\beta := 1/(\alpha-1)$, as well as

$$I_n(\lambda) := \mathbb{E}\left(\exp -\lambda\, n^{-\beta}\left(Y_n + \sum_{i=1}^{n-1} H_i\right)\right).$$

Then

$$I_n(\lambda) = \int_0^\infty \mathbb{P}(Y_n \in dz) e^{-2\lambda n^{-\beta} z}\left(\mathbb{E}\left(e^{-\lambda n^{-\beta} H'_z}\right)\right)^{n-2},$$

where H'_z has the law of H conditioned on being smaller than z. Next, we have

$$\mathbb{P}(Y_n \in dz) = \left(\frac{cz^{\alpha-1}}{1+cz^{\alpha-1}}\right)^{n-2} \frac{c(n-1)(\alpha-1)z^{\alpha-2}}{(1+cz^{\alpha-1})^2}\, dz \qquad z > 0$$

and

$$\mathbb{P}(H'_z \in dx) = \frac{c(\alpha-1)x^{\alpha-2}}{(1+cx^{\alpha-1})^2}\, \frac{1+cz^{\alpha-1}}{cz^{\alpha-1}}\, dx \qquad 0 < x < z,$$

so we get

$$I_n(\lambda) = \int_0^\infty dz\, \frac{c(n-1)(\alpha-1)z^{\alpha-2}}{(1+cz^{\alpha-1})^2} e^{-2\lambda n^{-\beta} z}\left(\int_0^z dx\, \frac{c(\alpha-1)x^{\alpha-2}}{(1+cx^{\alpha-1})^2}\, e^{-\lambda n^{-\beta} x}\right)^{n-2}$$

Changing variables, this also reads

$$I_n(\lambda) = c^{-1}(1-n^{-1})(\alpha-1)\lambda^{\alpha-1}\int_0^\infty dv\, \frac{v^{-\alpha}\, e^{-2v}}{(1+n^{-1}c^{-1}\lambda^{\alpha-1}v^{1-\alpha})^2} J_n(v;\lambda)^{n-2}$$

where

$$\begin{aligned}
J_n(v;\lambda) &= (\alpha-1)cn\lambda^{1-\alpha}\int_0^v du\,\frac{u^{\alpha-2}\,e^{-u}}{(1+cn\lambda^{1-\alpha}u^{\alpha-1})^2}\\
&= \left[\frac{-e^{-u}}{1+cn\lambda^{1-\alpha}u^{\alpha-1}}\right]_0^v - \int_0^v du\,\frac{e^{-u}}{1+cn\lambda^{1-\alpha}u^{\alpha-1}}\\
&= 1-\frac{e^{-v}}{1+cn\lambda^{1-\alpha}v^{\alpha-1}} - \int_0^v du\,\frac{e^{-u}}{1+cn\lambda^{1-\alpha}u^{\alpha-1}}\\
&= 1-n^{-1}K_n(v;\lambda),
\end{aligned}$$

where $K_n(v;\lambda)$ is positive and converges to $c^{-1}\lambda^{\alpha-1}\varphi(v)$ as $n\to\infty$. By the Lebesgue convergence theorem, we get the convergence of $I_n(\lambda)$ to

$$c^{-1}(\alpha-1)\lambda^{\alpha-1}\int_0^\infty dv\,v^{-\alpha}\,e^{-2v}\exp(-c^{-1}\lambda^{\alpha-1}\varphi(v)).$$

Integrating by parts with $\varphi'(v)=(1-\alpha)v^{-\alpha}e^{-v}$, we finally get

$$\lim_{n\to\infty} I_n(\lambda) = \int_0^\infty dv\,e^{-v}\exp(-c^{-1}\lambda^{\alpha-1}\varphi(v)).$$

The last step is the same as in the foregoing proof, that is

$$\begin{aligned}
\lim_{n\to\infty}\mathbb{E}\left(\exp-\lambda\,n^{-1/(\alpha-1)}S_n\right) &= \lim_{n\to\infty}\mathbb{E}\left(\exp-\theta\,\left(1-e^{-\lambda n^{-1/(\alpha-1)}}\right)\left(Y_n+\sum_{i=1}^{n-1}H_i\right)\right)\\
&= \lim_{n\to\infty} I_n(\theta\lambda)\\
&= \int_0^\infty dv\,e^{-v}\exp(-c^{-1}\theta^{\alpha-1}\lambda^{\alpha-1}\varphi(v)),
\end{aligned}$$

which is the desired result. □

3 Number of distinct haplotypes

3.1 The next branch with no extra mutation

We let $\mathcal{E}^\theta$ denote the set of individuals who *carry no more mutations than* individual 0 (some of and at most exactly the mutations carried by 0, but no other mutation). Set $K_0^\theta := 0$ and for $i\geq 1$, define K_i^θ as the i-th individual in $\mathcal{E}^\theta$, and $H_i^\theta := H_{K_i^\theta}$ the associated branch length. We write H^θ in lieu of H_1^θ and we define the function W_θ by

$$\mathbb{P}(H^\theta > x) = \frac{1}{W_\theta(x)} \qquad x\geq 0.$$

Proposition 3.1 *The bivariate sequence* $((K_i^\theta - K_{i-1}^\theta, H_i^\theta); i\geq 1)$ *is a sequence of i.i.d. random pairs. The function* W_θ *is given by*

$$W_\theta(x) = 1+\int_0^x W'(u)e^{-\theta u}\,du \qquad x\geq 0.$$

Remark 9 *In the case when the coalescent process is derived from a splitting tree with lifespan measure Λ, the calculation of W_θ is straightforward. Indeed, it can be seen in that case that the point process $(H_i^\theta; i \geq 1)$ is the coalescent point process of the splitting tree obtained from the initial splitting tree with mutations after throwing away all points above a mutation. But this new tree is again a splitting tree, since lifespans are i.i.d. and terminate either at death time or at the first point mutation, so the lifespan measure is now $\Lambda_\theta(dx) = e^{-\theta x}\,\Lambda(dx) + \theta e^{-\theta x}\Lambda((x,\infty))\,dx$. As a consequence, W_θ is here the scale function characterised as in (3) by its Laplace transform*

$$\begin{aligned}\int_0^\infty dx\, e^{-\lambda x}\, W_\theta(x) &= \left(\lambda - \int_0^\infty \Lambda_\theta(dx)(1-e^{-\lambda x})\right)^{-1} \\ &= \frac{\lambda+\theta}{\lambda}\left(\lambda+\theta - \int_0^\infty \Lambda(dx)(1-e^{-(\lambda+\theta)x})\right)^{-1} \\ &= \frac{\lambda+\theta}{\lambda}\int_0^\infty dx\, e^{-(\lambda+\theta)x}\, W(x),\end{aligned}$$

which yields the equality given in the statement.

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 individual with no mutation other than those carried by individual 0 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.

Now the event $\{H^\theta \in dx\}$ can be decomposed according to: the value of H_1; conditional on $H_1 = z$, the value of the age V_z of the oldest mutation on H_1; conditional on $V_z = y$, the value H'_y of the branch length associated with the first individual in $\mathcal{E}_1^\theta$ with branch length greater than y. Indeed, $H^\theta \in dx$ if: $H_1 \in dx$ and there is no mutation in H_1 (then $K_0^\theta = 1$); or $H_1 \in dx$, the age of the oldest mutation on $H_1 = x$ is $V_x = y < x$ and the next individual with no mutation other than those carried by individual 1 and branch length $H'_y > y$ has $H'_y < x$; or $H_1 = z < x$, the age of the oldest mutation on $H_1 = z$ is $V_z = y < z$ and the next individual with no mutation other than those carried by individual 1 and branch length $H'_y > y$ has $H'_y \in dx$.

$$\begin{aligned}\mathbb{P}(H^\theta \in dx) = \mathbb{P}(H_1 \in dx)e^{-\theta x} + \mathbb{P}(H_1 \in dx)\int_0^x \mathbb{P}(V_x \in dy)\mathbb{P}(H'_y < x) \\ + \int_0^x \mathbb{P}(H_1 \in dz)\int_0^z \mathbb{P}(V_z \in dy)\mathbb{P}(H'_y \in dx).\end{aligned}$$

Thanks to the first statement of the proposition, H'_y has the same law as H^θ conditioned on being greater than y. Then since $\mathbb{P}(V_z \in dy) = \theta\, e^{-\theta(z-y)}\, dy$, we get

$$\mathbb{P}(H^\theta \in dx) = \mathbb{P}(H_1 \in dx)(1 - \mathbb{P}(H^\theta > x)f(x)) + \mathbb{P}(H^\theta \in dx)\int_0^x \mathbb{P}(H_1 \in dz)f(z),$$

where we have set

$$f(x) := \int_0^x dy\, \theta\, e^{-\theta(x-y)}\, W_\theta(y) \qquad x \geq 0.$$

We can drop the index 1 of H_1, since only its law now matters. We can rewrite the last result as

$$\mathbb{P}(H \in dx) = \mathbb{P}(H^\theta \in dx)(1 - \int_0^x \mathbb{P}(H \in dz) f(z)) + \mathbb{P}(H \in dx)\mathbb{P}(H^\theta > x) f(x),$$

which can be integrated as

$$\mathbb{P}(H > x) = \mathbb{P}(H^\theta > x)(1 - \int_0^x \mathbb{P}(H \in dz) f(z)).$$

Defining now the function G as

$$G(x) := \mathbb{P}(H > x)(W_\theta(x) - f(x)),$$

we get, thanks to the last integration,

$$G(x) = 1 - \int_0^x \mathbb{P}(H \in dz) f(z) - \mathbb{P}(H > x) f(x).$$

Integrating by parts yields

$$G(x) = 1 - \int_0^x dz\, \mathbb{P}(H > z) f'(z) = 1 - \int_0^x dz\, \mathbb{P}(H > z)(-\theta f(z) + \theta W_\theta(z)) = 1 - \theta \int_0^x dz\, G(z),$$

which shows that $G(x) = e^{-\theta x}$. This reads

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

One differentiation and one integration provide the result. □

3.2 Main result

3.2.1 Statement

Recall that $A_n(k)$ denotes the number of haplotypes carried by k individuals in a sample of n.

Theorem 3.2 *For all $k \geq 1$, the following convergence holds a.s.*

$$\lim_{n\to\infty} n^{-1} A_n(k) = \int_0^\infty dx\, \theta\, e^{-\theta x} \frac{1}{W_\theta(x)^2} \left(1 - \frac{1}{W_\theta(x)}\right)^{k-1}.$$

In addition,

$$\lim_{n\to\infty} n^{-1} A_n = \int_0^\infty dx\, \theta\, e^{-\theta x} \frac{1}{W_\theta(x)} = \mathbb{E}\left(1 - e^{-\theta H^\theta}\right).$$

Before proving this statement, we insert a (sub)subsection in which we state and prove a preliminary key result.

3.2.2 The key lemma

Recall that ℓ_{1i} denotes the (time elapsed since the) i-th (most recent) mutation on the first branch length. In particular, the mutations carried by individual 1 and not by individual 0 are exactly those ℓ_{1i} such that $\ell_{1i} < H_1$ (the other points of the process are thrown away). Let N_i denote the number of individuals whose *most recent mutation* is ℓ_{1i}.

Lemma 3.3 *In an infinite sample, for any integer $k \geq 1$,*

$$\sum_{i\geq 1} \mathbb{P}(N_i = k) = \int_0^\infty \theta\, e^{-\theta z}\, dz\, \frac{1}{W_\theta(z)^2} \left(1 - \frac{1}{W_\theta(z)}\right)^{k-1}$$

Proof. In the first place, not to care for the fact that only mutations with $\ell_{1i} < H_1$ contribute, we consider the number N_i' of individuals whose most recent mutation is ℓ_{0i}, and we condition on $\ell_{0j} = v_j$, $j \geq 1$. We will use later the fact that the law of N_i conditional on $\ell_{1j} = v_j$, $j \geq 1$, is that of $N_i' \mathbf{1}_{v_i < H}$, where H is independent of N_i' and the point process $(\ell_{0i}; i \geq 1)$.

Recall from the previous subsection that $\mathcal{E}^\theta$ is the set of individuals who carry no more mutations than individual 0, that K_i^θ is the i-th individual in $\mathcal{E}^\theta$, and $H_i^\theta := H_{K_i^\theta}$. Then set $D_0 := 0$ and

$$D_i := \inf\{j \geq 1 : H_j^\theta > v_{i-1}\} \qquad i \geq 1.$$

Now observe that $N_i' = D_i - D_{i-1}$ for all $i \geq 1$ (for N_1', the count includes individual 0). As an application of Proposition 3.1, we get that conditional on $\ell_{0j} = v_j$, $j \geq 1$,

$$\mathbb{P}(N_1' = k) = \mathbb{P}(H^\theta < v_1)^{k-1} \mathbb{P}(H^\theta > v_1),$$

whereas for any $i \geq 2$,

$$\mathbb{P}(N_i' \neq 0) = \mathbb{P}(H^\theta < v_i \mid H^\theta > v_{i-1}) \quad \text{and} \quad \mathbb{P}(N_i' = k \mid N_i' \neq 0) = \mathbb{P}(H^\theta < v_i)^{k-1} \mathbb{P}(H^\theta > v_i).$$

Recalling the relation between the laws of N_i and N_i' mentioned in the beginning of the proof, we get that conditional on $\ell_{1j} = v_j$, $j \geq 1$,

$$\mathbb{P}(N_1 = k) = \mathbb{P}(H^\theta < v_1)^{k-1} \mathbb{P}(H^\theta > v_1) \mathbb{P}(H > v_1).$$

whereas for any $i \geq 2$,

$$\mathbb{P}(N_i \neq 0) = \mathbb{P}(H^\theta < v_i \mid H^\theta > v_{i-1}) \mathbb{P}(H > v_i).$$

Now $\mathbb{P}(N_i' = k \mid N_i' \neq 0) = \mathbb{P}(N_i = k \mid N_i \neq 0)$, so we finally get (for $i \geq 2$)

$$\begin{aligned} \mathbb{P}(N_i = k) &= \mathbb{P}(H^\theta < v_i)^{k-1} \mathbb{P}(H^\theta < v_i \mid H^\theta > v_{i-1}) \mathbb{P}(H^\theta > v_i) \mathbb{P}(H > v_i) \\ &= \left(1 - \frac{1}{W_\theta(v_i)}\right)^{k-1} \left(1 - \frac{W_\theta(v_{i-1})}{W_\theta(v_i)}\right) \frac{1}{W(v_i) W_\theta(v_i)}. \end{aligned}$$

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

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

so that

$$\begin{aligned}\sum_{i\geq 2}\mathbb{P}(N_i=k) &= \sum_{i\geq 2}\int_0^\infty dz\int_0^z dx\,\frac{\theta^i x^{i-2}}{(i-2)!}\,e^{-\theta z}\left(1-\frac{1}{W_\theta(z)}\right)^{k-1}\left(1-\frac{W_\theta(x)}{W_\theta(z)}\right)\frac{1}{W(z)W_\theta(z)}\\ &= \int_0^\infty dz\,\theta\,e^{-\theta z}\left(1-\frac{1}{W_\theta(z)}\right)^{k-1}\frac{1}{W(z)W_\theta(z)}\int_0^z dx\,\theta\,e^{\theta x}\left(1-\frac{W_\theta(x)}{W_\theta(z)}\right).\end{aligned}$$

Now thanks to Proposition 3.1, we can perform the following integration by parts on the last integral in the last display

$$\begin{aligned}\int_0^z dx\,\theta\,e^{\theta x}\left(1-\frac{W_\theta(x)}{W_\theta(z)}\right) &= \left[e^{\theta x}\left(1-\frac{W_\theta(x)}{W_\theta(z)}\right)\right]_0^z+\frac{1}{W_\theta(z)}\int_0^z dx\,e^{\theta x}W_\theta'(x)\\ &= -1+\frac{1}{W_\theta(z)}+\frac{1}{W_\theta(z)}\int_0^z dx\,W'(x)\\ &= \frac{W(z)}{W_\theta(z)}-1.\end{aligned}$$

This entails

$$\sum_{i\geq 2}\mathbb{P}(N_i=k)=\int_0^\infty dz\,\theta\,e^{-\theta z}\left(1-\frac{1}{W_\theta(z)}\right)^{k-1}\frac{1}{W(z)W_\theta(z)}\left(\frac{W(z)}{W_\theta(z)}-1\right).$$

But since

$$\mathbb{P}(N_1=k)=\int_0^\infty dz\,\theta\,e^{-\theta z}\left(1-\frac{1}{W_\theta(z)}\right)^{k-1}\frac{1}{W(z)W_\theta(z)},$$

the result follows. □

3.2.3 Proof of Theorem 3.2

For each individual $i\geq 0$, we denote by $\mathcal{A}_{ij}$ the set of individuals bearing the unique haplotype whose most recent mutation is ℓ_{ij}. In particular, it is understood that $\mathcal{A}_{ij}=\emptyset$ whenever $\ell_{ij}>H_i$ (because no such haplotype exists).

Now fix $M\geq 1$. Similarly as in the proof of Theorem 2.3, we can define

$$G_M(\Xi_i):=\text{ Card }\{j\geq 1:\text{ Card }\mathcal{A}_{ij}\cap\{i,\ldots,i+M\}\geq k\},$$

where

$$\Xi_i:=((\mathcal{P}_i,H_i),\ldots,(\mathcal{P}_{i+M},H_{i+M})).$$

Observe that G_M is bounded from above, so that $G_M(\Xi_i)$ is integrable for all $i\geq 0$. Now for any $0\leq r\leq M$, the random variables $G_M(\Xi_i)$, for i such that $i=r\ [M+1]$ (standing for mod $(M+1)$), are i.i.d. and integrable, so by the strong law of large numbers, we have the following convergence a.s. (and in L^1)

$$\lim_{n\to\infty}n^{-1}\sum_{0\leq i=r[M+1]\leq n-M}G_M(\Xi_i)=\frac{1}{M+1}\,\mathbb{E}(G_M(\Xi_1)).$$

Summing over r these $M+1$ equalities, we get the following convergence a.s. (and in L^1)

$$\lim_{n\to\infty} n^{-1} \sum_{i=0}^{n-M} G_M(\Xi_i) = \mathbb{E}(G_M(\Xi_1)).$$

Our goal is now to let $M \to \infty$. First define

$$A'_n(k) := \sum_{i=0}^{n} \text{ Card } \{j \geq 1 : \text{ Card } \mathcal{A}_{ij} \cap \{i, \ldots, n\} \geq k\}.$$

Notice that

$$A'_n(k) = \sum_{h \geq k} A_n(h).$$

Then for any $i = 0, \ldots, n-M$, for any $j \geq 1$, if Card $\mathcal{A}_{ij} \cap \{i, \ldots, i+M\} \geq k$, then Card $\mathcal{A}_{ij} \cap \{i, \ldots, n\} \geq k$, so that $A'_n(k) \geq \sum_{i=1}^{n-M} G_M(\Xi_i)$, and

$$\liminf_{n\to\infty} n^{-1} A'_n(k) \geq \liminf_{n\to\infty} n^{-1} \sum_{i=1}^{n-M} G_M(\Xi_i) = \mathbb{E}(G_M(\Xi_1)).$$

Letting $M \to \infty$, Beppo Levi's theorem yields

$$\liminf_{n\to\infty} n^{-1} A'_n(k) \geq \mathbb{E}\left[\text{Card } \{j \geq 1 : \text{ Card } \mathcal{A}_{1j} \geq k\}\right] = \sum_{j \geq 1} \mathbb{P}(\text{Card } \mathcal{A}_{1j} \geq k) =: y_k.$$

In the notation of the previous subsection Card $\mathcal{A}_{1j} = N_j$, so by Fubini–Tonelli's theorem,

$$y_k = \sum_{j \geq 1} \mathbb{P}(N_j \geq k) = \sum_{j \geq 1} \sum_{h \geq k} \mathbb{P}(N_j = h) = \sum_{h \geq k} x_h,$$

where $x_k := \sum_{j \geq 1} \mathbb{P}(N_j = k)$. Thanks to Lemma 3.3 we have the following explicit expression for x_k

$$x_k = \int_0^\infty \theta\, e^{-\theta z}\, dz\, \frac{1}{W_\theta(z)^2} \left(1 - \frac{1}{W_\theta(z)}\right)^{k-1}.$$

Now recall that $\sum_{k \geq 1} A'_n(k) = \sum_{h \geq 1} h A_n(h) = n$. Since it is easily seen that $\sum_{k \geq 1} y_k = \sum_{h \geq 1} h x_h = 1$, by Fatou's lemma

$$1 = \sum_{k \geq 1} y_k \leq \sum_{k \geq 1} \liminf_n n^{-1} A'_n(k) \leq \liminf_n n^{-1} \sum_{k \geq 1} A'_n(k) = 1.$$

Then we would get a contradiction if there was k_0 such that $\liminf_n n^{-1} A'_n(k_0) > y_{k_0}$, so that for all $k \geq 1$ a.s.,

$$\lim_{n\to\infty} n^{-1} A'_n(k) = y_k.$$

The first equation of the theorem stems from the fact that $A_n(k) = A'_n(k) - A'_n(k+1)$ and the second one by taking $k = 1$ in the last display. It takes an elementary integration by parts to check that

$$y_1 = \int_0^\infty dx\, \theta\, e^{-\theta x} \frac{1}{W_\theta(x)} = \mathbb{E}\left(1 - e^{-\theta H^\theta}\right).$$

Acknowledgments. This work was partially funded by the project MAEV 'Modèles Aléatoires de l'Évolution du Vivant' of ANR (French national research agency).

References

[1] Abraham, R., Delmas, J.F. (2008)
Williams' decomposition of the Lévy continuous random tree and simultaneous extinction probability for populations with neutral mutations. *Stoch. Proc. Appl.* doi:10.1016/j.spa.2008.06.001

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

[3] Arratia, R., Barbour, A.D., Tavaré, S. (1992)
Poisson process approximations for the Ewens sampling formula. *Ann. Appl. Probab.* **2** 519–535.

[4] Basdevant, A.L., Goldschmidt, C. (2008)
Asymptotics of the allele frequency spectrum associated with the Bolthausen–Sznitman coalescent. Preprint arXiv:0706.2808v1

[5] Berestycki, J., Berestycki, N., Schweinsberg, J. (2007)
Beta-coalescents and continuous stable random trees. *Ann. Probab.* **35** 1835–1887.

[6] Bertoin, J. (1996)
Lévy processes. Cambridge University Press, Cambridge.

[7] Bertoin, J. (2008)
The structure of the allelic partition of the total population for Galton-Watson processes with neutral mutations. Preprint arXiv:0711.3852

[8] Donnelly, P., Tavaré, S. (1986)
The ages of alleles and a coalescent. *Adv. Appl. Probab.* **18** 1–19.

[9] Durrett, R. (2008)
Probability Models for DNA Sequence Evolution. Springer–Verlag, Berlin. 2nd revised ed.

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

[11] Fisher, R.A., Corbet, S.A., Williams, C.B. (1943)
The relation between the number of species and the number of individuals in a random sample of an animal population. *J. Anim. Ecol.* **12** 42–58.

[12] 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.

[13] Kingman, J.F.C. (1982)
The coalescent. *Stochastic Process. Appl.* **13** 235–248.

[14] Lambert, A. (2008)
The contour of splitting trees is a Lévy process. Preprint arXiv:0704.3098v1

[15] Lambert, A. (2008) Population Dynamics and Random Genealogies. *Stoch. Models* **24** 45–163.

[16] Möhle, M. (2006)
On the number of segregating sites for populations with large family sizes. *Adv. Appl. Prob.* **38** 750–767.

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