Genealogical properties of subsamples in highly fecund populations

Bjarki Eldon
Museum für Naturkunde
43 Invalidenstraße
10115 Berlin
Germany

Fabian Freund
University of Hohenheim
Institute 350b
Fruwirthstraße 21
D-70599 Stuttgart, Germany

January 9, 2018

Abstract We consider some genealogical properties of nested samples. The complete sample is assumed to have been drawn from a natural population characterised by high fecundity and sweepstakes reproduction (abbreviated HFSR). The random gene genealogies of the samples are — due to our assumption of HFSR — modelled by coalescent processes which admit multiple mergers of ancestral lineages looking back in time. Among the genealogical properties we consider are the probability that the most recent common ancestor is shared between the complete sample and the subsample nested within the complete sample; we also compare the lengths of 'internal' branches of nested genealogies between different coalescent processes. The results indicate how 'informative' a subsample is about the properties of the larger complete sample, how much information is gained by increasing the sample size, and how the 'informativeness' of the subsample varies between different coalescent processes.

keywords: coalescent; high fecundity; nested samples; multiple mergers; time to most recent common ancestor

AMS subject classification: 92D15, 60J28

Contents

1 Introduction . . . 1
2 Sharing the MRCA . . . 3
3 Relative times and lengths . . . 12
4 Proofs . . . 22
5 Conclusion and open questions . . . 30
A1 Population models . . . 35
A2 Coalescent processes . . . 37
A3 Goldschmidt and Martin's construction of the Bolthausen-Sznitman n-coalescent . . . 39

1 Introduction

The study of the evolutionary history of natural populations usually proceeds by drawing inference from a random sample of DNA sequences. To this end the coalescent approach initiated by [56,58,57,83,52] - i.e. the probabilistic modeling of the random ancestral relations of the sampled DNA sequences - has proved to be very useful [85, cf.]. Inference based on the coalescent relies on the key assumption, as in standard statistical inference, that the evolutionary history of the (finite) sample approximates, or is informative about, the evolutionary history of the population from which the sample is drawn. We would like to know how much some basic genealogical sample-based statistics tell us about the population in a multiple-merger coalescent framework. Does the 'informativeness' of the various genealogical statistics depend on the underlying coalescent process? A more practical approach to this question is, instead of comparing a sample with the population, to ask how much of the genetic information of a sample is already contained in a subsample, i.e. what is gained by enlarging the sample? A related question concerns the size of the sample; i.e. how large does our sample need to be for a reliable inference? Do standard genetic approaches or guidelines, e.g. about

sample size for population genetic studies, still hold true for populations characterised by high fecundity and sweepstakes reproduction (abbreviated HFSR)? In Sec. A1 we give a brief overview of population models of reproduction which admit HFSR. The coalescent processes derived from HFSR population models admit multiple-mergers of ancestral lineages (see Sec. A2).

We approach these problems by studying some genealogical properties of nested samples, by which we mean where a sample (a subsample) is drawn (uniformly at random without replacement) from a larger sample (the complete sample). By way of an example, [74] consider nested samples whose ancestries are governed by the Kingman n-coalescent [56,58,57]. One of the results of [74] concerns the probability that a subsample shares its most recent common ancestor (abbreviated MRCA) with the complete sample. In the case the complete sample and the subsample share the MRCA they also, with high probability, share the oldest genealogical information. Thus, the effects on the genetic structure of the complete sample of the oldest part of the genealogy are also present in the subsample. In addition, the complete sample and the subsample have had exactly the same timespan to collect mutations. [74] show that the probability that a subsample of a fixed size m shares the MRCA with the complete sample of arbitrarily large size n ($n \to \infty$) converges to $(m-1)/(m+1)$. Even a subsample of size 2 shares the MRCA with probability $1/3$, while a sample of size 19 already shares with probability 0.9. This shows that by this measure (the probability of sharing the MRCA) even a rather small subsample drawn from a large complete sample whose ancestry is governed by the Kingman coalescent captures properties of the complete sample quite well.

The outline of the paper is as follows. In Section 2 we introduce our key object: the probability that the subsample shares the MRCA with the complete sample (see Eq. (3)). In Section 2 we present results for finite sample size, namely Prop. 1 regarding comparing the probability (3) between certain coalescent processes (see Sec. A2 for a precise description of the coalescent processes we consider), Eq. (4) for a recursion to compute (3) exactly for any Λ-coalescent (see Eq. (A33)), Prop. 2 which gives a general representation of probability (3) for any Ξ-coalescent (see Eq. A32) — and thus for any Λ-coalescent — and Prop. 3 which gives a representation of (3) for the Bolthausen-Sznitman coalescent (a specific multiple-merger coalescent, the Beta-coalescent (A36) with $\alpha = 1$). In Section 2.3 we present our main mathematical result (Thm. 1), a representation of the probability (3) as sample size $n \to \infty$ for the Beta-coalescent (see Eq. (A36)). We also give a criterion for when the limit of (3), as $n \to \infty$, stays positive under a general Ξ-coalescent. In Sec. 2.2 we discuss the probability of sharing the oldest allele between the subsample and the larger sample, and the probability of monophyly of the subsample, and we present recursions for these probabilities. In Sec. 3 we investigate by simulations the fraction of internal branch lengths covered by the subsample. Proofs of our mathematical results are presented in Section 4. A brief discussion of the implication of our results, and open problems, is given in Section 5. Section A1 contains a brief description of the population models underlying the coalescent processes we consider, Section A2 contains a detailed description of the coalescent processes, and Section A3 a review of Goldschmidt and Martin's construction of the Bolthausen-Sznitman n-coalescent [39].

For ease of reference we include a table (Table 1) of notation and terminology.

Table 1 Notation and terminology.

symbol	explanation
HFSR	high fecundity and sweepstakes reproduction
MMC	multiple-merger coalescent
MRCA	most recent common ancestor
TMRCA	time to MRCA
leaves	special kind of vertices in a random graph (genealogy); correspond to sampled DNA sequences
n-coalescent	a coalescent process started from n leaves
$\mathbb{N}$	the set of the natural numbers $\mathbb{N} := \{1,2,\ldots\}$
$[n]$	$[n] := \{1,2,\ldots,n\}$, $n \in \mathbb{N}$
$[n]_a$	$[n]_a := \{a,a+1,\ldots,n\}$ for $n,a \in \{0\} \cup \mathbb{N}$, $a \leq n$
$\mathscr{P}_n$	set of all partitions of $[n]$
$\mathbb{1}_{(A)}$	$\mathbb{1}_{(A)} = 1$ if A holds, and zero otherwise
$x \wedge y$	$\min\{x,y\}$
$T_{\mathrm{MRCA}}^{(\infty)}$	the random TMRCA of the population current at some stated time
$T_{\mathrm{MRCA}}^{(n)}$	the random TMRCA of a sample of size $n \in [2,\infty)$
$T_{\mathrm{MRCA}}^{(m;n)}$	the random TMRCA of a subsample of size m taken from a complete sample of size $n > m$
$T_{\mathrm{MRCA}}^{(M)}$	the random TMRCA of a finite sample $M \subset \mathbb{N}$
Π	coalescent process; $\Pi \equiv \Pi_t := \{\Pi(t), t \geq 0\}$
$\Pi^{(n)}$	Π restricted to $[n]$
$\Pi^{(\Lambda)}$	Λ-coalescent
$\Pi^{(\Xi)}$	Ξ-coalescent
$\mathbb{P}^{(\Pi)}(A)$	probability of event A under Π
$p_{n,m}^{(\Pi)}$	$p_{n,m}^{(\Pi)} := \mathbb{P}^{(\Pi)}\left(T_{\mathrm{MRCA}}^{(m;n)} = T_{\mathrm{MRCA}}^{(n)}\right)$; the probability that subsample and complete sample share the MRCA
Δ	the infinite simplex $\Delta := \{(x_1,x_2,\ldots) \vert x_i \in [0,1],\ \sum_{i\in\mathbb{N}} x_i \leq 1\}$
$\rho_T^{(m;n)}$	the ratio $T_{\mathrm{MRCA}}^{(m;n)}/T_{\mathrm{MRCA}}^{(n)}$; see Sec. 3.1
$\rho_I^{(m;n)}$	the ratio of 'internal' edge lengths between subsample and complete sample; see Sec. 3.1

2 Sharing the MRCA

We consider a Ξ- or Λ-n-coalescent with a starting partition $\pi = \{\{1\},\ldots,\{n\}\}$, i.e. initially all the blocks $\pi_i \in \pi$ are singleton blocks. We refer to the elements of the starting partition as 'leaves'. A common ancestor of a set $A \subset \mathbb{N}$ of leaves is any block containing A. A set A of leaves has a common ancestor if and only if the coalescent passes through a partition with a block containing A. This allows us to identify the common ancestor with blocks of the partition-valued states of the coalescent. The MRCA of a set A of leaves is the smallest block which contains A (whenever that block appears). Given that we start from a finite set $[n]$ of leaves ($n < \infty$) we will eventually (i.e. in finite time almost surely) observe the partition $\{[n]\}$ containing only the block $[n]$. Write $\Pi_t^{(n)}$ for the partition reached at time t in the case when the coalescent process is started from n leaves. Let $T_{\mathrm{MRCA}}^{(n)}$ denote the random time to the MRCA (abbreviated TMRCA) of the set $[n]$ of leaves, i.e. we define

$$T_{\mathrm{MRCA}}^{(n)} := \inf\left\{t \geq 0 : \Pi_t^{(n)} = \{[n]\}\right\}. \tag{1}$$

The $T^{(n)}_{\mathrm{MRCA}}$ is therefore the first time $\Pi^{(n)}_t$ arrives at partition $\{[n]\} \in \mathscr{P}_n$, where $\mathscr{P}_n$ denotes the set of all partitions of $[n]$. Write $T^{(\infty)}_{\mathrm{MRCA}}$ for the TMRCA of the whole population. By a finite sample we mean a finite set A of leaves.

A subsample is a subset of a given sample (a given set of leaves). We let m denote the size of the subsample. For convenience and w.l.o.g. we assume leaves 1 to m are the leaves of the subsample, and we assume block π_1 in any partition always contains element 1. A common ancestor of the subsample is any block containing $[m]$; the MRCA of the subsample is the smallest block containing $[m]$ (whenever it appears). We define the TMRCA of a subsample of size m of a sample of size $n \geq m$ as

$$T^{(m;n)}_{\mathrm{MRCA}} := \inf\left\{t \geq 0 : [m] \subseteq \pi_1 \in \Pi^{(n)}_t\right\}; \tag{2}$$

i.e. $T^{(m;n)}_{\mathrm{MRCA}}$ is the time of first occurrence of the subset $[m]$ in block π_1 in a partition of $\Pi^{(n)}$. The sample and the subsample share the MRCA if the smallest block containing $[m]$ ever observed in $\Pi^{(n)}$ is $[n]$; this happens almost surely if $T^{(m;n)}_{\mathrm{MRCA}} = T^{(n)}_{\mathrm{MRCA}}$.

Our main mathematical results concern the probability

$$p^{(\Pi)}_{n,m} := \mathbb{P}^{(\Pi)}\left(T^{(m;n)}_{\mathrm{MRCA}} = T^{(n)}_{\mathrm{MRCA}}\right), \tag{3}$$

which is the probability that the sample (of size n) and the nested subsample (of size $m < n$) share their MRCA under the coalescent process Π. From now on it should be understood that we always look at nested samples. We are able to obtain representations of $p^{(\Pi)}_{n,m}$ both for finite n and m and also for the limit $\lim_{n\to\infty} p^{(\Pi)}_{n,m}$, m fixed, for some multiple-merger coalescent processes. We will let $p^{(\Xi)}_{n,m}$ denote $p^{(\Pi)}_{n,m}$ in (3) when Π is a Ξ-coalescent, and $p^{(\Lambda)}_{n,m}$ denote $p^{(\Pi)}_{n,m}$ when Π is a Λ-coalescent.

2.1 Finite n

Our main focus is to compare genealogical properties of nested samples between different coalescent processes in order to learn what is gained by enlarging the sample size. In this context, a natural question to address is which n-coalescent Π maximises $p^{(\Pi)}_{n,m}$ for a given finite sample size n and subsample size m? In the context of Λ-coalescents (see Eq. (A33) in Sec. A2) this is the 'star-shaped' coalescent with Λ-measure $\Lambda(dx) = \delta_1(x)dx$ so that $\Lambda(\{1\}) = 1$, all n blocks merge after an exponential waiting time, and $p^{(\delta_1)}_{n,m} = 1$. We now compare $p^{(\mathrm{Kingman})}_{n,m}$ (meaning $p^{(\Pi)}_{n,m}$ when Π is the Kingman-coalescent) to all $p^{(\Lambda)}_{n,m}$ with $\Lambda(\{1\}) = 0$. We can show the following (see Sec. 4.1 for a proof).

Proposition 1 *For any given sample size n and subsample size $m < n$ there is a Λ' with $\Lambda'(\{1\}) = 0$ which fulfills $p^{(\Lambda')}_{n,m} > p^{(Kingman)}_{n,m}$.*

One can think of Λ' as given by $\Lambda = \delta_\psi$ for some fixed $\psi \in (0,1)$ and very close to 1. Prop. 1 holds for any finite sample size n and subsample size m. Regarding the limit $p_m^{(\Pi)} = \lim_{n\to\infty} p_{n,m}^{(\Pi)}$ with m fixed we conjecture that $p_m^{(\text{Kingman})} > p_m^{(\Lambda)}$ for every Λ-coalescent with $\Lambda(\{1\}) = 0$. Should our conjecture be true, the limits compare in the opposite way to the comparison of the non-limit probabilities given in Prop. 1.

The result in Prop. 1 holds for a very special Λ-coalescent. One can numerically evaluate $p_{n,m}^{(\Lambda)}$ for any Λ-coalescent with a recursion (see Sec. 4.7.1 for a proof), and thus compare $p_{n,m}^{(\Lambda)}$ for different Λ-coalescents. Let $\lambda(n)$ (see Eq. (A34)) denote the total rate of mergers given n blocks, and $\lambda_k(n) = \binom{n}{k}\lambda_{n,k}$ (see Eq. (A33)) denote the rate at which any k of n blocks merge. Write $\beta(n, n-k+1) := \lambda_k(n)/\lambda(n)$ for the probability of a single merger of k blocks (a k-merger) given n blocks ($2 \le k \le n$). Then

$$p_{n,m}^{(\Lambda)} = \sum_{k=2}^{n} \beta(n, n-k+1) \sum_{\ell=0}^{k\wedge m} \frac{\binom{n-m}{k-\ell}\binom{m}{\ell}}{\binom{n}{k}} p_{n-k+1,m'}^{(\Lambda)}. \tag{4}$$

where $\binom{n-m}{k-\ell} := 0$ if $n-m < k-\ell$ and $m' = (m-\ell+1)\mathbb{1}_{(\ell>1)} + m\mathbb{1}_{(\ell\le 1)}$. In the case $m = 2$ recursion (4) simplifies to

$$\begin{aligned} p_{n,2}^{(\Lambda)} = &\sum_{k=2}^{n-2} \beta(n, n-k+1) \frac{(n-k)(n+k-1)}{n(n-1)} p_{n-k+1,2}^{(\Lambda)} \\ &+ \beta(n,2)\frac{2}{n} + \beta(n,1). \end{aligned} \tag{5}$$

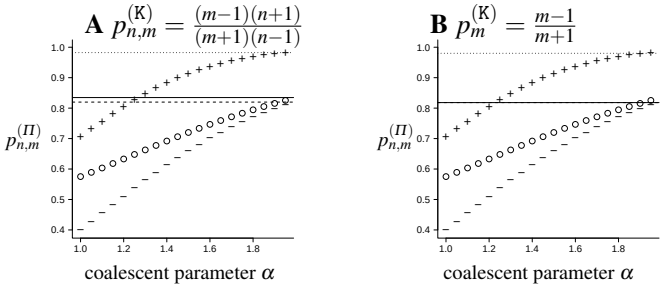
Recursion (4) further simplifies in the case of the Kingman coalescent, since then $\beta(n, n-1) = 1$ for $n \ge 2$. [74] obtain

$$p_{n,m}^{(\text{Kingman})} = \frac{m-1}{m+1}\frac{n+1}{n-1}. \tag{6}$$

Since the representation (6) only depends on which mergers are possible, the result (6) holds for a time-changed Kingman-coalescent as derived for example in [54] from a population model of 'modest' changes in population size.

The Beta-coalescent (see Eq. (A35)) with coalescent parameter $\alpha \in [1,2)$, is an example of a Λ-coalescent (see Eq. (A33)) and can be derived from population model (A28). Figure 1 shows graphs of $p_{n,m}^{(\Pi)}$ when Π is the Beta-coalescent (see Eq. (A35)) as a function of α; the results indicate that $p_{n,m}^{(\text{Beta-coal})} < p_{n,m}^{(\text{Kingman})}$ for n large enough and any m. This shows that one needs a larger subsample under the Beta-coalescent than under the Kingman-coalescent for a given sample size to have the same value of $p_{n,m}^{(\Pi)}$. By implication, one gains more information by enlarging the sample under the Beta-coalescent than under the Kingman-coalescent.

Fig. 1 Graphs of $p_{n,m}^{(\mathrm{Beta-coal})}$ (see Eq. (4)) as a function of α for $(n,m) = (10^2, 10^1)$ (circles); $(10^3, 10^1)$ $(-)$; $(10^3, 10^2)$ $(+)$. The corresponding results for the Kingman-coalescent, $p_{n,m}^{(\mathrm{K})}$ (**A**) and $p_m^{(\mathrm{K})}$ (**B**) are shown as lines.



168

We conclude this subsection with two closed-form representations of $p_{n,m}^{(\Pi)}$. To prepare for the first one we recall the concept of 'coming down from infinity'. This property is defined as follows. If a Ξ-n-coalescent (see Eq. (A32)) $(\Pi_t^{(\Xi)})_{t\geq 0}$ comes down from infinity then, with probability 1, the number of blocks is finite for any $t>0$, which is equivalent to $\lim_{n\to\infty} T_{\text{MRCA}}^{(n)} < \infty$ a.s. If $\Pi_t^{(\Xi)}$, for all $t>0$, has infinitely many blocks with probability 1, we say that the coalescent 'stays infinite'. Conditions for Ξ to fall into one of these two classes are available, see e.g. [78,77,49]. If $\Xi(\{\boldsymbol{x}\in\Delta \,|\, \sum_{i=1}^k x_i = 1 \text{ for } k\in\mathbb{N}\}) > 0$, the Ξ-coalescent does not stay infinite [77, p.39], but does not necessarily come down from infinity. In fact, there is a.s. a finite (random) time $T\geq 0$ so that the number of blocks is finite for all $t>T$ (see [76, p. 39]). This means that for such a coalescent, $\lim_{n\to\infty} T_{\text{MRCA}}^{(n)}$ is finite almost surely. For processes that stay infinite $(\tilde{\Pi})$, $\lim_{n\to\infty} p_{n,m}^{(\tilde{\Pi})} = 0$ since the MRCA of the set $\mathbb{N}$ of leaves in the starting partition $\{\{1\},\{2\},\ldots\}$ is never reached.

We have a representation of $p_{n,m}^{(\Xi)}$ (see Sec. 4.2 for a proof). This representation allows us to later derive characterisations of $\lim_{n\to\infty} p_{n,m}^{(\Pi)}$ for different multiple-merger coalescents Π, see Thm. 1 and Prop. 5. For example, we use Eq. 8 in Prop. 2 to prove Theorem 1.

Proposition 2 *For any finite measure Ξ on Δ, we have*

$$p_{n,m}^{(\Xi)} = 1 - \mathbb{E}\left[\sum_{i\in\mathbb{N}} \prod_{\ell=0}^{m-1} \frac{B_{[i]}^{(n)} - \ell}{n-\ell}\right] > 0, \tag{7}$$

where $B_{[1]}^{(n)}, B_{[2]}^{(n)}, \ldots$ are the sizes of the blocks of $\Pi_{T_{MRCA}^{(n)}-}^{(n)}$, ordered by size from biggest to smallest where the sequence $B_{[1]}^{(n)}, B_{[2]}^{(n)}, \ldots$ is extended to an infinite sequence by taking $B_{[i]}^{(n)} = 0$ for $i > \#\Pi_{T_{MRCA}^{(n)}-}^{(n)}$. If the Ξ-coalescent comes down from infinity, we have

$$p_{n,m}^{(\Xi)} \to 1 - \mathbb{E}\left[\sum_{i\in\mathbb{N}} P_{[i]}^m\right] = 1 - \mathbb{E}\left[X^{m-1}\right] = 1 - \frac{\mathbb{E}\left[Y^m\right]}{\mathbb{E}\left[Y\right]} > 0 \tag{8}$$

for fixed m and $n\to\infty$, where $P_{[i]} := \lim_{n\to\infty} B_{[i]}^{(n)}/n$ is the (almost surely existing) asymptotic frequency of the ith biggest block of $\Pi_{T_{MRCA}^{(\infty)}-}$, X is the asymptotic frequency of a size-biased pick from the blocks of $\Pi_{T_{MRCA}^{(\infty)}-}$, while Y is the asymptotic frequency of a block picked uniformly at random from $\Pi_{T_{MRCA}^{(\infty)}-}$.

In the case of the Bolthausen-Sznitman (BS-coal) n-coalescent [20], which is a Λ-n-coalescent with $\Lambda(dx) = dx$ (see Eq. (A33)), i.e. the density associated with the uniform distribution on $[0,1]$, we can give a characterisation of $p_{n,m}^{(\Pi)}$ in terms of independent Bernoulli r.v.'s (see Sec. 4.3 for a proof). We use Eq. 9 in Prop. 3 to prove Prop. 5.

Proposition 3 *Let $B_1, \ldots, B_{n-1}$ be independent Bernoulli random variables with $\mathbb{P}(B_i = 1) = 1/i$. Let Π denote the Bolthausen-Sznitman n-coalescent. For $2 \leq m < n$,*

$$p_{n,m}^{(\Pi)} = \mathbb{E}\left[\frac{B_1 + \ldots + B_{m-1}}{B_1 + \ldots + B_{n-1}}\right]. \tag{9}$$

Moreover, $\log n p_{n,m}^{(\Pi)} \to \sum_{i=1}^{m-1} i^{-1}$ *for $n \to \infty$ and m fixed.*

2.2 Two variants of $p_{n,m}^{(\Pi)}$

2.2.1 Including the oldest allele

The probability $p_{n,m}^{(\Pi)}$ (see Eq. (4)) is an indication of how likely it is that the 'oldest' genealogical branches, or the edges connected directly to the MRCA, are (partially) shared between the subsample and the complete sample. We remark that the complete sample and the subsample may share the MRCA without sharing any of the 'internal' edges — i.e. an edge subtended by at least 2 leaves (e.g. the marked edges in Fig. 2A) — if the associated coalescent admits multiple mergers (see Fig. 2C for an example). Such events are highly unlikely though for n large enough if the Λ-coalescent comes down from infinity, see Corollary 1. If the complete sample and the subsample share the MRCA then the subsample is more likely to include the 'oldest allele' — i.e. the allele that arose closest to the root — of the complete sample. To derive the actual probability of the event that the subsample carries the oldest allele of the complete sample one needs to include mutation. Consider a Λ-n-coalescent with neutral mutation. Mutations are modelled by a homogeneous Poisson point process on the branches of the Λ-n-coalescent with (scaled) mutation rate $\theta > 0$. We assume the infinitely-many-alleles model. This means that the allelic type of each individual is seen by tracing its ancestral line back to the first mutation on it. The ancestral line shares the type of the MRCA if there is no mutation on the line before the MRCA is reached. We are interested in the event that the oldest allele from the complete sample is also found in the subsample. The probability of this event has been discussed in case of the Kingman's n-coalescent [74] (see Eq. 5.13). For multiple-merger coalescents this probability can be expressed by using the concept of 'frozen' and 'active' ancestral lines in a n-coalescent with mutation [23]. At a given time t, an ancestral lineage is called frozen if there has been a mutation on it, otherwise it is called active. The age of a sampled allele (i, say) is the waiting time τ_i until its' ancestral lineage is frozen. For consistency we prolong the n-coalescent after reaching the MRCA (at time $T_{\mathrm{MRCA}}^{(n)}$) by a single ancestral line. The first mutation on the prolonged line is seen after an additional $\mathrm{Exp}(\theta/2)$ time which freezes the line. Thus, the oldest allele of a sample is given by the ancestral lineage which is frozen last (active the longest), and this age is $\max\{\tau_i : i \in [n]\}$ for the sample and $\max\{\tau_i : i \in [m]\}$ for the subsample. Let $A^{(n)}(t)$ denote the count of active ancestral lineages in the sample at time t. We write

$$p_{n,m}^{(\Pi,\theta)} := \mathbb{P}^{(\Pi,\theta)}\left(A^{(n)}(\max\{\tau_i : i \in [m]\}) = 0\right) \tag{10}$$

for the probability that the subsample includes the oldest allele of the sample.

We consider $p_{n,m}^{(\Lambda,\theta)} \equiv p_{n,m}^{(\Pi^{(\Lambda)},\theta)}$ for $n,m \in \mathbb{N}_0$, $\theta > 0$. The case $n = m = 1$ (or $n > m = 1$) means we trace back a single lineage until it is hit by a mutation (either in the sample and/or subsample). The boundary conditions are $p_{n,n}^{(\Lambda,\theta)} = 1$ and $p_{n,0}^{(\Lambda,\theta)} = 0$ for $n > 0$. The recursion for $p_{n,m}^{(\Lambda,\theta)}$ is

$$\begin{aligned} p_{n,m}^{(\Lambda,\theta)} &= \frac{\theta m}{2\lambda(n)+\theta n} p_{n-1,m-1}^{(\Lambda,\theta)} + \frac{\theta(n-m)}{2\lambda(n)+\theta n} p_{n-1,m}^{(\Lambda,\theta)} \\ &\quad + \frac{2\lambda(n)}{2\lambda(n)+\theta n} \sum_{k=2}^{n} \beta(n,n-k+1) \sum_{\ell=0}^{k\wedge m} \frac{\binom{n-m}{k-\ell}\binom{m}{\ell}}{\binom{n}{k}} p_{n-k+1,m'}^{(\Lambda,\theta)}, \end{aligned} \tag{11}$$

where $m' = (m-\ell+1)\mathbb{1}_{(\ell>1)} + m\mathbb{1}_{(\ell\leq 1)}$ (see Sec. 4.7.3 for a proof),

The probability $p_{n,m}^{(\Lambda,\theta)}$ is a function of the scaled mutation rate θ. Here, and in most models in population genetics which include mutation, $\theta := \mu_N/c_N$ where μ_N is the rate of mutation per locus per generation, and c_N is the pairwise coalescence probability, or the probability that 2 distinct individuals sampled at the same time from a population of size N have the same parent. Since (usually) one arranges things so that $c_N \to 0$ as $N \to \infty$ to ensure convergence to a continuous-time limit [71,64], and since θ is usually assumed to be of order $\mathscr{O}(1)$, we let μ_N depend on N. The key point here is that θ depends on c_N. By way of an example, $c_N = 1/N$ for the haploid Wright-Fisher model, while $c_N = \mathscr{O}(N^{1-\alpha})$ for the Beta$(2-\alpha,\alpha)$-coalescent, $1 < \alpha < 2$ [72]. This means that the scaled mutation rates (θ) are not directly comparable between different coalescent processes; this again means that expressions ($p_{n,m}^{(\Lambda,\theta)}$, defined in (10), for example) which depend on the mutation rate cannot be directly compared between different coalescent processes that may have different timescales. We further remark that we must define θ to be proportional to $1/c_N$ since the branch lengths on which the mutation process runs are in units of $1/c_N$ (i.e. 1 coalescent time unit corresponds to $\lfloor 1/c_N \rfloor$ generations); thus if we don't rescale the mutation rate μ_N with $1/c_N$ we would never see any mutations. It is therefore the mutation rate μ_N, which must be determined from molecular (or DNA sequence) data, which determines the timescale; the quantity c_N comes from the model.

2.2.2 The smallest block containing $[m]$

The probability $p_{n,m}^{(\Pi)}$ is also the probability of the event that the MRCA of the subsample (of size m) subtends all the n leaves ($[n]$ is the smallest block containing $[m]$). A related more general question is to ask about the distribution of the size (number of elements) of the smallest block which contains $[m]$. This is the same as asking about the distribution of the number of leaves subtended by the MRCA of the subsample. For Kingman's n-coalescent, the distribution is computed in [82, Thm. 1]. The probability of the event that the MRCA of the subsample subtends only the leaves of the subsample is especially interesting, see e.g. [88, p. 184, Eq. 2], where this probability is described recursively in the case of the Kingman-coalescent. This recursion can be easily extended to Λ-coalescents. Define $T_{\mathrm{MRCA}}^{(A)}$

to be the first time that A is completely contained within a block of Π_t. Write

$$q_{n,m}^{(\Lambda)} := \mathbb{P}^{\left(\Pi^{(\Lambda)}\right)}\left(T_{\text{MRCA}}^{([m]\cup\{i\})} > T_{\text{MRCA}}^{([m])}\ \forall\, i \in \{m+1,\ldots,n\}\right) \tag{12}$$

for the probability that the MRCA of the subsample subtends only the leaves of the subsample (leaves 1 to m). Let $\beta(n,n-k+1)$ be the probability of a k-merger $(2 \le k \le n)$ given n active lines. The recursion for $q_{n,m}^{(\Lambda)}$ is (see Sec. 4.7.2 for a proof)

$$q_{n,m}^{(\Lambda)} = \sum_{k=2}^{(n-m)\wedge m} \frac{\beta(n,n-k+1)}{\binom{n}{k}}\left(\binom{m}{k} q_{n-k+1,m-k+1}^{(\Lambda)} + \binom{n-m}{k} q_{n-k+1,m}^{(\Lambda)}\right) \tag{13}$$

with boundary conditions $q_{n,n}^{(\Lambda)} = q_{n,1}^{(\Lambda)} = 1$ for $n \in \mathbb{N}$. One may use $q_{n,m}^{(\Lambda)}$ to calculate the p-value of a test for monophyly or non-random mating (see Discussion in [82]), i.e. to calculate the p-value of a test for observing block $[m]$ under the null-hypothesis that the Λ-coalescent models the genealogy.

As one might expect (see Sec. 4.3 for a proof), for m fixed and for any Λ-coalescent,

$$\lim_{n\to\infty} q_{n,m}^{(\Lambda)} = 0. \tag{14}$$

In the case of the Bolthausen-Sznitman (BS-coal) n-coalescent we obtain an exact representation of $q_{n,m}^{(\text{BS-coal})}$ (see Sec. 4.4 for a proof).

Proposition 4 *Let $B_1,\ldots,B_{n-1},B'_1,\ldots,B'_{n-m}$ be independent Bernoulli variables with $\mathbb{P}(B_i=1) = \mathbb{P}(B'_i=1) = i^{-1}$. For the Bolthausen-Sznitman n-coalescent we have, for $2 \le m < n$,*

$$q_{n,m}^{(BS\text{-}coal)} = \binom{n-1}{m-1}^{-1} \mathbb{E}\left[\begin{pmatrix} \sum_{i\in[m-1]} B_i + \sum_{i\in[n-m]} B'_i \\ \sum_{i\in[m-1]} B_i \end{pmatrix}^{-1}\right]. \tag{15}$$

2.3 The limit $\lim_{n\to\infty} p_{n,m}^{(\Pi)}$

As we stated in the Introduction the aim of modelling the random genealogy of a sample of DNA sequences drawn from some population is to learn about the evolutionary history of the population. We are therefore interested in investigating the behaviour of the genealogical statistics within our framework of nested samples as the size of the complete sample is allowed to be arbitrarily large, but keeping the size of the subsample fixed. In this subsection we discuss the limit $\lim_{n\to\infty} p_{n,m}^{(\Pi)}$ with m fixed. For a fixed $m \in \mathbb{N}$, write

$$p_m^{(\Pi)} := \lim_{n\to\infty} \mathbb{P}^{(\Pi)}\left(T_{\text{MRCA}}^{(m;n)} = T_{\text{MRCA}}^{(n)}\right) \tag{16}$$

for the probability, under coalescent Π, that a subsample of size m shares the MRCA with an arbitrarily large sample. The limit $\lim_{n\to\infty} T_{\text{MRCA}}^{(n)}$ is a valid limit for any coalescent (even if it diverges) and therefore (16) is well defined. For any

Ξ-coalescent $\left(p_{n,m}^{(\Xi)}\right)_{n>m}$ is monotonically decreasing as n increases. The limit $p_m^{(\Xi)}$ is derived under the assumption that the same Ξ-coalescent is obtained for arbitrarily large sample size. This assumption may not hold when one wants to relate to finite real populations. The quantity $p_m^{(\Pi)}$ should therefore only be regarded as a limit. See further discussion on this point in Sec. 5.

For the Kingman-coalescent we have the following result, first obtained in [74] by solving a recursion,

$$p_m^{(\text{Kingman})} = \frac{m-1}{m+1}. \tag{17}$$

To see (17) without solving a recursion, we consider the process forwards in time from the MRCA. Label the two ancestral lines generated by the first split (of the MRCA) as a_1 and a_2. The fraction of the population that is a descendant of a_1 is distributed as a uniform random variable on the unit interval, see e.g. the remark after Thm. 1.2 in [8]. Therefore, with U a uniform r.v. on $[0,1]$, and any finite $m \in \mathbb{N}$,

$$p_m^{(\text{Kingman})} = 1 - \mathbb{E}\left[U^m\right] - \mathbb{E}\left[(1-U)^m\right] = 1 - 2\int_0^1 x^m dx = \frac{m-1}{m+1}. \tag{18}$$

For the Bolthausen-Sznitman coalescent (BS-coal) $\lim_{n\to\infty} p_{n,m}^{(\text{BS-coal})} = 0$ for m fixed, see Prop. 3. We remark in this context that the Bolthausen-Sznitman coalescent does not come down from infinity.

Result (17) indicates that $T_{\text{MRCA}}^{(n)}$ is a good statistic for capturing a property of the population with a small sample, at least under the Kingman coalescent. We remark that the Kingman coalescent comes down from infinity. Result (17) (and (18)) is the 'spark' for the current work.

Our main mathematical result, Thm. 1, is a representation of $p_m^{(\text{Beta-coal})}$, i.e. $p_m^{(\Pi^{(\Lambda)})}$ when $\Pi^{(\Lambda)}$ is the Beta$(2-\alpha,\alpha)$-coalescent [79] (Beta-coalescent; see Eq. (A36)). For $\alpha \in (1,2)$, the Beta-coalescent comes down from infinity. The representation of $p_m^{(\text{Beta-coal})}$ given in Thm. 1 can be directly derived from [8, Thm. 1.2], which is a result based on the connection between the Beta-coalescent and a continuous-state branching process (see Sec. 4.5 for a proof).

Theorem 1 *Define $p_m^{(\textit{Beta-coal})} \equiv p_m^{(\Pi)}$ (see Eq. (16)) when Π is the Beta-coalescent for $\alpha \in (1,2)$. Let K denote the random number of blocks involved in the merger upon which the MRCA of $[n]$ is reached; K has generating function $\mathbb{E}\left[u^K\right] = \alpha u \int_0^1 (1-x^{1-\alpha})^{-1}((1-ux)^{\alpha-1}-1)dx$ for $u \in [0,1]$ [48, Thm. 3.5]. Let $(Y_i)_{i\in\mathbb{N}}$ be a sequence of i.i.d. r.v. with Slack's distribution on $[0,\infty)$, i.e. Y_1 has Laplace transform $\mathbb{E}\left[e^{-\lambda Y_1}\right] = 1-(1+\lambda^{1-\alpha})^{-1/(\alpha-1)}$ [81]. We have the representation*

$$p_m^{(\textit{Beta-coal})} = 1 - \sum_{k\in\mathbb{N}} k\mathbb{E}\left[(Y_1+\ldots+Y_k)^{1-\alpha}\right]^{-1} \mathbb{E}\left[\frac{Y_1^m}{(Y_1+\ldots+Y_k)^{\alpha+m-1}}\right] \mathbb{P}\left(K=k\right). \tag{19}$$

We discuss the relevance of $p_m^{(\Pi)}$ for biology in Sec. 5.

We close this subsection with a consideration of the limit $\lim_{n\to\infty} p_{n,m}^{(\Pi)}$ when Π is a Ξ-coalescent (A32). We give a criterion for when $p_m^{(\Xi)} > 0$. This question is closely related to the question of coming down from infinity for a Ξ-coalescent. We have the following result (see Sec. 4.6 for a proof).

Proposition 5 *Consider any Ξ-coalescent. For any fixed $m \in \mathbb{N}$, $m \geq 2$, $p_m^{(\Xi)}$ exists. If the coalescent comes down from infinity or $\Xi(\{x \in \Delta | \sum_{i=1}^k x_i = 1 \text{ for } k \in \mathbb{N}\}) > 0$ then $p_m^{(\Xi)} > 0$. If it stays infinite then $p_m^{(\Xi)} = 0$.*

3 Relative times and lengths

In this section we use simulations to assess how well the subsample's genealogy 'covers' the genealogy of the complete sample containing the subsample. We consider the relative times $\rho_T^{(m;n)} := T_{\text{MRCA}}^{(m;n)}/T_{\text{MRCA}}^{(n)}$ and the relative lengths $\rho_I^{(m;n)} := L_{\text{int}}^{(m;n)}/L_{\text{int}}^{(n)}$ where $L_{\text{int}}^{(m;n)}$ is the sum of the lengths of the 'internal' edges associated with the subsample and $L_{\text{int}}^{(n)}$ is the sum of the lengths of internal edges of the complete sample. An edge (ancestral line) is internal if it is subtended by at least two leaves, else it is 'external'. An edge is associated with the subsample if at least one of the leaves subtending it belongs to the subsample — we call such a line a subsample line. By way of example, the continuing line of the first merger in Fig. 2D counts as an internal subsample line, although it is only subtended by a single leaf of the subsample. The ratio $\rho_I^{(m;n)}$ keeps track of the fraction of internal edges of the sample's genealogy that are covered by edges of the subsample's genealogy. The statistic $\rho_I^{(m;n)}$ indicates how much of the 'ancestral variation', or mutations present in at least 2 copies in the sample, are captured by the subsample. The ratio $\rho_T^{(m;n)}$ indicates how likely we are to capture with the subsample the ancestral variation in the complete sample. We compare $\rho_T^{(m;n)}$ and $\rho_I^{(m;n)}$ between the Beta-coalescent, a time-changed Kingman coalescent representing Wright-Fisher (or a similar) sampling with exponential population growth (see Sec. A1), and the classical Kingman-coalescent.

3.1 Simulation method

We simulate realisations of $\rho_I^{(m;n)}$ and $\rho_T^{(m;n)}$ for the classical and time-changed Kingman-coalescents and for Beta-coalescents. All processes have a Markovian jump chain and waiting times between the jumps are dependent on the current state of the jump chain, more precisely on the number of ancestral lines present. We simulate sample genealogies for a sample of size n by first generating the jump chain, i.e. choosing how many ancestral lines are merged. Given the size of a merger we draw the number of internal and external lines to merge. Given a number of sample lines we draw the waiting time until the next merger.

Let j denote the current number of lines of the complete sample. Let $M \in \{2,\ldots,n\}$ be the size of the next merger. Under a (time-changed) Kingman-n-coalescent $M = 2$ regardles of the value of $j \geq 2$. Under a Λ-n-coalescent $M = k$ with probability $\mathbb{P}(M = k) = \lambda_k(n)/\lambda(n)$ (see Eq. (A33) and (A34)) for $2 \leq k \leq n$. The k lines picked are merged into a single ancestral line, so $j-k+1$ lines are left after the merger. We draw subsequent mergers starting with n lines until there is only a single line left, when the MRCA of the sample is reached.

Consider $j = m_{\text{ext}} + m_{\text{int}} + m_{\text{ext}}^{(c)} + m_{\text{int}}^{(c)}$ sample ancestral lines present, from which m_{ext}, m_{int} are external and internal subsample lines, whereas $m_{\text{ext}}^{(c)}$ and $m_{\text{int}}^{(c)}$ are external and internal lines not subtended by leaves from the subsample. As an example, we start (before any merger) with n lines distributed as $m_{\text{ext}} = m$, $m_{\text{int}} = 0$, $m_{\text{ext}}^{(c)} = n - m$ and $m_{\text{int}}^{(c)} = 0$. All n-coalescents are exchangeable, and we always pick lines to merge at random from the lines present without replacement. This leads to drawing numbers of lines x_1, x_2, x_3, and x_4 from the four categories m_{ext}, m_{int}, $m_{\text{ext}}^{(c)}$ and $m_{\text{int}}^{(c)}$ following a multivariate hypergeometric distribution $(\underline{X} = (X_1,\ldots,X_4))$

$$\mathbb{P}(\underline{X} = \underline{x}) = \frac{\binom{m_{\text{ext}}}{x_1}\binom{m_{\text{ext}}^{(c)}}{x_2}\binom{m_{\text{int}}}{x_3}\binom{m_{\text{int}}^{(c)}}{x_4}}{\binom{m_{\text{ext}}+m_{\text{ext}}^{(c)}+m_{\text{int}}+m_{\text{int}}^{(c)}}{k}}, \quad x_1 + \cdots + x_4 = k, \tag{20}$$

where k denotes the given merger size. All lines are merged into a single ancestral line, which is a subsample line if and only if at least one subsample line was picked in the merger ($x_1 + x_3 \geq 1$), so the numbers of lines belonging to the four categories change from before to after the merger as

$$\begin{aligned} m_{\text{ext}} &\to m_{\text{ext}} - x_1, \\ m_{\text{ext}}^{(c)} &\to m_{\text{ext}}^{(c)} - x_2, \\ m_{\text{int}} &\to m_{\text{int}} - x_3 + \mathbb{1}_{(x_1+x_3\geq 1)}, \\ m_{\text{int}}^{(c)} &\to m_{\text{int}}^{(c)} - x_4 + \mathbb{1}_{(x_2+x_4=k)}. \end{aligned} \tag{21}$$

The transitions shown in Eq. (21) reflect our assumption that if at least 1 subsample line is involved in a given merger, the continuing ancestral line is considered to belong to the subsample; mutations that arise on the continuing line will then be carried by the subsample, and visible in the subsample unless all the subsample lines were involved in the merger ($x_1 = m_{\text{ext}}$ and $x_3 = m_{\text{int}}$ and $x_1 + x_3 \geq 1$). Therefore, if a single external subsample line, and no other subsample line, is involved in a merger ($x_1 = 1$, $x_3 = 0$) we regard the continuing line as an 'internal' line of the subsample. An external line of the subsample therefore remains so only until it is involved in a merger. By way of example, the continuing line of the first merger in Fig. 2D counts as an internal line of the subsample. The continuing line of the first merger in 2B is <u>not</u> a subsample line since the MRCA of the subsample leaves is reached in the first merger; the continuing line of the first merger in 2B counts as an internal line of the complete sample.

Denote by T_j the random waiting time for the first merger of the j-coalescent. The coalescent process under exponential population growth is a time-changed

Kingman-coalescent (see e.g. [25,30]). [41] give a way of sampling T_j under exponential growth. Let $\beta > 0$ denote the growth rate under exponential growth. Write $S_j = T_n + \cdots + T_j$ for $2 \le j \le n$, with $S_{n+1} = 0$ a.s. If $\{U_j : 2 \le j \le n\}$ denotes a collection of i.i.d. uniform $(0,1]$ random variables, then [41]

$$S_j = T_j + S_{j+1} = \frac{1}{\beta} \log\left(\exp(\beta S_{j+1}) - \tfrac{2\beta}{j(j-1)} \log(U_j)\right), \quad 2 \le j \le n. \tag{22}$$

Eq. (22) tells us that if β is very large, the time intervals T_j near the MRCA become quite small. The time intervals near the leaves are much less affected. We choose the grid of values for β as

$$\beta \in \{0.1, 0.5, 1, 10, 50, 100, 500, 1000, 5000, 10000\}.$$

Recall in this context the growth model $N_k = N_0(1+\beta/N_0)^k$ for the population size in generation $k \ge 0$ going forward in time, and where N_0 is the population size at the start of the growth $(k = 0)$. Our choice of grid values for β should reflect the range of growth from weak $(\beta = 0.1)$ to very strong $(\beta = 10^4)$ and most estimates of β obtained for natural populations should fall within this range.

Under the Beta-coalescent without growth T_j is an exponential with rate $\lambda(j) = \lambda_2(j) + \cdots + \lambda_j(j)$ where $\lambda_i(j)$ is given in Eq. (A36).

A realisation of $\rho_I^{(m;n)}$ is obtained as follows. Given $j = m_{\text{ext}} + m_{\text{int}} + m_{\text{ext}}^{(c)} + m_{\text{int}}^{(c)}$ current sample lines, let t_j denote a realisation of T_j, the random time during which there are j lines of the complete sample. We update the total lengths $\ell_{\text{int}}^{(m;n)}$ of internal subsample lines, and $\ell_{\text{int}}^{(n)}$ of internal lines of the complete sample, as

$$\begin{aligned} \ell_{\text{int}}^{(m;n)} &\to \ell_{\text{int}}^{(m;n)} + \mathbb{1}_{(m_{\text{ext}}+m_{\text{int}}>1)} m_{\text{int}} t_j, \\ \ell_{\text{int}}^{(n)} &\to \ell_{\text{int}}^{(n)} + \mathbb{1}_{(j>1)} \left(m_{\text{int}} + m_{\text{int}}^{(c)}\right) t_j. \end{aligned} \tag{23}$$

The updating rule for $\ell_{\text{int}}^{(m;n)}$ in Eq. (23) reflects the fact that mutations on the common ancestor line of the subsample, for example the continuing line after the merger of all 3 subsample lines in Fig. 2D, are not visible in the subsample. The updating rule for $\ell_{\text{int}}^{(n)}$ in Eq. (23) similarly reflects the fact that mutations on the continuing line of the MRCA of the complete sample are not visible in the sample; but once the MRCA of the complete sample is reached we stop the process.

A realisation of $\rho_I^{(m;n)}$ is then recorded as $r_I^{(m;n)} := \ell_{\text{int}}^{(m;n)} / \ell_{\text{int}}^{(n)}$. By way of example, the edges marked with a black dot in Fig. 2B are internal edges of the complete sample while the edges marked with a circle in Fig. 2A are internal edges associated with the subsample as well as the complete sample, and we have $\rho_I^{(m;n)} = (T_5 + T_4 + T_3)/(T_5 + 2T_4 + T_3)$ for the genealogy in Fig. 2A. There are no internal edges associated with the subsample in Fig. 2B and 2C; therefore $\rho_I^{(m;n)} = 0$ for the genealogies in Fig. 2B and 2C. The sample and the subsample share all the internal edges in the genealogy shown in Fig. 2D and therefore $\rho_I^{(m;n)} = 1$.

Realisations of $T^{(m;n)}_{\text{MRCA}}$ $(t^{(m;n)})$ and $T^{(n)}_{\text{MRCA}}$ $(t^{(n)})$ are recorded as

$$\begin{aligned} t^{(m;n)} &= \inf\{t \geq 0 : m_{\text{ext}} + m_{\text{int}} = 1\}, \\ t^{(n)} &= \inf\{t \geq 0 : m_{\text{ext}} + m_{\text{int}} + m^{(c)}_{\text{ext}} + m^{(c)}_{\text{int}} = 1\}, \end{aligned} \tag{24}$$

by adding up the realised waiting times t_j of T_j. We record a realisation of $\rho_T^{(m;n)}$ as $r_T^{(m;n)} := t^{(m;n)}/t^{(n)}$.

3.2 Simulation results

Figures 3 and 4 show estimates, in the form of violin plots[44], of the distributions of $\rho_T^{(m;n)}$ (left column) and $\rho_I^{(m;n)}$ (right column); under the Beta-coalescent as a function of α (Figure 3) and under exponential growth as a function of β (Figure 4). We see that under exponential growth the distribution of $\rho_T^{(m;n)}$ can be rather concentrated (recall Eq. (22)). The estimates shown in Fig. 4 of the distribution of $\rho_T^{(m;n)}$ indicate that $\rho_T^{(m;n)}$ becomes more concentrated at 1 as β increases. Recall in this context that $p_{n,m}^{(\text{exp. growth})} = (m-1)(n+1)/((m+1)(n-1))$ since exp. growth results in a time-changed Kingman-coalescent.

In Figure 3 we see a gradual shift in the distribution of $\rho_T^{(m;n)}$ as subsample size increases; from being skewed to the right (ie. towards higher values) to being skewed to the left (ie. towards smaller values). This is in sharp contrast to the distribution under exponential growth (Figure 4) where the distribution of $\rho_T^{(m;n)}$ is always skewed to the left. This indicates that under a multiple-merger coalescent process a subsample is much less informative about the complete sample than under exponential growth. In contrast, under exponential growth, even a small subsample can be very informative about the complete sample, especially in a strongly growing (large β) population. Estimates of the means $\mathbb{E}^{(\Pi)}\left[\rho_T^{(m;n)}\right]$, shown in Figure 5 (circles) for the Beta-coalescent, and in Figure 6 (circles) for exponential growth, further strengthen our conclusion.

The distribution of $\rho_I^{(m;n)}$, the relative lengths of internal edges, also behaves differently between the Beta-coalescent and exponential growth. The distribution of $\rho_I^{(m;n)}$ becomes more concentrated around smaller values as growth becomes stronger (β increases) while it stays highly variable as α tends to 1, although the median decreases as skewness increases (α tends to 1). This indicates that we capture less and less of the 'ancestral variation' (mutations observed in at least 2 copies in the sample) in the complete sample as growth or skewness increase. Estimates of $\mathbb{E}^{(\Pi)}\left[\rho_I^{(m;n)}\right]$ (Figures 5 and 6, '+') also indicate that one would need a large sample to capture at least half of the ancestral variation if growth or skewness is high.

To conclude, $\rho_T^{(m;n)}$ and $\rho_I^{(m;n)}$ seem to tend to opposite values under exponential growth; $\rho_T^{(m;n)}$ to 1 and $\rho_I^{(m;n)}$ to small values, as β increases. Thus, even if we are sharing the MRCA with higher probability as β increases (recall that the samples are nested), we are capturing less and less of the ancestral variation.

Essentially the opposite trend is seen for both $\rho_T^{(m;n)}$ and $\rho_I^{(m;n)}$ under the Beta-coalescent; the distributions of both statistics stay highly variable as $\alpha \to 1$.

Fig. 2 Examples of genealogies. Thick edges denote lineages ancestral to the subsample of size $m = 3$; sample size $n = 7$. The marked edges in **A** denote internal ancestral lineages to both the subsample and the complete sample; the marked edges in **B** denote lineages internal only to the complete sample. In **C** the complete sample and subsample share the MRCA without sharing any of the internal edges. The genealogies are shown from the time of sampling (present) until the MRCA of the complete sample is reached (past). In **C** the complete sample and the subsample share the MRCA without sharing any internal edges. In **D** the complete sample and the subsample share the MRCA and all the internal edges.

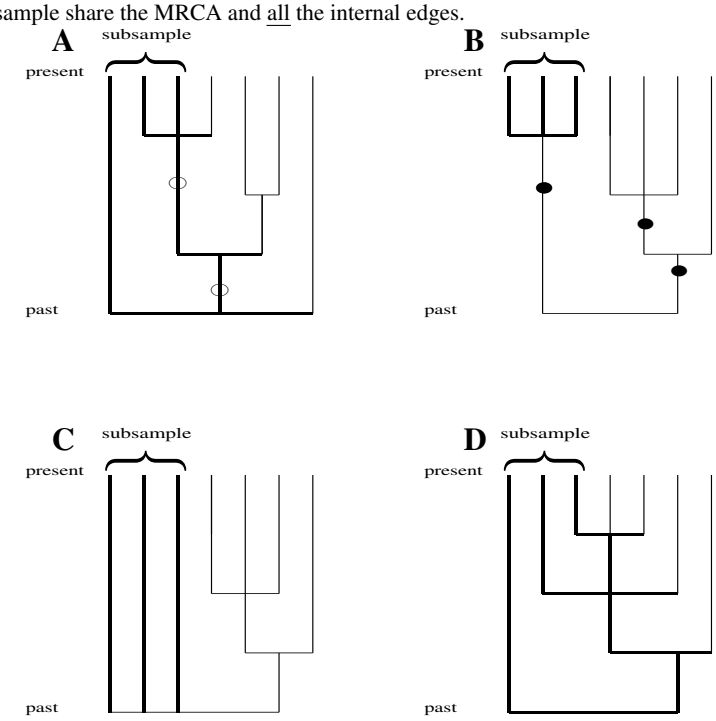


Fig. 3 Estimates, shown in the form of violin plots [44] of the distributions of $\rho_T^{(m;n)}$ and of $\rho_I^{(m;n)}$ as functions of the coalescent parameter α of the Beta$(2-\alpha,\alpha)$-coalescent for values of sample size $n = 10^4$ and subsample size m as shown. The coalescent process at $\alpha = 2$ is the Kingman-coalescent. For explanation of symbols see Subsection 3.1. Shown are results from 10^5 replicates.

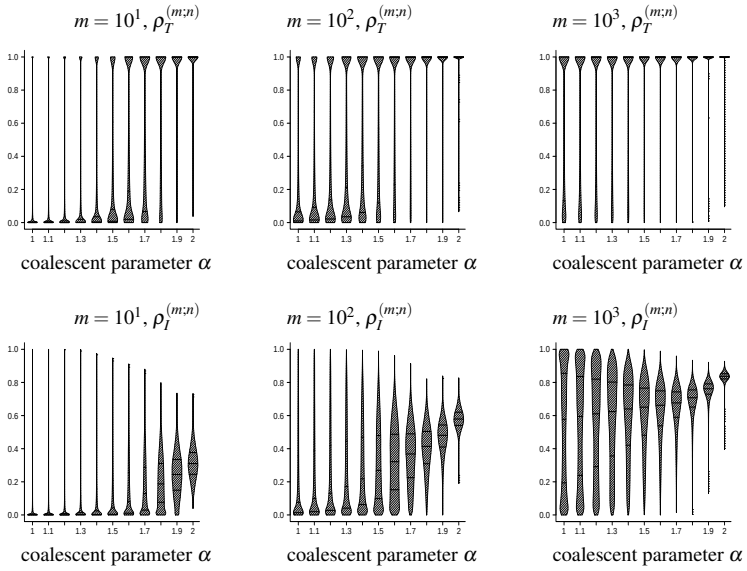


Fig. 4 Estimates, shown in the form of violin plots [44] of the distributions of $\rho_T^{(m;n)}$ and of $\rho_I^{(m;n)}$ as functions of the exponential growth parameter β for values of sample size $n = 10^4$ and subsample size m as shown. The coalescent process at $\beta = 0$ is the Kingman-coalescent. For explanation of symbols see Subsection 3.1. The grid of values of β is $\{0.1, 0.5, 1.0, 10.0, 50.0, 100.0, 500.0, 1000.0, 5000.0, 10000.0\}$. Shown are results from 10^5 replicates.

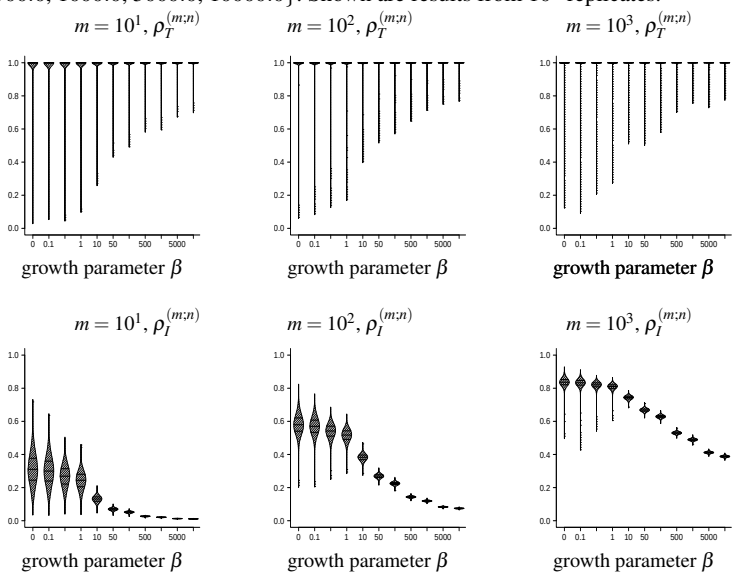


Fig. 5 Estimates of $\mathbb{E}\left[\rho_T^{(m;n)}\right]$ ($\circ$), and of $\mathbb{E}\left[\rho_I^{(m;n)}\right]$ ($+$) as functions of the coalescent parameter α for values of sample size $n = 10^4$ and subsample size $m = 10^1$ (solid lines); $m = 10^2$ (dashed lines); $m = 10^3$ (dotted lines). The coalescent process at $\alpha = 2$ is the Kingman-coalescent. For explanation of symbols see Subsection 3.1. Shown are results from 10^5 replicates.

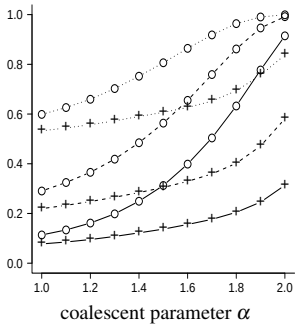
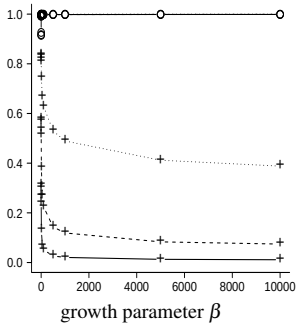


Fig. 6 Estimates of $\mathbb{E}\left[\rho_T^{(m;n)}\right]$ ($\circ$) and of $\mathbb{E}\left[\rho_I^{(m;n)}\right]$ ($+$) as functions of the exponential growth parameter β for values of sample size $n = 10^4$ and subsample size $m = 10^1$ (solid lines); $m = 10^2$ (dashed lines); $m = 10^3$ (dotted lines). The coalescent process at $\beta = 0$ is the Kingman-coalescent. For explanation of symbols see Subsection 3.1. The grid of values of β is $\{0.1, 0.5, 1.0, 10.0, 50.0, 100.0, 500.0, 1000.0, 5000.0, 10000.0\}$. Shown are results from 10^5 replicates.



4 Proofs

4.1 Proof of Prop. 1

Proof Let $\Lambda = \delta_p$ for $p \in (0,1)$ which fulfills $\Lambda(\{0\}) = 0$. Clearly

$$p_{m,n}^{(\Pi)} > \mathbb{P}\left(\text{tree is star-shaped}\right).$$

The probability that the associated Λ-n-coalescent is star-shaped, i.e. all blocks merge at the first (and then only) collision, is

$$\frac{p^{n-2}}{p^{-2}(1-(1-p)^n - np(1-p)^{n-1})} = \frac{p^n}{\sum_{i=2}^n \binom{n}{i} p^i (1-p)^{n-i}} > p^n.$$

For any star-shaped path of a n-coalescent, we have $T_{\text{MRCA}}^{(m;n)} = T_{\text{MRCA}}^{(n)}$ for any $m < n$. Thus, we can choose Λ' s.t. $\Lambda' = \delta_p$ with

$$p = \left(\mathbb{P}^{(\delta_0)}\left(T_{\text{MRCA}}^{(m;n)} = T_{\text{MRCA}}^{(n)}\right)\right)^{\frac{1}{n}}.$$

□

4.2 Proof of Prop. 2

Proof Assume Π is a Ξ-coalescent. The event $\left\{T_{\text{MRCA}}^{(m;n)} = T_{\text{MRCA}}^{(n)}\right\}$ is the complement of the event

$$A_{m,n} := \left\{[m] \subseteq \pi_1,\, \pi_1 \text{ is a block of } \Pi^{(n)}_{T_{\text{MRCA}}^{(n)}-}\right\}. \tag{25}$$

Due to the exchangeability of the Ξ-coalescent, $\Pi^{(n)}_{T_{\text{MRCA}}^{(n)}-}$ is an exchangeable partition of $[n]$. Given the (ordered) block sizes $\left(B_{[i]}^{(n)}\right)_{i\in\mathbb{N}}$, the probability that a given block of size $B_{[i]}^{(n)}$ contains $[m]$ is given by drawing without replacement, i.e.

$$\mathbb{P}\left([m] \subseteq \text{ a given block of size } B_{[i]}^{(n)} \,|\, B_{[i]}^{(n)}\right) = \prod_{\ell=0}^{m-1} \frac{B_{[i]}^{(n)} - \ell}{n-\ell}.$$

Summing this up over all blocks and taking the expectation yields $\mathbb{P}(A_{m,n}) = 1 - p_{n,m}^{(\Xi)}$, thus establishing Eq. (7) (by definition there is more than one block at time $T_{\text{MRCA}}^{(n)}-$, i.e. at time infinitely close to $T_{\text{MRCA}}^{(n)}$, so $p_{n,m}^{(\Xi)} > 0$.)

To show the convergence in Eq. (8) we first establish that all objects are well defined. Assume now that the Ξ-coalescent comes down from infinity, so at any time $t > 0$, there are only finitely many blocks in the partition Π_t almost surely. For $n \to \infty$, Kingman's correspondence [56, Thm. 2] ensures that the asymptotic frequencies of the blocks in the partition Π_t of $\mathbb{N}$ exist almost surely and are limits

of the block frequencies in the n-coalescent as written in the proposition. Pick an arbitrarily small $t > 0$. Then, consider only paths where Π_t has more than one block. Since the number of blocks of Π_t is finite a.s. we can find $n_0 \in \mathbb{N}$ so that $\Pi_t^{(n_0)}$ has at least one individual in any block of Π_t (thus has the same number of blocks). By construction, from time t onwards, the Ξ-n_0-coalescent merges the blocks in exactly the same (Markovian) manner as the Ξ-coalescent. So if Π_t has more than one block, $T_{\mathrm{MRCA}}^{(n)} = T_{\mathrm{MRCA}}^{(\infty)}$ for $n \geq n_0$ and the asymptotic frequencies at $T_{\mathrm{MRCA}}^{(\infty)}-$ exist (since their corresponding blocks are a specific merger of the blocks of Π_t whose block frequencies exist). Now $T_{\mathrm{MRCA}}^{(2)} \leq T_{\mathrm{MRCA}}^{(\infty)}$ almost surely and $T_{\mathrm{MRCA}}^{(2)}$ is $\mathrm{Exp}(\Xi(\Delta))$-distributed. Therefore, for almost every path, we can choose $t < T_{\mathrm{MRCA}}^{(2)}$ so that Π_t has more than one block.

We have established that all objects are well defined; now we show the actual convergence in (8). For $\boldsymbol{x} \in \Delta$ let

$$f_{n,m}(\boldsymbol{x}) := \sum_{i\in\mathbb{N}} \prod_{\ell=0}^{m-1} \frac{nx_i - \ell}{n-\ell}$$

and $f_m(\boldsymbol{x}) := \sum_{i\in\mathbb{N}} x_i^m$. We have $f_{n,m} \to f_m$ uniformly on Δ and that f_m is continuous on Δ in the ℓ^1-norm with $0 \leq f_m \leq 1$. We can rewrite, using Eq. (7),

$$p_{n,m}^{(\Xi)} = 1 - \mathbb{E}\left[f_{n,m}\left((\tfrac{1}{n}B_{[i]}^{(n)})_{i\in\mathbb{N}}\right)\right].$$

For any $\varepsilon > 0$ we find n_0 so that for $n \geq n_0$

$$\begin{aligned}
&|\mathbb{E}\left[f_{n,m}\left((\tfrac{1}{n}B_{[i]})_{i\in\mathbb{N}}\right)\right] - \mathbb{E}\left[f_m\left(P_{[i]}\right)_{i\in\mathbb{N}}\right]| \\
\leq &|\mathbb{E}\left[f_{n,m}\left((\tfrac{1}{n}B_{[i]})_{i\in\mathbb{N}}\right)\right] - \mathbb{E}\left[f_m\left((\tfrac{1}{n}B_{[i]})_{i\in\mathbb{N}}\right)\right]| \\
&+ |\mathbb{E}\left[f_m\left((\tfrac{1}{n}B_{[i]})_{i\in\mathbb{N}}\right)\right] - \mathbb{E}\left[f_m\left((P_{[i]})_{i\in\mathbb{N}}\right)\right]| \leq 2\varepsilon.
\end{aligned}$$

We have used uniform convergence of $f_{n,m}$ to f_m to control the first difference and the convergence (in law) of $(n^{-1}B_{[i]}^{(n)})_{i\in\mathbb{N}}$ to $(P_{[i]})_{i\in\mathbb{N}}$ to control the second.

The representation of the limit in Eq. (8) in terms of X and Y follows directly from the properties of exchangeable partitions (c.f. for example [9,10]). The first equality is [9, Eq. (1.4)], while the second equality uses the correspondence between the distribution of a size-biased and a uniform pick of a block, see [9, Eq. (1.2)]. By definition $\Pi_{T_{\mathrm{MRCA}}^{(\infty)}-}$ has more than one block almost surely so the limit in Eq. (8) is > 0.

□

Remark 1 Reordering the block frequencies, e.g. in order of least elements of blocks, does not change Eq. (8).

4.3 Proof of Prop. 3

Proof We use the construction of [39] in which the Bolthausen-Sznitman coalescent is obtained by cutting a random recursive tree $\mathbb{T}_n$ with n nodes at independent $\mathrm{Exp}(1)$ times, see Sec. A3. Consider the last merger in the Bolthausen-Sznitman n-coalescent. In terms of cutting edges of $\mathbb{T}_n$, the last merger is reached when the last edge connected to the root of $\mathbb{T}_n$ is cut. Let E_n be the number of such edges in $\mathbb{T}_n$. For $T^{(m;n)}_{\mathrm{MRCA}} = T^{(n)}_{\mathrm{MRCA}}$, we need that not all $i \in [m]$ are in a single block of the n-coalescent before the last merger (see proof of Prop. 2).

By construction, for any node with label in $[m]$, on the path to the node labelled 1 (root) in the uncut tree $\mathbb{T}_n$, the last node passed before reaching the root must also have a label from $[m]$. Thus, any node connected to the root of $\mathbb{T}_n$ that is labelled from $[n]_{m+1}$ cannot root a subtree that includes any nodes labelled from $[m]$.
Now, we consider the last edge of $\mathbb{T}_n$ cut in the construction of the Bolthausen-Sznitman n-coalescent, which causes the MRCA of the n-coalescent to be reached. It has to be connected to the root. Consider the two subtrees on both sides of the edge cut last. One subtree contains the root, thus includes at least the label 1 from $[m]$. If the other subtree is rooted in a node labelled from $[m]$, we have $T^{(m;n)}_{\mathrm{MRCA}} = T^{(n)}_{\mathrm{MRCA}}$, since both subtrees contain labels of $[m]$, thus not all $i \in [m]$ are in a single block of the n-coalescent before the last merger. If the subtree not containing the root has a root labelled from $[n]_{m+1}$, as argued above, it contains no labels from $[m]$. Additionally, since we are at the last cut, all other edges connected to the root of $\mathbb{T}_n$ have already been cut and all labels in the subtrees rooted by them joined with label 1. Thus, all labels in $[m]$ are labelling the root before the last cut, which corresponds to $[m]$ being a subset of a block of the n-coalescent before the last merger, hence $T^{(m;n)}_{\mathrm{MRCA}} \neq T^{(n)}_{\mathrm{MRCA}}$.
This shows $T^{(m;n)}_{\mathrm{MRCA}} = T^{(n)}_{\mathrm{MRCA}}$ if and only if the last edge cut is an edge connecting a node labelled from $[m]$ with the root. Let E_m be the count of edges of $\mathbb{T}_n$ connected to the root labelled from $[m]$ and E_n be the total count of edges connected to the root. Then,

$$\mathbb{P}\left(T^{(m;n)}_{\mathrm{MRCA}} = T^{(n)}_{\mathrm{MRCA}}\right) = \mathbb{E}\left[\frac{E_m}{E_n}\right], \tag{26}$$

because given $\mathbb{T}_n$, E_m/E_n is the probability that the edge cut last is connected to a node with a label from $[m]$; edges are cut at i.i.d. times, so the edge cut last is uniformly distributed among all edges connected to the root.
As we see from the sequential construction of $\mathbb{T}_n$, E_m is the number of edges connected to 1 when the first m nodes are set, the resulting tree is a random recursive tree $\mathbb{T}_m$ with n leaves. The numbers E_n and E_m can be described in terms of a Chinese restaurant process (CRP), see [39, p. 724]: The number of edges connected to node 1 is distributed as the number of tables in a CRP with n (resp. m) customers. This distribution is $E_i \overset{d}{=} B_1 + \ldots + B_i$ ($i \in \{m,n\}$), where $B_1, \ldots, B_i$ are independent Bernoulli variables with $P(B_j = 1) = j^{-1}$, see e.g [3, p. 10]. The sequential construction of the random recursive trees (and the connected CRPs) ensures that the $B_1, \ldots, B_m$ are identical for E_m and E_n. This establishes the equality of Equations (26) and (9).

From the proof of [38, Lemma 3], we have $\frac{\log(n)}{E_n} \to 1$ in L^1 for $n \to \infty$. The sequence $(E_m/E_n)_{n\in\mathbb{N}}$ is bounded a.s. Thus, bounded convergence ensures

$$\lim_{n\to\infty} \mathbb{E}\left[\log(n)\frac{E_m}{E_n}\right] = \mathbb{E}\left[\lim_{n\to\infty} \frac{\log(n)E_m}{E_n}\right] = \mathbb{E}[E_m] = 1 + \tfrac{1}{2} + \cdots + \tfrac{1}{m-1}.$$

□

4.4 Proof of Prop. 4

Proof As in Subsection 3 we use the construction of the Bolthausen-Sznitman n-coalescent described in [39]. We wish to establish the probability that the MRCA of a subsample of size m from a sample of size n is an ancestor of only the subsample in the n-coalescent. We will also use the Bernoulli variables B_i, $i \in [n]$ of $\mathbb{T}_n$ as in Prop. 3, where $B_i = 1$ if the node labelled i is directly connected to the root (node labelled 1). If we look at the cutting procedure which constructs the Bolthausen-Sznitman n-coalescent from $\mathbb{T}_n$, we observe that no path of $\mathbb{T}_n$ can contribute positive probability to $q_{n,m}^{(\text{BS-coal})}$ that attaches any node labelled from $[n]_{m+1}$ to a node labelled from $[m]_2$. If we do attach a node labelled $i \in [n]_{m+1}$ to a node labelled from $[m]_2$, when constructing the Bolthausen-Sznitman n-coalescent we will cut an edge on the path from the node labelled i to the root labelled 1 before the MRCA of $[m]$ is reached, thus i would subtend the MRCA of $[m]$. The probability that a node labelled $i \in [n]_{m+1}$ is not connected to a node labelled from $[m]_2$ in $\mathbb{T}_n$ is

$$\prod_{i=1}^{n-m} \frac{i}{m+i-1} = \binom{n-1}{m-1}^{-1}.$$

Even when there is no edge connecting a node labelled from $[n]_{m+1}$ directly with a node labelled from $[m]_2$, not all such paths of $\mathbb{T}_n$ will contribute to $q_{n,m}^{(\text{BS-coal})}$. To contribute, we need that the cutting procedure does not lead to any $i \in [n]_{m+1}$ being subtended by the MRCA of $[m]$. For the mentioned paths, this happens if and only if we cut all edges connecting nodes labelled from $[m]_2$ to 1 before cutting any edge connecting 1 to nodes labelled from $[n]_{m+1}$. We have $\sum_{i\in[m-1]} B_i$ edges adjacent to node 1, see the proof of Prop. 3. With the constraint that no edge connects a node labelled from $[n]_{m+1}$ directly with a node labelled from $[m]_2$, the sequential construction yields that, after relabelling, the nodes labelled with $\{1\} \cup [n]_{m+1}$ form a $\mathbb{T}_{n-m+1}$ and thus there are $\sum_{i\in[n-m]} B_i'$ edges adjacent to the root of $\mathbb{T}_n$ connecting to the nodes labelled with $\{1\} \cup [n]_{m+1}$, where $B_i' \stackrel{d}{=} B_i$ for independent B_i'. All edges adjacent to the node labelled 1 need to be cut before the MRCA of $[n]$ is reached and they are cut at independent $\text{Exp}(1)$ times. This means that the probability of cutting all edges connecting 1 to nodes labelled from $[m]_2$ first is just drawing $\sum_{i\in[m-1]} B_i$ times without replacement from $\sum_{i\in[m-1]} B_i + \sum_{i\in[n-m]} B_i'$ edges, where all $\sum_{i\in[m-1]} B_i$ edges connecting nodes labelled from to $[m]_2$ have to be drawn. This probability equals

$$\binom{\sum_{i\in[m-1]} B_i + \sum_{i\in[n-m]} B_i'}{\sum_{i\in[m-1]} B_i}^{-1}.$$

Integrating over all contributing paths of $\mathbb{T}_n$ with the cutting constraint described above finishes the proof.

□

4.5 Proof of Thm. 1

Proof We track the asymptotic frequencies $\left(P_{[i]}(t)\right)_{t\geq 0}$ of the ith biggest block for all $t>0$ and $i\in\mathbb{N}$. Consider a non-negative and measurable $[0,\infty)$-valued function g on the k-dimensional simplex

$$\Delta_k := \{(x_1,\ldots,x_k) : x_1 \geq x_2 \geq \ldots \geq x_k \geq 0, \sum_{i\in[k]} x_i = 1\}$$

that is invariant under permutations $(x_1,\ldots,x_k)\mapsto(x_{\sigma(1)},\ldots,x_{\sigma(k)})$. [8, Thm. 1.2] shows that

$$\begin{aligned}
&\mathbb{E}\left[g\left((P_{[i]}(T_k))_{i\in\mathbb{N}}\right)|N(T_k)=k\right]\\
=&\,\mathbb{E}\left[(Y_1+\ldots+Y_k)^{1-\alpha}\right]^{-1}\mathbb{E}\left[(Y_1+\ldots+Y_k)^{1-\alpha}g\left(\left(\frac{Y_i}{Y_1+\ldots+Y_k}\right)_{i\in[k]}\right)\right],
\end{aligned}$$

where T_k is the waiting time until a state with $\leq k$ blocks is hit by the Beta-coalescent and $N(t)$ is the number of blocks of Π_t, thus we condition on the coalescent to hit a state with exactly k blocks.
We can apply this formula to compute $\mathbb{E}\left[\sum_{i\in[K]}P_i^m\right]$ from Eq. (8), where K is the number of blocks at the last collision of the Beta-coalescent. For this, condition on $K=k$. With $\{K=k\}=\{N(T_k)=k\}\cap\{\text{all blocks of } \Pi_{T_k} \text{ merge at the next merger}\}$, the strong Markov property shows that the block frequencies at T_k are independent of them merging at the next collision. However, these frequencies are, conditioned on K, just $(P_i)_{i\in[K]}$. For $\underline{x}\in\Delta_k$ we set $g_m(\underline{x})=\sum_{i=1}^k x_i^m$ (which fulfills all necessary conditions to apply [8, Thm. 1.2]) and compute

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i\in[K]}P_i^m\right]\\
=&\sum_{k\in\mathbb{N}}\mathbb{E}\left[\sum_{i\in[K]}P_i^m\,|\,K=k\right]\mathbb{P}(K=k)\\
=&\sum_{k\in\mathbb{N}}\mathbb{E}\left[g_m\left((P_i(T_k))_{i\in\mathbb{N}}\right)|N(T_k)=k\right]\mathbb{P}(K=k)\\
=&\sum_{k\in\mathbb{N}}\mathbb{E}\left[(Y_1+\ldots+Y_k)^{1-\alpha}\right]^{-1}\mathbb{E}\left[\sum_{i=1}^k\frac{Y_i^m}{(Y_1+\ldots+Y_k)^{\alpha+m-1}}\right]\mathbb{P}(K=k).
\end{aligned}$$

The distribution of K for the Beta-coalescent is known from [48, Thm. 3.5]. Using that

$$\left(\frac{Y_i^m}{(Y_1+\ldots+Y_k)^{\alpha+m-1}}\right)_{i\in[k]}$$

are identically distributed and $p_m^{(\text{Beta-coal})}=1-\mathbb{E}\left[\sum_{i\in[K]}P_i^m\right]$ completes the proof.

□

4.6 Proof of Prop. 5

Proof Consider any Ξ-coalescent (and its restrictions to $[n]$, $n \in \mathbb{N}$). Since, for nested samples, $T_{\text{MRCA}}^{(m;n)} \leq T_{\text{MRCA}}^{(n)}$ almost surely for any $n \geq m$, we have

$$\left\{T_{\text{MRCA}}^{(m;m+i)} = T_{\text{MRCA}}^{(m+i)}\right\} \supseteq \left\{T_{\text{MRCA}}^{(m;m+i+1)} = T_{\text{MRCA}}^{(m+i+1)}\right\}$$

for any $i \in \mathbb{N}$. Thus, $p_m^{(\Xi)} = \lim_{n\to\infty} p_{n,m}^{(\Xi)} = \mathbb{P}^{(\Xi)}\left(T_{\text{MRCA}}^{(m;n)} = T_{\text{MRCA}}^{(n)} \ \forall\, n > m\right)$ exists. Suppose first that the Ξ-coalescent comes down from infinity. Then, Eq. (8) shows $p_m^{(\Xi)} > 0$. If the Ξ-coalescent stays infinite, τ_m is almost surely finite, while $T_{\text{MRCA}}^{(n)} \to \infty$ almost surely. Thus, $p_m^{(\Xi)} = 0$.

Consider a Ξ-coalescent that neither comes down from infinity nor stays infinite. Then, $\Xi(\{\boldsymbol{x} \in \Delta | \sum_{i=1}^k x_i = 1 \text{ for } k \in \mathbb{N}\}) > 0$. As stated in the introduction, in this case there is an almost surely finite waiting time T with $\#\Pi_T < \infty$ almost surely. Let n_T be the finite number of blocks at time T. Again, exchangeability ensures, as in proving Eq. (7), that there is a positive probability that not all $i \in [m]$ are in the same block of Π_T (so in particular, with positive probability, $T_{\text{MRCA}}^{(m)} > T$). The strong Markov property of the Ξ-coalescent ensures that, given n_T, Π_T evolves like a Ξ-n_T-coalescent, which can have at most n_T mergers. In summary, with positive probability, more than one of the n_T blocks at time T includes individuals from the subset $[m]$ and the n_T blocks are merged following a Ξ-n_T-coalescent. Then, Eq. (7) shows that with positive probability, conditioned on the event that $k > 1$ blocks of Π_T contain individuals from $[m]$, also more than one block of the Ξ-coalescent at its last collision contains individuals of $[m]$.

□

Remark 2 Prop. 5 shows that $\mathbb{P}^{(\Xi)}\left(T_{\text{MRCA}}^{(m;n)} = T_{\text{MRCA}}^{(n)}\right) \to 0$ for fixed m and $n \to \infty$ if the Ξ-coalescent stays infinite. The Bolthausen-Sznitman coalescent stays infinite [78, Example 15]; however, convergence to 0 is only of order $\mathscr{O}(1/\log(n))$.

4.7 Proof of recursions (4), (13), and (11)

The strong Markov property of a Λ-coalescent together with a natural coupling which we will introduce below allows us to describe many functionals of multiple-merger n-coalescents recursively by conditioning on their first jump, e.g. see [40] or [62]. We use this to prove recursions (4) and (11).

4.7.1 Proof of Eq. (4)

Consider the probability $p_{n,m}^{(\Lambda)}$ (see Eq. (4)) that a sample of size n shares the MRCA with a subsample of size $m \in [n-1]_2$. The boundary conditions $p_{m,m} = 1$ and $p_{n,1} = 0$ for $n > 1$ follow directly from the definition. We record how many individuals are merged at the first jump of the n-coalescent. Suppose a k-merger occurs which happens with probability $\beta(n, n-k+1)$. Conditional on a k-merger,

$\ell \leq m$ of individuals that merge are taken from the subsample and $n-\ell$ are not with probability $\binom{m}{\ell}\binom{n-m}{k-\ell}/\binom{n}{k}$, since the individuals that merge are picked uniformly at random without replacement. For $p^{(\Lambda)}_{n,m} > 0$, we need that not all m individuals are merged unless all n individuals are merged, thus $\ell < m$ or $k = n$. Writing $C(k,\ell)$ for the event that exactly ℓ lineages from the subsample are merged (with $\ell < m$ or $k = n$), the strong Markov property shows that

$$\mathbb{P}^{(\Pi^{(\Lambda)})}\left(T^{(m;n)}_{\mathrm{MRCA}} = T^{(n)}_{\mathrm{MRCA}} \,|\, C(k,\ell)\right) = \mathbb{P}^{(\Pi^{\Lambda})}\left(T^{(m';n-k+1)}_{\mathrm{MRCA}} = T^{(n-k+1)}_{\mathrm{MRCA}}\right)$$

with $m' = (m-\ell+1)\mathbb{1}_{(\ell>1)} + m\mathbb{1}_{(\ell\leq 1)}$, since among the ancestral lines (blocks) after the first collision, m' are subtended by the subsample. Summing over all possible values (recall the boundary conditions) yields recursion (4). □

4.7.2 Proof of Eq. (13)

We again condition on the event that k blocks are merged at the first jump. Only k-mergers where either all merged individuals are picked from the subsample $[m]$ or none is sampled from $[m]$ contribute positive probability to $q^{(\Lambda)}_{n,m}$. After the jump, we thus have $n-k+1$ ancestral lineages present, from which either $m-k+1$ or m are connected to the subsample. The strong Markov property and sampling without replacement for the k-merger then yields Eq. (13).

4.7.3 Proof of Eq. (11)

Recall the natural coupling: if we restrict an n-coalescent with mutation rate θ to any ℓ-sized subset $L \subseteq [n]$, the restriction is an ℓ-coalescent with mutation with the same rate θ. To prove recursion (11) we partition over three possible outcomes of the first event: it is a mutation on a lineage subtending the subsample (E_1), it is a mutation on a lineage not subtending the subsample (E_2), or it is a merger (E_3). Naturally, before any mutation occurs, all edges are active.

We recall a few elementary facts. The time to the first mutation on any lineage is $\mathrm{Exp}(\theta/2)$-distributed (mutations on different/disjoint lineages are independent) and independent of the waiting time for the first merger. The minimum of independent exponential r.v.'s $X_1,\ldots,X_i$ with parameters $\alpha_1,\ldots,\alpha_i$ is again exponentially distributed with parameter $\sum_{j=1}^{i}\alpha_j$. Finally, $\mathbb{P}(X_1 \leq X_2) = \alpha_1/(\alpha_1+\alpha_2)$.

The waiting times X_i for events E_i for $1 \leq i \leq 3$ are all exponential; the one for E_1 with rate $\theta m/2$, for E_2 with rate $\theta(n-m)/2$, and for E_3 with rate $\lambda(n)$. The probability of event E_1 is $\mathbb{P}(E_1) = \theta m/(2\lambda(n)+\theta n)$ and, conditional on E_1,

$$\left\{A^{(n)}(\max\{\tau_i : i \in [m]\}) = 0\right\}$$

is determined by the $n-1$ active lineages after the event. The memorylessness of the exponential distribution and natural coupling imply that after the first event,

conditional on that event being E_1, the remaining $n-1$ lineages, of which $(m-1)$ subtend the subsample, follow an $(n-1)$-coalescent with mutation rate θ. Thus,

$$\mathbb{P}\left(A^{(n)}(\max\{\tau_i : i \in [m]\}) = 0 \,|\, E_1\right) = p^{(\Pi^{(\Lambda)})}_{n-1,m-1}.$$

Analogously, we have $\mathbb{P}(E_2) = \theta(n-m)/(2\lambda(n)+\theta n)$. Given E_2, we need to follow the coalescent of $n-1$ lineages, of which m are from the subsample, which gives

$$\mathbb{P}\left(A^{(n)}(\max\{\tau_i : i \in [m]\}) = 0 \,|\, E_2\right) = p^{(\Pi^{(\Lambda)},\theta)}_{n,m-1}.$$

We have $P(E_3) = 1 - P(E_1) - P(E_2) = 2\lambda(n)/(2\lambda(n)+\theta n)$. To compute

$$\mathbb{P}\left(A^{(n)}(\max\{\tau_i : i \in [m]\}) = 0 \,|\, E_3\right),$$

proceed exactly as in the proof of recursion (4) by partitioning over the number of mutant lineages involved in the merger, but with changed boundary conditions since $p^{(\Pi^{(\Lambda)},\theta)}_{i,1} > 0$, while $p^{(\Pi^{(\Lambda)})}_{i,1} = 0$ for $i > 1$. Summing over E_1, E_2, E_3 yields Eq. (11). □

4.8 Proof of Eq. (14)

Recall our assumption that block π_1 always contains element 1. To see (14), we will show that, for n large enough,

$$\mathbb{P}^{(\Pi^{(\Lambda)})}\left(T^{(m;n)}_{\text{MRCA}} \geq \inf\{t \geq 0 : \pi_1 \cap [n]_{m+1} \neq \emptyset, \quad \pi_1 \in \Pi_t\}\right) = 1. \tag{27}$$

In words, the smallest block containing $[m]$ appearing in the n-coalescent will always contain at least $m+1$ elements; block $[m]$ will almost never be observed. Hence, $\lim_{n\to\infty} q^{(\Pi^{(\Lambda)})}_{n,m} = 0$.

Consider first Λ with $\int_{[0,1]} x^{-1}\Lambda(dx) = \infty$, which makes the Λ-coalescent dust-free (no singleton blocks almost surely for $t > 0$) - see the proof of [70, Lemma 25]. For $t > 0$, [70, Prop. 30] shows that the partition block $\pi_1 \in \Pi^{(n,\Lambda)}_t$ containing individual 1 at time t in the Λ-n-coalescent $\{\Pi^{(n,\Lambda)}_t, t \geq 0\}$ fulfills $\lim_{n\to\infty} \#\pi_1/n > 0$ almost surely. Thus, individual 1 has already merged before any time $t > 0$ if $n > N'$, where N' is a random variable on $\mathbb{N}$ almost surely. However, within the subsample of fixed size m, we wait an exponential time with rate $\lambda(m)$ for any merger of individuals in $[m]$. Thus, for n large enough individual 1 has almost surely already merged with individuals of $[n]_{m+1}$ before merging with another individual in the subsample. Consider now Λ with $\int_{[0,1]} x^{-1}\Lambda(dx) < \infty$, which shows that the coalescent has dust, i.e. there is a positive probability that there is a positive fraction of singleton blocks at any time t, see [70, Prop. 26]. In this case [37, Corollary 2.3] shows that at its first merger, for $n \to \infty$, individual 1 merges with a positive fraction of all individuals $\mathbb{N}$ almost surely, which has to include individuals in $[n]_{m+1}$. Since this is the earliest merger where the MRCA of $[m]$ can be reached, the proof is complete. □

Analogously we have the following:

Corollary 1 *Consider any Ξ-coalescent which comes down from infinity and its restrictions to $[n]$, $n \in \mathbb{N}$. Fix subsample size $m \in \mathbb{N}$. Let $\widetilde{T}_m^{(n)}$ be the first time that any $i \in [m]$ is involved in a merger in the Ξ-n-coalescent for $n \geq m$. We have*

$$\lim_{n\to\infty} \mathbb{P}\left(\widetilde{T}_m^{(n)} = T_{MRCA}^{(n)}\right) = 0.$$

Proof If the Ξ-coalescent comes down from infinity, it fulfills $\int_{[0,1]} x^{-1}\Lambda(dx) = \infty$, since it has to be dust-free. As above, we see that individual 1 has already merged before $T_{\mathrm{MRCA}}^{(n)}$ for $n \to \infty$, which establishes the corollary.

□

5 Conclusion and open questions

By studying properties of nested samples we have aimed at understanding how much information about the evolutionary history of a population can be extracted from a sample, i.e. how the genealogical information increases if we enlarge the sample. In particular, we have focussed on multiple-merger coalescent (abbreviated MMC) processes derived from population models characterised by high fecundity and sweepstakes reproduction (abbreviated HFSR). In comparison with the Kingman-coalescent the general conclusion, at least for the statistics we consider, is that a subsample represents less well the 'population' or the complete sample from which the subsample was drawn when the underlying coalescent mechanism admits multiple mergers. The subsample reaches its most recent common ancestor (abbreviated MRCA; see Table 1 for definition of acronyms) sooner and shares less of the ancestral genetic variants (internal branches) with the complete sample under a MMC process than under the Kingman-coalescent. A similar conclusion can be broadly reached in comparison with exponential population growth. This seems to imply that one would need a larger sample for inference under a MMC than under a (time-changed) Kingman-coalescent. Large sample size has been shown to impact inference under the Wright-Fisher model [11], in particular if the sample size exceeds the effective size [86]. The main effect is that when sample size is large enough, one starts to notice multiple and/or simultaneous mergers in the trees — events which would not be possible under the assumption that the sample size is fixed and the population size is arbitrarily large (and thus much larger than the sample size). The implication is that for any finite population, a large enough sample will 'break down' the coalescent approximation. One would also expect an impact of large sample size on inference under MMC.

The effective size in HFSR populations can be much smaller than in a Wright-Fisher population with the same census size [79, 47, 87]. Therefore, for almost any finite population, the genealogy of the whole population is not well approximated by the genealogy one derives under the assumption of a fixed sample size and an arbitrarily large population size. This therefore leaves the question of what one is making an inference about when one applies a coalescent-based inference method — and how to evaluate whether the sample size is small enough that the coalescent approximation holds.

Naturally, this also means that our asymptotic results, most notably the representation for $p_m^{(\text{Beta-coal})}$ from Thm. 1, are not readily applied to real populations. Our asymptotic results rely on the assumption that the coalescent approximation holds for arbitrarily large sample sizes. However, our asymptotic results for $p_m^{(\Xi)}$ for any Ξ are still valid lower bounds for $p_{n,m}^{(\Xi)}$ for any n until the coalescent approximation breaks down.

We focussed mostly on how genealogical properties (MRCA, internal branches) are shared between the complete sample and the nested subsample. These properties cannot be directly observed in genomic data, but they do reflect how much (and how old) polymorphisms can be potentially shared between the samples. We did not discuss similar quantities involving genetic variation (mutations) since, as we discuss in Sec. 2.2.1, a comparison of such quantities is confounded by the differences between different coalescent processes in the way time is measured.

All our results are applicable to a single non-recombining locus. A natural question to ask is if and how our results might change if we considered multiple unlinked loci. How would the statistics we consider, averaged over many unlinked loci, behave under multiple-merger coalescents in comparison with a (time-changed) Kingman-coalescent? DNA sequencing technology has advanced to the degree that sequencing whole genomes is now almost routine (see eg. [43, 4]). One could ask how large a sample from a HFSR population does one need to be confident to have sampled a significant fraction of the genome-wide ancestral variation? In this context, let $T_{\text{MRCA}}^{(n,\ell)}$ denote the TMRCA of the complete sample of size n at a non-recombining locus $\ell \in [L]$, and $T_{\text{MRCA}}^{(m;n,\ell)}$ the TMRCA of a nested subsample of size m at same locus. Then we would like to compare the probability

$$\mathbb{P}^{(\Pi)}\left(\bigcap_{\ell\in[L]}\left\{T_{\text{MRCA}}^{(m;n,\ell)} = T_{\text{MRCA}}^{(n,\ell)}\right\}\right)$$

between different coalescent processes. And in fact, the independence of the genealogies at unlinked loci under the Kingman-coalescent, and Eq. (6), gives

$$\mathbb{P}^{(\text{Kingman})}\left(\bigcap_{\ell\in[L]}\left\{T_{\text{MRCA}}^{(m;n,\ell)} = T_{\text{MRCA}}^{(n,\ell)}\right\}\right) = \left(\frac{(m-1)(n+1)}{(m+1)(n-1)}\right)^L.$$

Under a multiple-merger coalescent process the genealogies at unlinked loci are not independent (see e.g. [32, 15]).

We compared results from single-locus multiple-merger coalescent models with a time-changed Kingman-coalescent derived from a single-locus model of exponential population growth. Naturally one would like to compare results between genomic (multi-locus) models of HFSR with population growth to genomic models of HFSR without growth, and to genomic models of growth without HFSR. Some mathematical handle on the distributions of the quantities we simulated would (obviously) also be nice. However, these will have to remain important open tasks.

Acknowledgements We thank Alison Etheridge for many and very valuable comments and suggestions, especially regarding Theorem 1. BE was funded by DFG grant STE 325/17-1 to Wolfgang Stephan through Priority Programme SPP1819: Rapid Evolutionary Adaptation. FF was funded by DFG grant FR 3633/2-1 through Priority Program 1590: Probabilistic Structures in Evolution.

References

1. Agrios, G.: Plant pathology. Academic Press, Amsterdam (2005)
2. Árnason, E., Halldórsdóttir, K.: Nucleotide variation and balancing selection at the Ckma gene in Atlantic cod: analysis with multiple merger coalescent models. PeerJ **3**, e786 (2015). DOI 10.7717/peerj.786. URL http://dx.doi.org/10.7717/peerj.786
3. Arratia, R., Barbour, A.D., Tavaré, S.: Logarithmic Combinatorial Structures: A Probabilistic Approach. European Mathematical Society (EMS), Zürich (2003)
4. Barney, B.T., Munkholm, C., Walt, D.R., Palumbi, S.R.: Highly localized divergence within supergenes in atlantic cod (gadus morhua) within the gulf of maine. BMC Genomics **18**(1) (2017). DOI 10.1186/s12864-017-3660-3. URL https://doi.org/10.1186/s12864-017-3660-3
5. Barton, N.H., Etheridge, A.M., Véber, A.: Modelling evolution in a spatial continuum. Journal of Statistical Mechanics: Theory and Experiment **2013**(01), P01,002 (2013). URL http://stacks.iop.org/1742-5468/2013/i=01/a=P01002
6. Basu, A., Majumder, P.P.: A comparison of two popular statistical methods for estimating the time to most recent common ancestor (tmrca) from a sample of DNA sequences. Journal of genetics **82**(1-2), 7–12 (2003)
7. Berestycki, J., Berestycki, N., Schweinsberg, J.: Beta-coalescents and continuous stable random trees. Ann Probab **35**, 1835–1887 (2007)
8. Berestycki, J., Berestycki, N., Schweinsberg, J.: Small-time behavior of beta coalescents. Ann Inst H Poincaré Probab Statist **44**, 214–238 (2008)
9. Berestycki, N.: Recent progress in coalescent theory. Ensaios Mathématicos **16**, 1–193 (2009)
10. Bertoin, J.: Exchangeable coalescents. Cours d'école doctorale pp. 20–24 (2010)
11. Bhaskar, A., Clark, A., Song, Y.: Distortion of genealogical properties when the sample size is very large. PNAS **111**, 2385–2390 (2014)
12. Birkner, M., Blath, J.: Computing likelihoods for coalescents with multiple collisions in the infinitely many sites model. J Math Biol **57**, 435–465 (2008)
13. Birkner, M., Blath, J.: coalescents and population genetic inference. Trends in stochastic analysis (353), 329 (2009)
14. Birkner, M., Blath, J., Capaldo, M., Etheridge, A.M., Möhle, M., Schweinsberg, J., Wakolbinger, A.: Alpha-stable branching and beta-coalescents. Electron. J. Probab **10**, 303–325 (2005)
15. Birkner, M., Blath, J., Eldon, B.: An ancestral recombination graph for diploid populations with skewed offspring distribution. Genetics **193**, 255–290 (2013)
16. Birkner, M., Blath, J., Eldon, B.: Statistical properties of the site-frequency spectrum associated with Λ-coalescents. Genetics **195**, 1037–1053 (2013)
17. Birkner, M., Blath, J., Möhle, M., Steinrücken, M., Tams, J.: A modified lookdown construction for the Xi-Fleming-Viot process with mutation and populations with recurrent bottlenecks. ALEA Lat. Am. J. Probab. Math. Stat. **6**, 25–61 (2009)
18. Birkner, M., Blath, J., Steinrücken, M.: Analysis of DNA sequence variation within marine species using Beta-coalescents. Theor Popul Biol **87**, 15–24 (2013)
19. Blath, J., Cronjäger, M.C., Eldon, B., Hammer, M.: The site-frequency spectrum associated with Ξ-coalescents. Theoretical Population Biology **110**, 36–50 (2016). DOI 10.1016/j.tpb.2016.04.002
20. Bolthausen, E., Sznitman, A.: On Ruelle's probability cascades and an abstract cavity method. Comm Math Phys **197**, 247–276 (1998)
21. Capra, J.A., Stolzer, M., Durand, D., Pollard, K.S.: How old is my gene? Trends in Genetics **29**(11), 659–668 (2013)
22. Desai, M.M., Walczak, A.M., Fisher, D.S.: Genetic diversity and the structure of genealogies in rapidly adapting populations. Genetics **193**(2), 565–585 (2013)

23. Dong, R., Gnedin, A., Pitman, J.: Exchangeable partitions derived from markovian coalescents. The Annals of Applied Probability pp. 1172–1201 (2007)
24. Donnelly, P., Kurtz, T.G.: Particle representations for measure-valued population models. Ann Probab **27**, 166–205 (1999)
25. Donnelly, P., Tavare, S.: Coalescents and genealogical structure under neutrality. Annual review of genetics **29**(1), 401–421 (1995)
26. Durrett, R.: Probability models for DNA sequence evolution, 2nd edn. Springer, New York (2008)
27. Durrett, R., Schweinsberg, J.: Approximating selective sweeps. Theor Popul Biol **66**, 129–138 (2004)
28. Durrett, R., Schweinsberg, J.: A coalescent model for the effect of advantageous mutations on the genealogy of a population. Stoch Proc Appl **115**, 1628–1657 (2005)
29. Eldon, B.: Inference methods for multiple merger coalescents. In: P. Pontarotti (ed.) Evolutionary Biology: convergent evolution, evolution of complex traits, concepts and methods, pp. 347–371. Springer (2016)
30. Eldon, B., Birkner, M., Blath, J., Freund, F.: Can the site-frequency spectrum distinguish exponential population growth from multiple-merger coalescents. Genetics **199**, 841–856 (2015)
31. Eldon, B., Wakeley, J.: Coalescent processes when the distribution of offspring number among individuals is highly skewed. Genetics **172**, 2621–2633 (2006)
32. Eldon, B., Wakeley, J.: Linkage disequilibrium under skewed offspring distribution among individuals in a population. Genetics **178**, 1517–1532 (2008)
33. Etheridge, A.: Some Mathematical Models from Population Genetics. Springer Berlin Heidelberg (2011). DOI 10.1007/978-3-642-16632-7. URL http://dx.doi.org/10.1007/978-3-642-16632-7
34. Etheridge, A., Griffiths, R.: A coalescent dual process in a Moran model with genic selection. Theor Popul Biol **75**, 320–330 (2009)
35. Etheridge, A.M., Griffiths, R.C., Taylor, J.E.: A coalescent dual process in a Moran model with genic selection, and the Lambda coalescent limit. Theor Popul Biol **78**, 77–92 (2010)
36. Ewens, W.J.: Mathematical population genetics 1: theoretical introduction, vol. 27. Springer Science & Business Media (2012)
37. Freund, F., Möhle, M.: On the size of the block of 1 for Ξ-coalescents with dust. ArXiv e-prints (2017)
38. Freund, F., Siri-Jégousse, A.: Minimal clade size in the bolthausen-sznitman coalescent. Journal of Applied Probability **51**(3), 657–668 (2014)
39. Goldschmidt, C., Martin, J.B.: Random recursive trees and the bolthausen-sznitman coalescent. Electron. J. Probab **10**(21), 718–745 (2005)
40. Griffiths, R.C., Tavare, S.: Monte carlo inference methods in population genetics. Mathematical and computer modelling **23**(8-9), 141–158 (1996)
41. Griffiths, R.C., Tavaré, S.: The age of a mutation in a general coalescent tree. Comm Statistic Stoch Models **14**, 273–295 (1998)
42. Griswold, C.K., Baker, A.J.: Time to the most recent common ancestor and divergence times of populations of common chaffinches (Fringilla coelebs) in Europe and North Africa: insights into Pleistocene refugia and current levels of migration. Evolution **56**(1), 143–153 (2002)
43. Halldórsdóttir, K., Árnason, E.: Whole-genome sequencing uncovers cryptic and hybrid species among Atlantic and Pacific cod-fish (2015). DOI 10.1101/034926. Http://dx.doi.org/10.1101/034926
44. Hintze, J.L., Nelson, R.D.: Violin plots: A box plot-density trace synergism. The American Statistician **52**(2), 181–184 (1998). DOI 10.1080/00031305.1998.10480559
45. Hedgecock, D.: Does variance in reproductive success limit effective population sizes of marine organisms? In: A. Beaumont (ed.) Genetics and evolution of Aquatic Organisms, pp. 1222–1344. Chapman and Hall, London (1994)
46. Hedgecock, D., Pudovkin, A.I.: Sweepstakes reproductive success in highly fecund marine fish and shellfish: a review and commentary. Bull Marine Science **87**, 971–1002 (2011)
47. Hedrick, P.: Large variance in reproductive success and the N_e/N ratio. Evolution **59**(7), 1596 (2005). DOI 10.1554/05-009
48. Hénard, O.: The fixation line in the Λ-coalescent. The Annals of Applied Probability **25**(5), 3007–3032 (2015)

49. Herriger, P., Möhle, M.: Conditions for exchangeable coalescents to come down from infinity. Alea **9**(2), 637–665 (2012)
50. Hird, S., Kubatko, L., Carstens, B.: Rapid and accurate species tree estimation for phylogeographic investigations using replicated subsampling. Molecular Phylogenetics and Evolution **57**(2), 888–898 (2010)
51. Hovmøller, M.S., Sørensen, C.K., Walter, S., Justesen, A.F.: Diversity of Puccinia striiformis on cereals and grasses. Annual review of phytopathology **49**, 197–217 (2011)
52. Hudson, R.R.: Properties of a neutral allele model with intragenic recombination. Theor Popul Biol **23**, 183–201 (1983)
53. Huillet, T., Möhle, M.: On the extended Moran model and its relation to coalescents with multiple collisions. Theor Popul Biol **87**, 5–14 (2013)
54. Kaj, I., Krone, S.M.: The coalescent process in a population with stochastically varying size. Journal of Applied Probability **40**(01), 33–48 (2003)
55. King, L., Wakeley, J.: Empirical bayes estimation of coalescence times from nucleotide sequence data. Genetics **204**(1), 249–257 (2016). DOI 10.1534/genetics.115.185751
56. Kingman, J.F.C.: The coalescent. Stoch Proc Appl **13**, 235–248 (1982)
57. Kingman, J.F.C.: Exchangeability and the evolution of large populations. In: G. Koch, F. Spizzichino (eds.) Exchangeability in Probability and Statistics, pp. 97–112. North-Holland, Amsterdam (1982)
58. Kingman, J.F.C.: On the genealogy of large populations. J App Probab **19A**, 27–43 (1982)
59. Li, G., Hedgecock, D.: Genetic heterogeneity, detected by PCR-SSCP, among samples of larval Pacific oysters (Crassostrea gigas) supports the hypothesis of large variance in reproductive success. Can. J. Fish. Aquat. Sci. **55**(4), 1025–1033 (1998). DOI 10.1139/f97-312
60. May, A.W.: Fecundity of Atlantic cod. J Fish Res Brd Can **24**, 1531–1551 (1967)
61. Möhle, M.: Robustness results for the coalescent. Journal of Applied Probability **35**(02), 438–447 (1998)
62. Möhle, M.: On sampling distributions for coalescent processes with simultaneous multiple collisions. Bernoulli **12**(1), 35–53 (2006)
63. Möhle, M.: Coalescent processes derived from some compound Poisson population models. Elect Comm Probab **16**, 567–582 (2011)
64. Möhle, M., Sagitov, S.: A classification of coalescent processes for haploid exchangeable population models. Ann Probab **29**, 1547–1562 (2001)
65. Möhle, M., Sagitov, S.: Coalescent patterns in diploid exchangeable population models. J Math Biol **47**, 337–352 (2003)
66. Neher, R.A., Hallatschek, O.: Genealogies of rapidly adapting populations. Proceedings of the National Academy of Sciences **110**(2), 437–442 (2013)
67. Niwa, H.S., Nashida, K., Yanagimoto, T.: Reproductive skew in japanese sardine inferred from DNA sequences. ICES Journal of Marine Science: Journal du Conseil **73**(9), 2181–2189 (2016). DOI 10.1093/icesjms/fsw070. URL http://dx.doi.org/10.1093/icesjms/fsw070
68. Oosthuizen, E., Daan, N.: Egg fecundity and maturity of North Sea cod, Gadus morhua. Netherlands Journal of Sea Research **8**(4), 378–397 (1974)
69. Pettengill, J.B.: The time to most recent common ancestor does not (usually) approximate the date of divergence. PloS one **10**(8), e0128,407 (2015)
70. Pitman, J.: Coalescents with multiple collisions. Ann Probab **27**, 1870–1902 (1999)
71. Sagitov, S.: The general coalescent with asynchronous mergers of ancestral lines. J Appl Probab **36**, 1116–1125 (1999)
72. Sagitov, S.: Convergence to the coalescent with simultaneous mergers. J Appl Probab **40**, 839–854 (2003)
73. Sargsyan, O., Wakeley, J.: A coalescent process with simultaneous multiple mergers for approximating the gene genealogies of many marine organisms. Theor Pop Biol **74**, 104–114 (2008)
74. Saunders, I.W., Tavaré, S., Watterson, G.A.: On the genealogy of nested subsamples from a haploid population. Advances in Applied Probability **16**(3), 471 (1984). DOI 10.2307/1427285
75. Schweinsberg, J.: Rigorous results for a population model with selection II: genealogy of the population. ArXiv:1507.00394
76. Schweinsberg, J.: Coalescents with simultaneous multiple collisions. Electron J Probab **5**, 1–50 (2000)
77. Schweinsberg, J.: Coalescents with simultaneous multiple collisions. Electronic Journal of Probability **5**, 1–50 (2000)

78. Schweinsberg, J.: A necessary and sufficient condition for the-coalescent to come down from the infinity. Electronic Communications in Probability [electronic only] **5**, 1–11 (2000)
79. Schweinsberg, J.: Coalescent processes obtained from supercritical Galton-Watson processes. Stoch Proc Appl **106**, 107–139 (2003)
80. Simon, M., Cordo, C.: Inheritance of partial resistance to Septoria tritici in wheat (Triticum aestivum): limitation of pycnidia and spore production. Agronomie **17**(6-7), 343–347 (1997)
81. Slack, R.: A branching process with mean one and possibly infinite variance. Probability Theory and Related Fields **9**(2), 139–145 (1968)
82. Spouge, J.L.: Within a sample from a population, the distribution of the number of descendants of a subsample's most recent common ancestor. Theoretical population biology **92**, 51–54 (2014)
83. Tajima, F.: Evolutionary relationships of DNA sequences in finite populations. Genetics **105**, 437–460 (1983)
84. Timm, A., Yin, J.: Kinetics of virus production from single cells. Virology **424**(1), 11–17 (2012)
85. Wakeley, J.: Coalescent theory. Roberts & Co (2007)
86. Wakeley, J., Takahashi, T.: Gene genealogies when the sample size exceeds the effective size of the population. Mol Biol Evol **20**, 208–2013 (2003)
87. Waples, R.S.: Tiny estimates of the N_e/N ratio in marine fishes: Are they real? Journal of Fish Biology **89**(6), 2479–2504 (2016). DOI 10.1111/jfb.13143
88. Wiuf, C., Donnelly, P.: Conditional genealogies and the age of a neutral mutant. Theoretical Population Biology **56**(2), 183 – 201 (1999). DOI http://dx.doi.org/10.1006/tpbi.1998.1411. URL http://www.sciencedirect.com/science/article/pii/S0040580998914113
89. Zhou, J., Teo, Y.Y.: Estimating time to the most recent common ancestor (tmrca): comparison and application of eight methods. European Journal of Human Genetics (2015)

A1 Population models

In this section we provide a brief overview of the population models behind the coalescent processes we consider, and why we think they are interesting. A detailed description of the coalescent processes is given in Sec. A2.

A universal mechanism among all biological populations is reproduction and inheritance. Reproduction refers to the generation of offspring, and inheritance refers to the transmission of information necessary for viability and reproduction. Mendel's laws on independent segregation of chromosomes into gametes describe the transmission of information from a parent to an offspring in a diploid population. For our purposes, however, it suffices to think of haploid populations where one can think of an individual as a single gene copy. By tracing gene copies as they are passed on from one generation to the next one automatically stores two sets of information. On the one hand one stores how frequencies of genetic types change going forwards in time; on the other hand one keeps track of the ancestral, or genealogical, relations among the different copies. This duality has been successfully exploited for example in modeling selection [34,35]. To model genetic variation in natural populations one requires a mathematically tractable model of how genetic information is passed from parents to offspring. In the Wright-Fisher model offspring choose their parents independently and uniformly at random. Suppose we are tracing the ancestry of $n \geq 2$ gene copies in a haploid Wright-Fisher population of N gene copies in total. For any pair, the chance that they have a common ancestor in the previous generation is $1/N$. Informally, we trace the genealogy of our gene copies on the order of $\mathcal{O}(N)$ generations until we see the first merger, i.e. when at least 2 gene copies (or their ancestral lines) find a common ancestor.

If n is small relative to N, when a merger occurs, with probability $1-\mathcal{O}(1/N)$ it involves just two ancestral lineages. This means that if we measure time in units of N generations, and assume N is very large, the random ancestral relations of our sampled gene copies can be described by a continuous-time Markov chain in which each pair of ancestral lines merges at rate 1 and no other mergers are possible. We have, in an informal way, arrived at the Kingman-coalescent [56,58,57]. One can derive the Kingman-coalescent not just from the Wright-Fisher model but from any population model which satisfies certain assumptions on the offspring distribution [61,71,64]. These assumptions mainly dictate that higher moments of the offspring number distribution are small relative to (an appropriate power of) the population size. The Kingman-coalescent, and its various extensions, are used almost universally as the 'null model' for a gene genealogy in population genetics. The Kingman-coalescent is a remarkably good model for populations characterised by low fecundity, i.e. whose individuals have small numbers of offspring relative to the population size.

The classical Kingman-coalescent is derived from a population model in which the population size is constant between generations. Extensions to stochastically varying population size, in which the population size does not vary 'too much' between generations, have been made [54]; the result is a time-changed Kingman-coalescent. Probably the most commonly applied model of deterministically changing population size is the model of exponential population growth (see eg. [25,41, 30]). In each generation the population size is multiplied by a factor $(1+\beta/N)$, where $\beta>0$. Therefore, the population size in generation k going forward in time is given by $N_k = N(1+\beta/N)^k$ where N is taken as the 'initial' population size. It follows that the population size $\lfloor Nt \rfloor$ generations ago is $Ne^{-\beta t}$. [30] show that exponential population growth can be distinguished from multiple-merger coalescents (in which at least three ancestral lineages can merge simultaneously), derived from population models of high fecundity and sweepstakes reproduction, using population genetic data from a single locus, provided that sample size and number of mutations (segregating sites) are not too small.

A diverse group of natural populations, including some marine organisms [46], fungi [1,80,51], and viruses [84] are highly fecund. By way of example, individual Atlantic codfish [60,68] and Pacific oysters [59] can lay millions of eggs. This high fecundity counteracts the high mortality rate among the larvae (juveniles) of these populations (Type III survivorship). The term 'sweepstakes reproduction' has been proposed to describe the reproduction mode of highly fecund populations with Type III survivorship [45]. Population models which admit high fecundity and sweepstakes reproduction (HFSR) through skewed or heavy-tailed offspring number distributions have been developed [64,65,79,31,73,53]. In the haploid model of [79], each individual independently contributes a random number X of juveniles where $(C,\alpha>0)$

$$\mathbb{P}(X \geq k) \sim \frac{C}{k^{\alpha}}, \quad k \to \infty, \tag{A28}$$

and $x_n \sim y_n$ means $x_n/y_n \to 1$ as $n \to \infty$. The constant $C>0$ is a normalising constant, and the constant α determines the skewness of the distribution. The next generation of individuals is then formed by sampling (uniformly without replacement) from the pool of juveniles. In the case $\alpha<2$ the random ancestral relations

of gene copies can be described by specific forms of multiple-merger coalescent processes [72]. We remark that the fate of the juveniles need not be correlated to generate multiple-mergers in the genealogies — the heavy-tailed distribution of juveniles means that occasionally one 'lucky' individual contributes a huge number of juveniles while all others contribute only a small number of juveniles. Uniform sampling without replacement from the pool of juveniles means that the lucky individual leaves significantly more descendents in the next generation than anyone else, and this is what generates multiple mergers of ancestral lines.

Coalescent processes derived from population models of HFSR (see (A28) for an example) admit multiple mergers of ancestral lineages [24,70,71,76,65, 72,63]. Mathematically, we consider exchangeable n-coalescent processes, which are Markovian processes $(\Pi_t^{(n)})_{t\geq 0}$ on the set of partitions of $[n] := \{1,2,\ldots,n\}$ whose transitions are mergers of partition blocks (a 'block' is a subset of $[n]$, see Sec. A2) with rates specified in Sec. A2. The blocks of $\Pi_t^{(n)}$ show which individuals in $[n]$ share a common ancestor at time t measured from the time of sampling. Thus, the blocks of $\Pi_t^{(n)}$ can be interpreted as ancestral lineages. The specific structure of the transition rates allows to treat a multiple-merger n-coalescent as the restriction of an exchangeable Markovian process $(\Pi_t)_{t\geq 0}$ on the set of partitions of $\mathbb{N}$, which is called a multiple-merger coalescent (abbreviated MMC) process. MMC processes are referred to as Λ-coalescents (Λ a finite measure on $[0,1]$) [24,70,71] if any number of ancestral lineages can merge at any given time, but only one such merger occurs at a time. By way of an example, if $1 \leq \alpha < 2$ in (A28) one obtains a so-called Beta$(2-\alpha,\alpha)$-coalescent [72] (Beta-coalescent, see Eq. (A35)). Processes which admit at least two (multiple) mergers at a time are referred to as Ξ-coalescents (Ξ a finite measure on the infinite simplex Δ) [76, 64,65]. See Sec. A2 for details. Specific examples of these MMC processes have been shown to give a better fit to genetic data sampled from Atlantic cod [12,18,2, 16,19] and Japanese sardines [67] than the classical Kingman-coalescent. See e.g. [29] for an overview of inference methods for MMC processes. [46] review the evidence for sweepstakes reproduction among marine populations and conclude 'that it plays a major role in shaping marine biodiversity'.

MMC models also arise in contexts other than high fecundity. [17] show that repeated strong bottlenecks in a Wright-Fisher population lead to time-changed Kingman-coalescents which look like Ξ-coalescents. [27,28] show that the genealogy of a locus subjected to repeated beneficial mutations is well approximated by a Ξ-coalescent. [75] provides rigorous justification of the claims of [66,22] that the genealogy of a population subject to repeated beneficial mutations can be described by the Beta-coalescent with $\alpha = 1$ (also referred to as the Bolthausen-Sznitman coalescent [20]). These examples show that MMC processes are relevant for biology. We refer the interested reader to e.g. [10,25,5,33,9,13] for a more detailed background on coalescent theory.

A2 Coalescent processes

To keep our presentation self-contained a precise definition of the coalescent processes we will need will now be given. We follow the description of [19]. A coalescent process Π is a continuous-time Markov chain on the partitions of $\mathbb{N}$. Let

$\Pi^{(n)}$ denote the restriction to $[n]$, and write $\mathscr{P}_n$ for the space of partitions of $[n]$. A partition $\pi = \{\pi_1, \ldots, \pi_{\#\pi}\} \in \mathscr{P}_n$ has $\#\pi$ *blocks* which are disjoint subsets of $[n]$. We assume the blocks π_i are ordered by their smallest element; therefore we always have $1 \in \pi_1$. In general a merging event can involve r distinct groups of blocks merging simultaneously. We write $\underline{k} = (k_1, \ldots, k_r)$ where $k_i \geq 2$ denotes the number of blocks merging in group i. Here $r \in [\lfloor \#\pi/2 \rfloor]$, $k_1 + \cdots + k_r \in [\#\pi]_2$ and $i_1^{(a)}, \ldots, i_{k_a}^{(a)}$ will denote the indices of the blocks in the ath group. By $\pi' \prec_{\#\pi,\underline{k}} \pi$ we denote a transition from π to $\pi' = A \cup B$ where

$$\begin{aligned} A &= \left\{ \pi_\ell : \ell \in [\#\pi], \ell \notin \bigcup_{a=1}^{r} \left\{ i_1^{(a)}, \ldots, i_{k_a}^{(a)} \right\} \right\}, \\ B &= \bigcup_{b=1}^{r} \left\{ \pi_{i_1^{(b)}}, \ldots, \pi_{i_{k_b}^{(b)}} \right\}. \end{aligned} \tag{A29}$$

In (A29), set A (possibly empty) contains the blocks not involved in a merger, and B lists the blocks involved in each of the r mergers. By $\pi' \prec_{\#\pi,k} \pi$ we denote the transition in a Λ-coalescent where $k \in [\#\pi]_2$ merge in a single merger and π' is given as in (A29) with $r = 1$; ie. only one group of blocks merges in each transition. By $\pi' \prec_{\#\pi} \pi$ we denote a transition in the Kingman-coalescent where $r = 1$ and 2 blocks merge in each transition.

Now that we have specified the possible transitions, we can state the rates of the transitions. Let Δ denote the infinite simplex $\Delta = \{(x_1, x_2, \ldots) : x_1 \geq x_2 \geq \ldots \geq 0, \sum_i x_i \leq 1\}$; let $\boldsymbol{x}$ denote an element of Δ. Define the functions $f(\boldsymbol{x}; \#\pi, \underline{k})$ and $g(\boldsymbol{x}; \#\pi, \underline{k})$ on $\Delta_{\mathbf{0}} := \Delta \setminus \{(0,0,\ldots)\}$ where $\left(\prod_{m=1}^{0} x_{i_{r+m}} := 1\right)$, and $s = \#\pi - k_1 - \ldots - k_r$, by

$$\begin{aligned} f(\boldsymbol{x}; \#\pi, \underline{k}) &= \frac{1}{\sum_j x_j^2} \sum_{\ell=0}^{s} \sum_{i_1 \neq \ldots \neq i_{r+\ell}} \binom{s}{\ell} x_{i_1}^{k_1} \cdots x_{i_r}^{k_r} \prod_{m=1}^{\ell} x_{i_{r+m}} \left(1 - \sum_j x_j\right)^{s-\ell}, \\ g(\boldsymbol{x}; n) &= \frac{1 - \sum_{\ell=0}^{n} \sum_{i_1 \neq \ldots \neq i_\ell} \binom{n}{\ell} x_{i_1} \cdots x_{i_\ell} \left(1 - \sum_j x_j\right)^{n-\ell}}{\sum_j x_j^2}. \end{aligned} \tag{A30}$$

where $x_{i_0} := 1$. For a finite measure Ξ on Δ, set $\Xi_0 := \Xi(\cdot \cap \Delta_0)$ and $a := \Xi(\{(0,0,\ldots)\})$. Then, define

$$\begin{aligned} \lambda_{n,\underline{k}} &:= \int_{\Delta_{\mathbf{0}}} f(\boldsymbol{x}, n, \underline{k}) \Xi_{\mathbf{0}} d\boldsymbol{x} + a \mathbb{1}_{(r=1, k_1=2)}, \\ \lambda_n &:= \int_{\Delta_{\mathbf{0}}} g(\boldsymbol{x}, n) \Xi_{\mathbf{0}} d\boldsymbol{x} + a \binom{n}{2}. \end{aligned} \tag{A31}$$

A Ξ-coalescent [76] is a continuous-time $\mathscr{P}_n$-valued Markov chain with transitions $q_{\pi,\pi'}$ given by, where $\lambda_{n,\underline{k}}$ and λ_n are given in (A31),

$$q_{\pi,\pi'} = \begin{cases} \lambda_{n,\underline{k}} & \text{if } \pi' \prec_{\#\pi,\underline{k}} \pi, \#\pi = n, \\ -\lambda_n & \text{if } \pi' = \pi \text{ and } n = \#\pi, \\ 0 & \text{otherwise.} \end{cases} \tag{A32}$$

A Λ-coalescent [24,70,71] is a specific case of a Ξ-coalescent where $\Xi_{\mathbf{0}}$ only has support on $\Delta_0 := \Delta_{\mathbf{0}} \cap \{(x_1, x_2, \dots) : x_1 \in (0,1], x_{1+i} = 0 \,\forall\, i \in \mathbb{N}\}$ [76]. Let Λ denote the restriction of Ξ on its first coordinate (which makes Λ a finite measure on $[0,1]$). The transition rate of $\pi' \prec_{\#\pi,k} \pi$ becomes, where $\#\pi = n$, $2 \le k \le n$,

$$\lambda_{n,k} = \int_0^1 x^{k-2}(1-x)^{n-k}\Lambda(dx), \quad 2 \le k \le n. \tag{A33}$$

The total rate of k-mergers in a Λ-coalescent is given by $\lambda_k(n) = \binom{n}{k}\lambda_{n,k}$ for $2 \le k \le n$. The total rate of mergers given $n \ge 2$ active blocks is

$$\lambda(n) = \lambda_2(n) + \cdots + \lambda_n(n). \tag{A34}$$

An important example of a Λ-coalescent is the Beta$(2-\alpha,\alpha)$-coalescent [79] where the Λ measure is associated with the beta density, where $B(\cdot,\cdot)$ is the beta function,

$$\Lambda(dx) = \frac{x^{1-\alpha}(1-x)^{\alpha-1}}{B(2-\alpha,\alpha)}dx, \quad 1 \le \alpha < 2. \tag{A35}$$

The total rate of a k-merger $\lambda_k(n) = \binom{n}{k}\lambda_{n,k}$ (see Eq. (A33)) is then given by, for $2 \le k \le n$,

$$\lambda_k(n) = \binom{n}{k}\frac{B(k-\alpha, n-k+\alpha)}{B(2-\alpha,\alpha)}, \quad 1 \le \alpha < 2. \tag{A36}$$

For $\alpha = 1$ the Beta$(2-\alpha,\alpha)$-coalescent is the Bolthausen-Sznitman coalescent [20,39]. The Beta-coalescent is well-studied, there are connections to superprocesses, continuous-state branching processes (CSBP) and continuous stable random trees as described e.g. in [14] and [7].

A3 Goldschmidt and Martin's construction of the Bolthausen-Sznitman n-coalescent

From [39], we recall the construction of the Bolthausen-Sznitman n-coalescent by cutting the edges of a random recursive tree. Let $\mathbb{T}_n$ be a random recursive tree with n nodes. We can construct $\mathbb{T}_n$ sequentially as follows

(i) Start with a node labelled with 1 (the root) and no edges,
(ii) If $i < n$ nodes are present, add a node labelled with $i+1$ and one edge connecting it to a node in $[i]$ picked uniformly,
(iii) stop if n nodes are present.

The object $\mathbb{T}_n$ is a labelled tree, each node has a single label. We consider a realisation of $\mathbb{T}_n$ and transform this tree over time into labelled trees with fewer nodes with nodes amassing multiple labels.

(i) Each edge of $\mathbb{T}_n$ is linked to an exponential clock. The clocks are i.i.d. Exp(1)-distributed.
(ii) We wait for the first clock to ring. At this time, we cut/remove the edge whose clock rang first. The tree is thus split in two trees, one of these trees includes the node with label 1. We denote this tree by $\mathbb{T}^{(1)}$, the other tree by $\mathbb{T}^{(2)}$. Let e_1 be the node of $\mathbb{T}^{(1)}$ that was connected to the removed edge.

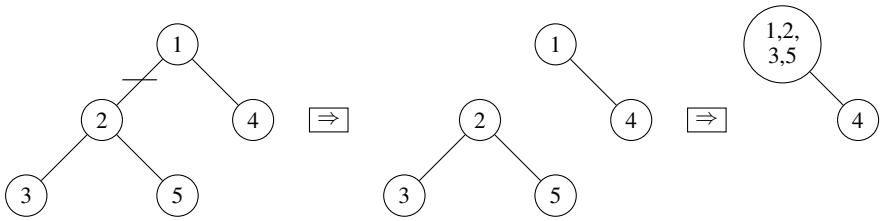


Fig. 7 Example for the first cutting and relabelling step (ii), (iii) for the construction from [39].

(iii) All labels of $\mathbb{T}^{(2)}$ are added to the set of labels of e_1. Remove $\mathbb{T}^{(2)}$ including its clocks.

(iv) Repeat from (ii), using $\mathbb{T}^{(1)}$ labelled as in (iii) with the (remaining) clocks from (i). Stop when $\mathbb{T}^{(1)}$ in step (iii) consists of only a single node and no edges.

(v) For any time t, label sets at the nodes of $\mathbb{T}^{(1)}$ ($\mathbb{T}_n$ before the first clock has rang) give a partition $\Pi_t^{(n)}$ of $[n]$. The process $(\Pi_t^{(n)})_{t\geq 0}$ is a Bolthausen-Sznitman n-coalescent (set $\Pi_t^{(n)} = [n]$ if t is bigger than the time at which we stopped the cutting procedure).

Figure A3 shows an illustration of steps (i)-(iii) for a realisation of $\mathbb{T}_5$.