bioRxiv preprint doi: https://doi.org/10.1101/025684; this version posted February 6, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The site-frequency spectrum associated with Ξ-coalescents

Jochen Blath[1], Mathias Christensen Cronjäger[2], Bjarki Eldon[3], Matthias Hammer[1]

February 4, 2016

Author affiliations:

1. TU Berlin, Institut für Mathematik
10623 Berlin, Germany

2. University of Oxford, Department of Statistics
OX1 3TG Oxford, UK

3. Museum für Naturkunde
Leibniz Institut für Evolutions- und Biodiversitätsforschung
10115 Berlin

bioRxiv preprint doi: https://doi.org/10.1101/025684; this version posted February 6, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Running title: Frequency spectrum and Xi-coalescents

Keywords: site-frequency spectrum, Xi-coalescents, diploidy, Atlantic cod, simultaneous mergers

Corresponding author:
Bjarki Eldon
Museum für Naturkunde,
Leibniz Institut für Evolutions- und Biodiversitätsforschung
Invalidenstraße 43
10115 Berlin, Germany
Email: eldon@math.tu-berlin.de, bjarki.eldon@mfn-berlin.de
Phone: +49(0)30209370379

bioRxiv preprint doi: https://doi.org/10.1101/025684; this version posted February 6, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Abstract

We give recursions for the expected site-frequency spectrum associated with so-called *Xi-coalescents*, that is exchangeable coalescents which admit *simultaneous multiple mergers* of ancestral lineages. Xi-coalescents arise, for example, in association with population models of skewed offspring distributions with diploidy, recurrent advantageous mutations, or strong bottlenecks. In contrast, the simpler *Lambda-coalescents* admit multiple mergers of lineages, but at most one such merger each time. Xi-coalescents, as well as Lambda-coalescents, can predict an excess of singletons, compared to the Kingman coalescent. We compare estimates of coalescent parameters when Xi-coalescents are applied to data generated by Lambda-coalescents, and vice versa. In general, Xi-coalescents predict fewer singletons than corresponding Lambda-coalescents, but a higher count of mutations of size larger than singletons. We fit examples of Xi-coalescents to unfolded site-frequency spectra obtained for autosomal loci of the diploid Atlantic cod, and obtain different coalescent parameter estimates than obtained with corresponding Lambda-coalescents. Our results provide new inference tools, and suggest that for autosomal population genetic data from diploid or polyploid highly fecund populations who may have skewed offspring distributions, one should not apply Lambda-coalescents, but Xi-coalescents.

bioRxiv preprint doi: https://doi.org/10.1101/025684; this version posted February 6, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Introduction

The coalescent approach, ie. the idea of considering the (random) ancestral relations of alleles sampled from natural populations, has provided rich mathematical theory [cf. 7], and very useful inference methods (cf. eg. [18, 45] for reviews). Initiated by the Kingman coalescent [28, 30, 29], coalescent models now include the family of Lambda-(Λ-)coalescents [35, 36, 17], and Xi-(Ξ-)coalescents [40, 33, 37]. Ξ-coalescents admit *simultaneous multiple mergers* of ancestral lineages. Thus, in each merger event, distinct groups of ancestral lineages can merge at the same time, and each group can have *more* than two lineages. Λ-coalescents, in contrast, only allow one group - possibly containing more than two lineages - to merge each time. Thus, due to multiple mergers, the derivation and application of inference methods becomes harder as one moves from the Kingman coalescent to Lambda-coalescent models, and from Λ-coalescents to Ξ-coalescents. Ξ-coalescents can be obtained from diploid population models [9, 34, 14]. They also arise in models of repeated strong bottlenecks [11], and in models of selective sweeps [19, 20].

[24] obtained closed-form expressions for the expected site-frequency spectrum, and (co)-variances, when associated with the Kingman coalescent. [10] obtain recursions for expected values and (co)-variances when associated with Λ-coalescents. However, the complexity of the recursions means that (co)-variances, when associated with Λ-coalescents, can only be computed for small sample sizes. The expected values can be applied in distance statistics [25], and in an approximate likelihood approach [10, 22].

Multiple merger coalescents can be obtained from population models that admit high fecundity and skewed offspring distributions, characteristics associated with many marine populations [1, 23, 8, 12, 13, 26, 27, 38]. Indeed, [2] find much better fit (ie. minimizing the sum of the squared distance between observed and expected values) between data on autosomal genes from Atlantic cod and Λ-coalescents than with the Kingman coalescent. Simulation results of [38] suggest that the site-frequency spectrum of (at least some) Ξ-coalescents is multi-modal, a pattern observed in data on the autosomal *Ckma* gene in Atlantic cod [2]. Based on this evidence, a way to compute expected values of the site-frequency spectrum

bioRxiv preprint doi: https://doi.org/10.1101/025684; this version posted February 6, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

associated with Ξ-coalescents should be a welcome and important addition to the set of inference methods for population genetics. Time-changed Kingman-coalescents obtained from models of population growth can predict an excess of singletons, a characteristic observed for some marine populations [cf. eg. 27]. However, the right tails of the site-frequency spectrum predicted by multiple-merger coalescents and time-changed Kingman-coalescents differ, which makes it possible to distinguish between them [22].

In this work, we obtain recursions for the expected site-frequency spectrum associated with general Ξ-coalescents, with an approach similar to the one applied by [10]. We compare estimates of coalescent parameters when applied to simulated data obtained under Λ-coalescents, and vice-versa. Since the recursions for the expected values are already fairly complex, and computationally expensive, we expect recursions for the (co)-variances to be even more so. The (co)-variances will therefore not be addressed.

We estimate coalescent parameters associated with specific examples of Xi-coalescents for the unfolded site-frequency spectrum of 3 autosomal loci [2] of the highly fecund Atlantic cod. Our simple method involves minimising the distance between observed and expected values, where the distance is not calibrated by the corresponding variance. Our estimates differ from previous estimates obtained with the use of Lambda-coalescents. The main biological implication of our results are that Xi-coalescents should be applied to autosomal data from highly fecund diploid (or polyploid) populations, and Lambda-coalescents to haploid data such as mitochondrial DNA.

The paper is structured as follows: First we give a precise mathematical description of the various coalescent models. We then state our main result on the expected site frequency spectrum of Ξ-coalescents, Theorem 2. This is followed by a discussion of some specific examples of Ξ-coalescents. Some numerical examples which illustrate the difference in the site-frequency spectrum between Lambda- and Xi-coalescents are then presented, followed by an application to autosomal Atlantic cod data. The proofs are collected in an Appendix.

bioRxiv preprint doi: https://doi.org/10.1101/025684; this version posted February 6, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Theory

For ease of reference, we collect the notation we use in the following table.

Table 1: Notation.

notation	explanation
$[n]$	$[n] := \{1, 2, \ldots, n\}$
$\mathcal{P}_n$	space of partitions of $[n]$
π	generic element of $\mathcal{P}_n$
$\#\pi$	number of blocks in π
$\pi' \prec \pi$	π' obtained from π by merging 2 blocks
$\pi' \prec_{m,k} \pi$	π' obtained from π by merging $k \leq m = \#\pi$ blocks
$\underline{k}$	$\underline{k} = (k_1, \ldots, k_r)$
$\pi' \prec_{m,\underline{k}} \pi$	π' obtained from π by a $\underline{k}$-merger when $m = \#\pi$
$\lvert\underline{k}\rvert$	$\lvert\underline{k}\rvert = k_1 + \cdots + k_r$
$\boldsymbol{x}$	$\boldsymbol{x} = (x_1, x_2, \ldots)$
Δ	$\Delta = \{\boldsymbol{x} : x_1 \geq x_2 \geq \cdots \geq 0, x_1 + x_2 + \cdots \leq 1\}$
$\tilde{n}$	the set of all partitions of integer n
$\lvert\nu\rvert = n$	means $\nu \in \tilde{n}$, ie. ν is a partition of n
$\lvert\nu\rvert \overset{k}{=} n$	$\nu \in \tilde{n}$ and $\#\nu = k$, ie. ν has k elements
$\nu = \langle \alpha_1, \alpha_2, \ldots \rangle$	α_j specifies how often j appears in integer partition ν
$\mu = \langle \beta_1, \beta_2, \ldots \rangle \subset \nu$	$\beta_j \leq \alpha_j$ for all j
$\pi^{\downarrow}$	$\pi \in \mathcal{P}_n$, $\lvert\pi^{\downarrow}\rvert \overset{\#\pi}{=} n$, $\pi_1^{\downarrow} \geq \pi_2^{\downarrow} \geq \cdots \geq \pi_{\#\pi}^{\downarrow}$, and the $\pi_j^{\downarrow}$ are the block sizes of π
$(\alpha) = (\alpha_2, \alpha_3, \ldots)$	α_j denotes the number of mergers of size j
$\mathbb{1}_{(A)}$	$\mathbb{1}_{(A)} = 1$ if A holds, and 0 otherwise
$(a)_m$	$(a)_m = a(a-1)\cdots(a-m+1)$, $(a)_0 := 1$ (falling factorial)
Π	coalescent process $\{\Pi_t, t \geq 0\}$
$\mathbb{E}^{(\Pi)}$	expected value relative to Π

Coalescent models

We briefly review the basic coalescent models, namely the Kingman-, Λ-, and Ξ-coalescents. They all have in common to be continuous-time Markov chains, taking values in the space of partitions of the natural numbers $\mathbb{N} := \{1, 2, \ldots, \}$, whose restriction to the first n integers can be described as follows: Let $\mathcal{P}_n$ denote the space of partitions of $[n] := \{1, \ldots, n\}$. We write π for a generic element of $\mathcal{P}_n$, and $\#\pi$ for the *size* of π, i.e. for the number of blocks

bioRxiv preprint doi: https://doi.org/10.1101/025684; this version posted February 6, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

$\pi_i \in \pi$. Thus for $\pi \in \mathcal{P}_n$ we have $\pi = \{\pi_1, \ldots, \pi_{\#\pi}\}$ with $\#\pi \leq n$.

Kingman coalescent

If $\pi, \pi' \in \mathcal{P}_n$ with $\#\pi = m \in \{2, \ldots, n\}$, we write $\pi' \prec \pi$ if there exist $i, j \in [m]$ with $\pi' = \{\pi_\ell : \ell \in [m], \ell \notin \{i, j\}\} \cup \{\pi_i \cup \pi_j\}$, ie. π' is obtained from π by merging blocks π_i and π_j. If the transition rates of the continuous-time Markov chain $\{\Pi_t^{(K,n)}, t \geq 0\}$ with values in $\mathcal{P}_n$, and starting from state $\{\{1\}, \ldots, \{n\}\}$ at time $t = 0$, are given by

$$q_{\pi,\pi'} = \begin{cases} 1 & \text{if } \pi' \prec \pi, \\ -\binom{m}{2} & \text{if } \pi' = \pi,\ \#\pi = m \geq 2, \\ 0 & \text{otherwise;} \end{cases} \tag{1}$$

we refer to $\{\Pi_t^{(K,n)}, t \geq 0\}$ as the Kingman-n-coalescent. The process is stopped at time $\inf\{t > 0 : \Pi_t^{(K,n)} = \{\{1, \ldots, n\}\}\}$, ie. when the most recent common ancestor of the n lineages has been reached.

Lambda-coalescent

If $\pi, \pi' \in \mathcal{P}_n$ with $\#\pi = m \in \{2, \ldots, n\}$, and there exist indices $i_1, \ldots, i_k \in [m]$ with $\pi' = \{\pi_\ell : \ell \in [m], \ell \notin \{i_1, \ldots, i_k\}\} \cup \{\pi_{i_1} \cup \cdots \cup \pi_{i_k}\}$, we write $\pi' \prec_{m,k} \pi$ and say that a k-merger has occurred, with $2 \leq k \leq m$. For a finite measure Λ on $[0, 1]$, define

$$\lambda_{m,k} := \int_0^1 x^{k-2}(1-x)^{m-k}\Lambda(dx), \tag{2}$$

$$\lambda_m := \int_0^1 [1 - (1-x)^m - mx(1-x)^{m-1}]x^{-2}\Lambda(dx) \tag{3}$$

if the integral in (3) exists, and

$$\lambda_m := \sum_{k=2}^{m} \binom{m}{k} \int_0^1 x^{k-2}(1-x)^{m-k}\Lambda(dx) \tag{4}$$

bioRxiv preprint doi: https://doi.org/10.1101/025684; this version posted February 6, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

otherwise. For example, the integral in (3) does not exist in case of the Beta$(2-\alpha,\alpha)$-coalescent introduced by [41].

A $\mathcal{P}_n$-valued continuous-time Markov chain $\{\Pi_t^{(\Lambda,n)}, t \geq 0\}$ with transition rates $q_{\pi,\pi'}$ from π to π' given by

$$q_{\pi,\pi'} = \begin{cases} \lambda_{m,k} & \text{if } \pi' \prec_{m,k} \pi, \\ -\lambda_m & \text{if } \pi' = \pi,\ \#\pi = m, \\ 0 & \text{otherwise;} \end{cases} \tag{5}$$

is referred to as a Λ-n-coalescent. The waiting time in state π is exponential with rate λ_m as in (3, 4).

Xi-coalescent

Now we specify the transition rates for a Ξ-n-coalescent. Let $\underline{k} \equiv (k_1, \ldots, k_r)$ denote a vector of positive integers of length $r \geq 1$. We write $\pi' \prec_{m,\underline{k}} \pi$ if $\pi, \pi' \in \mathcal{P}_n$ with $\#\pi = m \in \{2, \ldots, n\}$, and there exist $r \in [\lfloor \frac{m}{2} \rfloor]$ groups of indices $i_1^{(j)}, \ldots, i_{k_j}^{(j)}$, $j = 1, \ldots, r$, such that

$$\pi' = \left\{ \pi_\ell : \ell \in [m], \ell \notin \{i_1^{(1)}, \ldots, i_{k_1}^{(1)}\} \cup \cdots \cup \{i_1^{(r)}, \ldots, i_{k_r}^{(r)}\} \right\} \cup \left\{ \pi_{i_1^{(1)}} \cup \cdots \cup \pi_{i_{k_1}^{(1)}} \right\} \cup \cdots \cup \left\{ \pi_{i_1^{(r)}} \cup \cdots \cup \pi_{i_{k_r}^{(r)}} \right\},$$

by which we denote a transition where blocks with indices $i_1^{(j)}, \ldots, i_{k_j}^{(j)}$ merge into a single block, for $j \in [r]$. Thus, a transition denoted by $\pi' \prec_{m,\underline{k}} \pi$ is a *simultaneous multiple merger*, where $r \in [\lfloor \frac{m}{2} \rfloor]$ such mergers occur simultaneously. The vector $\underline{k} = (k_1, \ldots, k_r)$ specifies the merger sizes, and we write $|\underline{k}| := \sum_{j=1}^r k_j$.

Let Δ denote the infinite simplex

$$\Delta := \left\{ \boldsymbol{x} = (x_1, x_2, \ldots) : x_1 \geq x_2 \geq \cdots \geq 0, \quad \sum_i x_i \leq 1 \right\}. \tag{6}$$

Let $\boldsymbol{x} \in \Delta$, $m \in \mathbb{N}$, $\underline{k} = (k_1, \ldots, k_r)$ with $k_i \geq 2$ and $|\underline{k}| = k_1 + \cdots + k_r \leq m$ the sum of the r merger sizes, and $s := m - |\underline{k}|$ the number of blocks unaffected by the merger(s) specified by

bioRxiv preprint doi: https://doi.org/10.1101/025684; this version posted February 6, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

$\underline{k}$. Define the functions $f(\boldsymbol{x}, m, \underline{k})$ and $g(\boldsymbol{x}, m, \underline{k})$, with $\boldsymbol{x} \in \Delta_{\mathbf{0}} := \Delta \setminus \{(0,0,\ldots)\} = \Delta \setminus \{\mathbf{0}\}$,

$$f(\boldsymbol{x}, m, \underline{k}) := \frac{\sum_{\ell=0}^{s} \sum_{i_1 \neq \ldots \neq i_{r+\ell}} \binom{s}{\ell} x_{i_1}^{k_1} \cdots x_{i_r}^{k_r} x_{i_{r+1}} \cdots x_{i_{r+\ell}} \left(1 - \sum_j x_j\right)^{s-\ell}}{\sum_j x_j^2},$$
$$g(\boldsymbol{x}, m) := \frac{1 - \sum_{\ell=0}^{m} \sum_{i_1 \neq \ldots \neq i_\ell} \binom{m}{\ell} x_{i_1} \cdots x_{i_\ell} \left(1 - \sum_j x_j\right)^{m-\ell}}{\sum_j x_j^2}. \tag{7}$$

Let $\Xi_{\mathbf{0}}$ denote a finite measure on $\Delta_{\mathbf{0}}$, and write $\Xi := \Xi_{\mathbf{0}} + a\delta_{\{\mathbf{0}\}}$. Further, let $\mathcal{N}(m, \underline{k})$ denote the number of ways of arranging m items into r non-empty groups whose sizes are given by $\underline{k}$. With ℓ_j denoting the number of $k_1, \ldots, k_r$ equal to j, one checks that, with $s = m - |\underline{k}|$ [40],

$$\mathcal{N}(m, \underline{k}) = \binom{m}{k_1 \ldots k_r \; s} \frac{1}{\prod_{j=2}^{m} \ell_j!} \tag{8}$$

(recall that $0! = \Gamma(1) = 1$). Now define [40], with $f(\boldsymbol{x}, m, \underline{k})$ and $g(\boldsymbol{x}, m)$ given by (7),

$$\lambda_{m,\underline{k}} := \int_{\Delta_{\mathbf{0}}} f(\boldsymbol{x}, m, \underline{k})\, \Xi_{\mathbf{0}}(d\boldsymbol{x}) + a \mathbb{1}_{(r=1, k_1=2)}, \tag{9}$$

$$\lambda_m := \int_{\Delta_{\mathbf{0}}} g(\boldsymbol{x}, m)\, \Xi_{\mathbf{0}}(d\boldsymbol{x}) + a\binom{m}{2} \tag{10}$$

if the integral in (10) exists, and

$$\lambda_m = \sum_{n=1}^{m-1} \sum_{\substack{k_1 \geq \ldots \geq k_r \geq 2 \\ m - |\underline{k}| + r = n}} \mathcal{N}(m, \underline{k}) \lambda_{m,\underline{k}} \tag{11}$$

otherwise. A continuous-time $\mathcal{P}_n$-valued Markov chain with transitions $q_{\pi,\pi'}$ given by [40]

$$q_{\pi,\pi'} = \begin{cases} \lambda_{m,\underline{k}} & \text{if } \pi' \prec_{m,\underline{k}} \pi, \\ -\lambda_m & \text{if } \pi' = \pi,\ \#\pi = m, \\ 0 & \text{otherwise;} \end{cases} \tag{12}$$

bioRxiv preprint doi: https://doi.org/10.1101/025684; this version posted February 6, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

is referred to as a Ξ-n-coalescent, and denoted by $\{\Pi_t^{(\Xi,n)}, t \geq 0\}$. The waiting time in state π is exponential with rate λ_m as in (10, 11).

Specific examples of Xi-coalescents

Out of the rich class of Xi-coalescents, several special cases have been identified either for biological / modeling relevance or mathematical tractability. The example most relevant for us is concerned with diploidy.

Haploid population models are probably the most common models in mathematical population genetics. Diploidy, and other forms of polyploidy, are, however, widely found in nature. Atlantic cod is diploid, and oysters show both tetraploidy and triploidy [cf. eg. 32]. In polyploid models which admit skewed offspring distribution, one should observe up to $M \geq 2$ simultaneous mergers, where M is some fixed number which reflects the level of polyploidy. Indeed, [34] model diploidy in a population with a general skewed offspring distribution, and obtain a Xi-coalescent, which admits simultaneous multiple mergers in up to four groups.

Indeed, a mathematical description of a Xi-coalescent which admits up to M simultaneous mergers is as follows. Take a finite measure Λ on $[0,1]$ (which would normally describe a Λ-coalescent). For convenience, let $F(dx) := \frac{\Lambda(dx)}{\Lambda([0,1])}$ be the corresponding normalized probability measure. Then, with $M \geq 2$, define the measure Ξ on the simplex Δ by

$$\Xi(d\mathbf{y}) := \frac{1}{M} \int_{[0,1]} \delta_{(\underbrace{\frac{x}{M}, \ldots, \frac{x}{M}}_{M \text{ times}}, 0, 0, \ldots)}(d\mathbf{y})\, F(dx). \tag{13}$$

The interpetation is this: If the normalized Lambda-measure F produces a multiple merger event, in which individual active ancestral lineages (blocks of the current partition) take part with probability $x \in (0,1]$, then the participating lineages are randomly grouped into M simultaneous mergers (each with probability $\frac{1}{M}$). Observe that if $F(dx) = \delta_0(x)dx$, then (13) becomes $\Xi(d\mathbf{y}) = \frac{1}{M}\int_{[0,1]} \delta_{(0,\ldots)}(d\mathbf{y})\delta_0(x)dx = \frac{1}{M}\delta_{(0,\ldots)}(d\mathbf{y})$, which corresponds to a Kingman-coalescent with time scaled by a factor $\frac{1}{M}$.

bioRxiv preprint doi: https://doi.org/10.1101/025684; this version posted February 6, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

For a given merger of the Ξ-coalescent into $r \in [M]$ groups of sizes given by $\underline{k} = (k_1, \ldots, k_r)$, with $k_1 \geq \ldots \geq k_r \geq 2$ and $|\underline{k}| := k_1 + \cdots + k_r$, when m active ancestral lineages are present, with $m - |\underline{k}|$ lineages unaffected by the given merger, the transition rates are given by

$$\lambda_{m,\underline{k}} = \sum_{\ell=0}^{(m-|\underline{k}|)\wedge(M-r)} \binom{m-|\underline{k}|}{\ell} (M)_{r+\ell}\, M^{-(|\underline{k}|+\ell)} \int_{[0,1]} x^{|\underline{k}|+\ell-2}(1-x)^{m-(|\underline{k}|+\ell)} F(dx), \tag{14}$$

where $(M)_{r+\ell}$ denotes the falling factorial. The rates (14) depend on the choice of the probability measure F determined by the underlying population model. A proof of (14) is in the Appendix. We say that a Xi-coalescent is *M-fold* if the transition rates are given by (14).

Xi-coalescents which admit at most $M = 4$ simultaneous mergers arise from diploid Cannings population models with skewed offspring distribution as shown in [34, 9], and are thus relevant for population genetics. In fact, if a haploid model (for example for mitochondrial DNA) is governed by a Λ-coalescent, the corresponding diploid model (concerning the core genome) might naturally lead to Xi-coalescents. Indeed, [9] derive a 4-fold Xi-coalescent from a diploid model, in which exactly one pair of diploid parents contribute diploid offspring in each reproduction event. Hence, since 4 parental chromosomes are involved in each event, one can observe up to 4 simultaneous mergers. This was also observed by [34] in association with a diploid population model, but under a more general reproduction law than considered by [9]. Xi-coalescents are also classified by [14] for a very general diploid exchangeable Cannings model in which arbitrary pairs of diploid parents contribute offspring in each reproduction event. This generalises the model by [34], in which each individual forms at most one parental pair in each generation. For a detailed classification of coalescent limits, see [34, 37, 33, 14].

In truly diploid models, as considered by [34, 9], selfing is excluded, which leads to a 'separation of timescales' phenomenon in the ancestral process, in which blocks which reside

bioRxiv preprint doi: https://doi.org/10.1101/025684; this version posted February 6, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

in the same diploid individual instantaneously 'disperse'; thus the configuration of blocks in diploid individuals becomes irrelevant in the ancestral process (see Cor. 4.3 in [34]).

A natural candidate for F may be the beta distribution with parameters $\vartheta > 0$ and $\gamma > 0$ (cf. e.g. [9]), with density

$$\frac{\Gamma(\vartheta+\gamma)}{\Gamma(\vartheta)\Gamma(\gamma)} x^{\vartheta-1}(1-x)^{\gamma-1}, \quad x \in [0,1].$$

In this case, the rate $\lambda_{m,\underline{k}}$ in (14) takes the form

$$\lambda_{m,\underline{k}} = \sum_{\ell=0}^{(m-|\underline{k}|)\wedge(4-r)} \binom{m-|\underline{k}|}{\ell} (4)_{r+\ell}\, 4^{-|\underline{k}|+\ell}\, \frac{B(|\underline{k}|+\ell+\vartheta-2, m+\gamma-(|\underline{k}|+\ell))}{B(\vartheta,\gamma)}. \tag{15}$$

A different choice is based on a model of Eldon and Wakeley [23], where

$$\Lambda(dx) = F(dx) = \frac{2}{2+\psi^2}\delta_0(dx) + \frac{\psi^2}{2+\psi^2}\delta_\psi(dx), \quad \psi \in [0,1].$$

In this case, the rates reduce to

$$\lambda_{m,\underline{k}} = \frac{1}{2+\psi^2} \sum_{l=0}^{(m-|\underline{k}|)\wedge(4-r)} \binom{m-|\underline{k}|}{\ell} (4)_{r+\ell}\, 4^{-|\underline{k}|+\ell}\, (1-\psi)^{m-(|\underline{k}|+\ell)} \psi^{|\underline{k}|+\ell} + \mathbb{1}_{(r=1,\, k_1=2)}\, \frac{1}{2}\frac{1}{2+\psi^2}. \tag{16}$$

In (16) the parameter ψ has a clear biological interpretation as the fraction of the diploid population replaced by the offspring of the reproducing parental pair in one generation. The interpretation of the parameters in (15) is perhaps less clear.

A multi-loci ancestral recombination graph in which simultaneous mergers are admitted is obtained by [9] in which the framework of [34] is borrowed. There, one can think of the reproduction model as a two-atom Lambda-measure, one atom at zero, and another at some point $\psi \in (0,1)$. If the atom at ψ has mass of order at most N^{-2}, the limit process admits simultaneous mergers. The order N^{-2} represents the order of the expected time which two gene copies need to coalesce when only 1 diploid offspring is produced in each reproduction event. In [9], complete dispersion of chromosomes also occurs, and the configuration of blocks

bioRxiv preprint doi: https://doi.org/10.1101/025684; this version posted February 6, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

among diploid individuals becomes irrelevant in the limit process.

Xi-coalescents can also be obtained from a population model where the population size varies substantially due to recurrent bottlenecks. This has been introduced and discussed in [11], who obtain a randomly time-changed Kingman coalescent, which thus yields a Xi-coalescent.

Durrett and Schweinsberg [19, 20] show that a Xi-coalescent gives a good approximation [20, cf. Prop. 3.1] to the genealogy of a locus subject to recurrent beneficial mutations.

These examples suggest that Xi-coalescents form an important class of mathematical objects with which to study genetic diversity.

The site-frequency spectrum

The site-frequency spectrum is a simple summary statistic of the full DNA sequence data, but contains valuable information about variation among individuals. We assume the infinitely-many sites mutation model [46], in which mutations occur as independent Poisson processes on the branches of a given gene genealogy with rate θ (or $\theta/2$) for some constant $\theta > 0$, and no two mutations occur at the same site. The constant θ is determined by the ratio μ/c_N, where μ is the per-generation mutation rate, and c_N is the probability of two distinct individuals (gene copies) sharing a common ancestor in the previous generation. We refer to [22] for a discussion of the relation between mutation and timescales of different coalescent processes.

Given sample size n, we let $\xi_i^{(n)}$ denote the number of polymorphic sites at which one variant (the derived mutation) is observed in i copies. The random vector

$$\underline{\xi}^{(n)} = \left(\xi_1^{(n)}, \ldots, \xi_{n-1}^{(n)}\right) \tag{17}$$

is known as the (unfolded) *site-frequency spectrum*. If information about ancestral states is unavailable, so that one does not know which variant is new, one considers the *folded*

bioRxiv preprint doi: https://doi.org/10.1101/025684; this version posted February 6, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

spectrum $\underline{\eta}^{(n)} = \left(\eta_1^{(n)}, \ldots, \eta_{\lfloor \frac{n}{2} \rfloor}^{(n)}\right)$ in which

$$\eta_i^{(n)} = \begin{cases} \xi_i^{(n)} + \xi_{n-i}^{(n)} & \text{if } i < \frac{n}{2}, \\ \xi_i^{(n)} & \text{if } i = \frac{n}{2}. \end{cases} \tag{18}$$

[24] obtains closed-form solutions for expected values and (co)-variances of the site-frequency spectrum associated with the Kingman coalescent $\Pi^{(K)}$. Indeed [24],

$$\mathbb{E}^{(K)}\left[\xi_i^{(n)}\right] = \frac{\theta}{i}, \quad i \in [n-1], \tag{19}$$

where $\mathbb{E}^{(K)}$ denotes expectation with respect to Kingman coalescent.

Let $B_i^{(n)}$ denote the random total length of branches subtending $i \in [n-1]$ leaves. By leaves we refer to the special kind of vertices in the random graph generated by a given coalescent process which would represent the sampled DNA sequences after mutations have been added to the graph. Since the mutation process can be separated from the genealogy (the random graph) it can sometimes be useful to consider properties of the graph itself. Result (19) follows from [24]

$$\mathbb{E}^{(K)}\left[B_i^{(n)}\right] = \frac{2}{i}, \quad i \in [n-1], \tag{20}$$

and the infinitely-many sites mutation model.

Due to the multiple merger property of Λ- and Ξ-coalescents, closed-form expressions for $\mathbb{E}^{(\Pi)}\left[\xi_i^{(n)}\right]$ for $\Pi \in \{\Lambda, \Xi\}$ are quite hard to obtain. A key quantity in computing $\mathbb{E}^{(\Pi)}\left[B_i^{(n)}\right]$ for multiple merger coalescents is

$$p^{(n)}[k, i], \qquad i \in \{1, \ldots, n-k+1\},$$

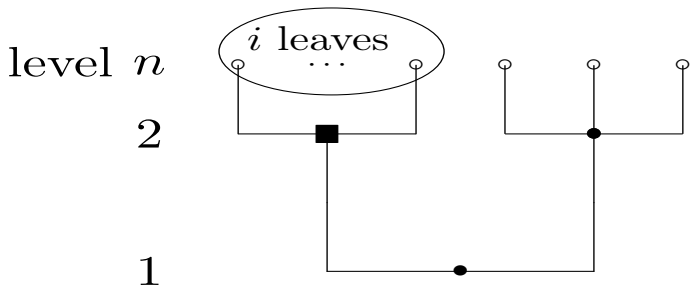
which can be described as the probability that starting from n blocks, conditioned that there are at some point in time exactly k blocks, one of them, sampled uniformly at random,

bioRxiv preprint doi: https://doi.org/10.1101/025684; this version posted February 6, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

subtends $i \in [n-k+1]$ leaves. See Figure 1 for an illustration.

bioRxiv preprint doi: https://doi.org/10.1101/025684; this version posted February 6, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 1: An illustration of an event with probability $p^{(n)}[k,i]$, with $k=2$, and $2 \leq i = n-3$ for n leaves (shown as open circles). In the example shown, the process reaches $k=2$ blocks in one transition, which is a simultaneous merger involving all of the i leaves (encircled) as shown. All the i leaves are subtended by the block represented by a black square. A 'level' refers to the values taken by the block-counting process; filled symbols represent ancestral blocks.



bioRxiv preprint doi: https://doi.org/10.1101/025684; this version posted February 6, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

With $g(n,k)$ we denote the expected length of time during which we see $k \in \{2,\dots,n\}$ blocks, given that we started from $n \geq 2$ blocks. Given $p^{(n)}[k,i]$ and $g(n,k)$, $\mathbb{E}^{(\Pi)}\left[B_i^{(n)}\right]$ can be computed as follows:

$$\mathbb{E}^{(\Pi)}\left[B_i^{(n)}\right] = \sum_{k=2}^{n-i+1} p^{(n)}[k,i] \cdot k \cdot g(n,k), \tag{21}$$

where moreover $g(n,k)$ can be computed recursively. This is shown in [10] for Λ-coalescents but in fact holds for Ξ-coalescents as well. Hence, it suffices to obtain a recursion for $p^{(n)}[k,i]$. For the Λ-case, [10] obtain the recursion

$$\begin{aligned} p^{(n)}[k,i] = \sum_{m=k}^{n-1} p_{n,m} \frac{g(m,k)}{g(n,k)} \Bigg[& \frac{i-n+m}{m} p^{(m)}[k,i-n+m] \mathbb{1}_{(i>n-m)} \\ & + \frac{m-i}{m} p^{(m)}[k,i] \mathbb{1}_{(i<m)} \Bigg] \end{aligned} \tag{22}$$

in which $p_{n,m}$ is the probability that the block-counting process jumps from n to m blocks.

Before we turn to the expected site-frequency spectrum associated with Ξ-coalescents, we give the recursion to compute $g(n,k)$.

Lemma 1. *Let $p_{n,m}$ denote the probability that the block-counting process associated with a Ξ-n-coalescent with transition rates (12) jumps from $n \geq 2$ to $m \in [n-1]$ blocks. For any $n > k \geq 1$, we have*

$$g(n,k) = \sum_{m=k}^{n-1} p_{n,m}\, g(m,k), \tag{23}$$

with the boundary condition

$$g(n,n) = \lambda_n^{-1} \tag{24}$$

for any $n \geq 2$, where $\lambda_n = -q_{\pi,\pi}$ with $\#\pi = n$, see (10, 11, 12).

A proof of Lemma 1 is given in the Appendix.

bioRxiv preprint doi: https://doi.org/10.1101/025684; this version posted February 6, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The expected site-frequency spectrum associated with Ξ-coalescents

Since formula (21) holds for any exchangeable coalescent, it suffices to obtain a recursion for $p^{(n)}[k,i]$ when associated with Ξ-coalescents. Before we state the main result, we review our notation for partitions of positive integers. Partitions of integers are helpful in enumerating the different ways in which the number of active blocks can change in one transition in a Xi-coalescent.

A *partition* of $n \in \mathbb{N}$ is a non-increasing sequence of positive integers whose sum is n. By $\tilde{n}$ we denote the set of all partitions of n, we denote by ν a generic element of $\tilde{n}$. By way of example,

$$\tilde{3} = \{(3), (2,1), (1,1,1)\}.$$

If $\nu \in \tilde{n}$, we write $|\nu| = n$. The *size* of a partition ν is defined as the length of the sequence, and is denoted by $\#\nu$. If $|\nu| = n$ with size $\#\nu = k \in [n]$, we write $|\nu| \overset{k}{=} n$. Thus, $|\nu| \overset{1}{=} n$ if and only if $\nu = (n)$; $|\nu| \overset{n}{=} n$ if and only if $\nu = (\underbrace{1, \ldots, 1}_{n \text{ times}})$.

Another way of representing an integer partition ν is by specifying how often each positive integer $i \in \mathbb{N}$ appears in ν (see [15] for details). Thus, we will also denote ν by $\langle \alpha_1, \alpha_2, \ldots \rangle$, where α_i denotes the number of times integer i appears in the given partition ν. A partition $\varrho = \langle \beta_1, \ldots \rangle$ is a *sub-partition* of $\nu = \langle \alpha_1, \ldots \rangle$, denoted $\varrho \subset \nu$, if and only if $\beta_i \leq \alpha_i$ for all i. For a set partition $\pi \in \mathcal{P}_n$, we define the *integer partition associated with* π, denoted $\pi^{\downarrow} \in \tilde{n}$, as the partition of n obtained by listing the block sizes of π in decreasing order. More detailed discussion of partitions of integers can be found eg. in [44].

The role of integer partitions in association with Ξ-coalescents should now be clear. We can enumerate all the possible ways the block counting process can jump from n to m active blocks by specifying the partitions of n (see eg. [15] for details). The elements of the sequence $\nu \in \tilde{n}$ specify the merger sizes, with the obvious exclusion of mergers of size 1. By way of example, integer partition $(3,2,1) \in \tilde{6}$ specifies a simultaneous merger of 3 blocks and 2 blocks, and one block remains unchanged, when we have 6 active blocks. By (12), any such transition happens at rate $\lambda_{6,(3,2)}$.

bioRxiv preprint doi: https://doi.org/10.1101/025684; this version posted February 6, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

More generally, given integer partition $|\nu| \overset{m}{=} n$ with $\nu = \langle \alpha_1, \alpha_2, \ldots \rangle$, put $r := m - \alpha_1$ for the number of elements of the sequence that are larger than 1, so that

$$\nu = (\nu_1, \nu_2, \ldots, \nu_r, \underbrace{1, \ldots, 1}_{\alpha_1 \text{ times}}).$$

Then $\nu_1 \geq \nu_2 \geq \cdots \geq \nu_r \geq 2$, and defining $\underline{k} := (\nu_1, \nu_2, \ldots, \nu_r)$, the corresponding transitions in which $\underline{k}$ specifies the sizes of the r mergers involved happen at rate $\lambda_{n,\underline{k}}$, see (12). Moreover, the probability $p_\nu^{(n)}$ of such a transition is given by [15, Lemma 2.2.2]

$$p_\nu^{(n)} = \frac{n!}{\prod_i \nu_i! \alpha_i!} \frac{\lambda_{n,\underline{k}}}{\lambda_n}, \tag{25}$$

where the combinatorial factor

$$\frac{n!}{\prod_i \nu_i! \alpha_i!}$$

(coinciding with $\mathcal{N}(n, \underline{k})$ as introduced in (8)) denotes the number of different ways of merging n blocks in r groups specified by $\underline{k}$, and $\lambda_{n,\underline{k}}$, λ_n are given by (9, 10, 11).

Now we state our main theorem, which contains the recursion for $p^{(n)}[k,i]$ needed to compute $\mathbb{E}^{(\Xi)}\left[B_i^{(n)}\right]$. The theorem holds for all Ξ-n-coalescents whose block-counting process visits every possible state with positive probability, which is true of all examples of Xi-coalescents that we consider. The assumption is not very restrictive, since it excludes only degenerate cases where the measure Ξ is concentrated on a vertex of the infinite simplex Δ_0, see Proposition 6 in the Appendix for a precise formulation.

Theorem 2. *[15] Let $\{\Pi_t^{(\Xi,n)}, t \geq 0\}$ be a Ξ-n-coalescent with transition rates (12) such that the corresponding block-counting process hits every $k \in [n]$ with positive probability. Then, for $2 \leq k \leq n$ and $1 \leq i \leq n-k+1$, we have*

$$p^{(n)}[k,i] = \sum_{m=k}^{n-1} \frac{g(m,k)}{g(n,k)} \sum_{\substack{|\nu| \overset{m}{=} n \\ \nu = \langle \alpha_1, \ldots \rangle}} p_\nu^{(n)} \sum_{j=1}^{i \wedge (m-k+1)} \sum_{\substack{\varrho \subset \nu \\ |\varrho| \overset{j}{=} i \\ \varrho = \langle \beta_1, \ldots \rangle}} p^{(m)}[k,j] \frac{\prod_\ell \binom{\alpha_\ell}{\beta_\ell}}{\binom{m}{j}}, \tag{26}$$

bioRxiv preprint doi: https://doi.org/10.1101/025684; this version posted February 6, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

with the boundary cases $p^{(n)}[n,i] = \mathbb{1}_{(i=1)}$.

A proof is provided in the Appendix, along with a necessary and sufficient condition for the block-counting process to hit state $k \in [n]$ with positive probability (see Proposition 6).

The recursion (26) for $p^{(n)}[k,i]$ simplifies when one restricts to Ξ-coalescents with at most $M \geq 2$ simultaneous mergers ($M = 1$ simply gives a Lambda-coalescent), as shown in Cor. 3. Before we state Cor. 3, we briefly review *ordered mergers*. The computations are more efficient for Xi-coalescents restricted to at most M simultaneous mergers, and when one considers ordered mergers rather than partitions of integers. Let

$$\mathcal{M}_M(n,m) := \{\underline{k} = (k_1, \ldots, k_r) : r \in [M], k_i \in \mathbb{N}, k_1 \geq \ldots \geq k_r \geq 2, m = n - |\underline{k}| + r\} \quad (27)$$

denote the set of single and up to M simultaneous ordered mergers by which the block-counting process can jump from $n \geq 2$ to $m \in [n-1]$ blocks in $r \in [M]$ mergers. The set $\mathcal{M}_M(n,m)$ corresponds to the set of all integer partitions $|\nu| \stackrel{m}{=} n$ such that, if $\nu = \langle \gamma_1, \gamma_2, \ldots \rangle$, we have $\sum_{j \geq 2} \gamma_j = m - \gamma_1 \in [M]$. Indeed, for $1 \leq m < n$, we have a bijection

$$\{\nu = \langle \gamma_1, \gamma_2, \ldots \rangle : |\nu| \stackrel{m}{=} n, m - \gamma_1 \leq M\} \ni \nu \mapsto (\nu_1, \ldots, \nu_{r_\nu}) \in \mathcal{M}_M(n,m),$$

with $r_\nu := \max\{j : \nu_j \geq 2\} = m - \gamma_1$, for $\nu = \langle \gamma_1, \gamma_2, \ldots \rangle$. Obviously, the inverse bijection is given by

$$\mathcal{M}_M(n,m) \ni \underline{k} \mapsto \nu := (k_1, k_2, \ldots, k_r, \underbrace{1, \ldots, 1}_{m-r \text{ times}}).$$

Informally, ordered mergers are just integer partitions where all elements equal to one are omitted.

For ease of presentation, we will also write $\mathcal{M}_M(n,m) \ni \mu = (\alpha) \equiv (\alpha_2, \alpha_3, \ldots)$ where α_j denotes the number of occurrences of mergers of size j in μ. By $\gamma = (\beta) \subset \mu = (\alpha)$ we denote a *submerger* γ of μ where $\beta_j \leq \alpha_j$ for all j, *including* the case $\beta_j = 0$. Finally, the *size* of $\mu = \underline{k} \in \mathcal{M}_M(n,m)$ is just the length of the sequence $\underline{k}$, i.e. $\#\mu = r$ for $\mu = \underline{k} = (k_1, \ldots, k_r)$.

bioRxiv preprint doi: https://doi.org/10.1101/025684; this version posted February 6, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

If μ is the ordered merger corresponding to some integer partition $\nu = \langle \gamma_1, \gamma_2, \ldots \rangle$, $|\nu| \overset{m}{=} n$ as above, then $\#\nu = m = \#\mu + \gamma_1$.

Corollary 3. *Let $\{\Pi_t^{(\Xi,n)}, t \geq 0\}$ be a Ξ-n-coalescent with transition rates (12) such that at most $M \geq 2$ simultaneous mergers are possible, and such that the corresponding block-counting process hits every $k \in [n]$ with positive probability. Then, for $2 \leq k \leq n$ and $1 \leq i \leq n-k+1$, we have*

$$p^{(n)}[k,i] = \sum_{m=k}^{n-1} \frac{g(m,k)}{g(n,k)} \sum_{\substack{\mu \in \mathcal{M}_M(n,m) \\ \mu = (\alpha)}} p_\mu^{(n)} \cdot \sum_{j=1}^{i \wedge (m-k+1)} \sum_{\substack{\gamma \subset \mu \\ \gamma \in \mathcal{M}_M(i,j) \\ \gamma = (\beta)}} p^{(m)}[k,j] \binom{m}{j}^{-1} \prod_{\ell \geq 2} \binom{\alpha_\ell}{\beta_\ell} \binom{m-\#\mu}{j-\#\gamma}, \tag{28}$$

with the boundary cases $p^{(n)}[n,i] = \mathbb{1}_{(i=1)}$.

Informally, the first sum in recursion (28) is over the number of blocks the block-counting process can jump to, given that it starts in n, and conditioned on it hits k. The second sum is over all (up to M simultaneous) mergers (μ) in which one can jump from n to m blocks, and the last sum is over all the ways mergers involving the i leaves can be nested within each given merger μ.

Corollary 3 is simply another way of representing $p^{(n)}[k,i]$ (26) (see Thm. 2) in terms of ordered mergers. The switch in focus to ordered mergers from integer partitions obliviates the need to keep track of all the 1s. Also, since we restrict to Ξ-coalescents which admit at most M simultaneous mergers, the required mergers can be generated quite efficiently.

The computation of $p^{(n)}[k,i]$ can be checked by noting that for each fixed $n \geq 2$, and with $2 \leq k \leq n$, $\sum_i p^{(n)}[k,i] = 1$. One can also compute $\mathbb{E}^{(\Xi)}\left[B_1^{(n)}\right]$ with a simple recursion as follows. Let $\mathbb{E}^{(\Xi)}\left[B_1^{(n,i)}\right]$ denote the expected length of i external branches, given $n \geq 2$ active blocks; ie. when we start from n active blocks, i of which are singleton blocks. Define,

bioRxiv preprint doi: https://doi.org/10.1101/025684; this version posted February 6, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

for all $n \in \mathbb{N}$, $n \geq 2$,

$$\begin{aligned} \mathcal{C}_n :&= \{\underline{k} = (k_1, \ldots, k_r) : r \in [M], k_i \in \mathbb{N}, k_1 \geq \ldots \geq k_r \geq 2, |\underline{k}| \leq n\} \\ &= \bigcup_{m=1}^{n-1} \mathcal{M}_M(n, m) \end{aligned} \tag{29}$$

as the set of all ordered (up to M simultaneous) mergers given n active blocks, where $\mathcal{M}_M(n, m)$ was defined in (27).

Lemma 4. *Let $\{\Pi_t^{(\Xi,n)}, t \geq 0\}$ be a Ξ-n-coalescent with transition rates (12) such that at most M simultaneous mergers are possible. With λ_n given by (10, 11), let $\mathcal{C}_n$ be given by (29), and let $p_{\underline{k}}^{(n)}$ denote the probability of merger $\underline{k}$ when the number of active blocks is n. Then,*

$$\mathbb{E}^{(\Xi)}\left[B_1^{(n,i)}\right] = \frac{i}{\lambda_n} + \sum_{\underline{k} \in \mathcal{C}_n} p_{\underline{k}}^{(n)} \sum_{j=(i-n+|\underline{k}|)^+}^{i \wedge |\underline{k}|} \binom{n}{|\underline{k}|}^{-1} \binom{i}{j} \binom{n-i}{|\underline{k}|-j} \mathbb{E}^{(\Xi)}\left[B_1^{(n-|\underline{k}|+r, i-j)}\right] \tag{30}$$

with the boundary condition $\mathbb{E}^{(\Xi)}\left[B_1^{(n,0)}\right] = 0$.

Numerical results

Expected branch lengths $\mathbb{E}^{(\Pi)}\left[B_i^{(n)}\right]$

Under the infinitely many sites mutation model, the expected site-frequency spectrum is given by $\mathbb{E}^{(\Pi)}\left[\xi_i^{(n)}\right] = \theta\, \mathbb{E}^{(\Pi)}\left[B_i^{(n)}\right]$ where $\theta > 0$ is the appropriately scaled mutation rate. Hence, it suffices to consider $\mathbb{E}^{(\Pi)}\left[B_i^{(n)}\right]$ in a comparison of the site-frequency spectrum associated with different coalescent models. In Figure 2, we consider the 4-fold Ξ-coalescent, when the measure F in (14) is associated with the beta-density, with $\alpha \in [1, 2)$ [41],

$$F(dx) = \frac{x^{1-\alpha}(1-x)^{\alpha-1}}{B(2-\alpha, \alpha)} dx, \quad x \in [0, 1]. \tag{31}$$

bioRxiv preprint doi: https://doi.org/10.1101/025684; this version posted February 6, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The range of α is the interval $[1, 2)$, since for $\alpha \in [1, 2)$, one obtains a Lambda-Beta coalescent from a supercritical branching population model [41]. We do not have a microscopic diploid population model which explicitly yields a Xi-Beta coalescent. However, the results of [34] indicate that such a process should exist. Also, the Lambda-Beta coalescent is one of the most studied examples of Lambda-coalescents. However, the existence of the Xi-Dirac coalescent was proved by [9]. The expected branch lengths associated with a Lambda-coalescent with F-measure (31), as well as the Kingman coalescent, are also shown in Figure 2 for comparison. We consider the normalised expected spectrum

$$\varphi_i^{(n,\Pi)} := \frac{\mathbb{E}^{(\Pi)}\left[B_i^{(n)}\right]}{\mathbb{E}^{(\Pi)}\left[B^{(n)}\right]}, \quad i \in [n-1], \tag{32}$$

in which $\mathbb{E}^{(\Pi)}\left[B^{(n)}\right]$ is the expected total size of the genealogy $\left(B^{(n)} := B_1^{(n)} + \cdots + B_{n-1}^{(n)}\right)$. Define $R_i^{(n)} := \frac{B_i^{(n)}}{B^{(n)}}$. Let $\xi^{(n)} := \xi_1^{(n)} + \cdots + \xi_{n-1}^{(n)}$ denote the (random) total number of segregating sites, and define the normalised spectrum $\zeta_i^{(n)} := \frac{\xi_i^{(n)}}{\xi^{(n)}}$ (with $\zeta_i^{(n)} \equiv 0$ if $\xi^{(n)} = 0$). The reasons for our preference for $\varphi_i^{(n,\Pi)}$ over $\mathbb{E}^{(\Pi)}\left[\xi_i^{(n)}\right]$ are the following. The quantity $\varphi_i^{(n,\Pi)}$, which is an approximation of the expected normalised spectrum $\mathbb{E}^{(\Pi)}\left[\zeta_i^{(n)}\right]$ is, clearly, *not* a function of the mutation rate θ. The expected normalised spectrum $\mathbb{E}^{(\Pi)}\left[\zeta_i^{(n)}\right]$ is well approximated by $\mathbb{E}^{(\Pi)}\left[R_i^{(n)}\right]$, and is also quite robust to changes in mutation rate, if the mutation rate is not very small [22]. Since $\varphi_i^{(n,\Pi)}$ is a decent approximation of $\mathbb{E}^{(\Pi)}\left[R_i^{(n)}\right]$ (results not shown), $\varphi_i^{(n,\Pi)}$ is therefore a good approximation of $\mathbb{E}^{(\Pi)}\left[\zeta_i^{(n)}\right]$. One can therefore use $\varphi_i^{(n,\Pi)}$ to estimate coalescent parameters, for example by minimising sum-of-squares, without the need to jointly estimate the mutation rate.

Refer to a Xi-coalescent with marginal (single-locus) rate described by Eq. (14) in [9] as the Xi-Dirac-Kingman coalescent. This corresponds to a process with coalescent rates given by (14), where the measure F has an atom at some $\psi \in (0, 1)$, but there is an additional atom at 0, which represents the 'Kingman component'. This process was derived by [9] in connection with ancestral recombination graphs involving multiple loci, in which simultaneous multiple mergers are admitted. In some cases, even for ψ close to 1, the

bioRxiv preprint doi: https://doi.org/10.1101/025684; this version posted February 6, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

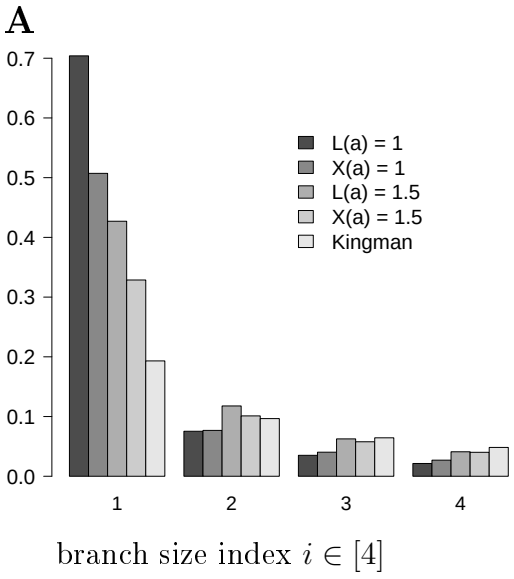
associated $\varphi_i^{(n,\Pi)}$ is very similar to $\varphi_i^{(n,K)}$ (results not shown).

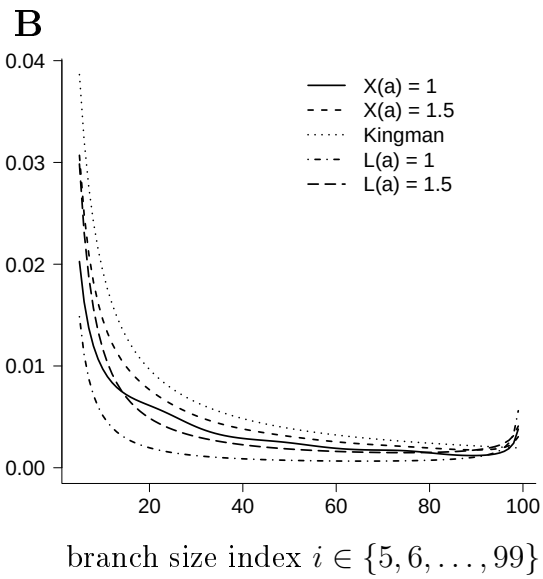
In Figure 2 the $\varphi_i^{(n,\Pi)}$ for corresponding Lambda-Beta and Xi-Beta coalescents are compared. As Figure 2 shows, the two processes predict different patterns of the site-frequency spectrum, at least for $\alpha \in \left[1, \frac{3}{2}\right]$. Both processes can predict a significant excess of singletons relative to the Kingman coalescent.

Similar conclusions can be reached from Figure 3, in which the measure F is given by the Dirac-measure $F(dx) = \delta_\psi(x)dx$ for some $\psi \in [0,1]$ (no Kingman component). The Xi-Dirac-coalescent, for $\psi = 0.95$, displays a multimodal graph of $\varphi_i^{(n,\Pi)}$, but the relative 'height' of the modes is small ($< 1\%$ of the total expected length). As well as an excess of singleton polymorphisms, [2] observe multi-modality, or small 'bumps', in the site-frequency spectrum associated with the *Ckma* gene in Atlantic cod.

bioRxiv preprint doi: https://doi.org/10.1101/025684; this version posted February 6, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

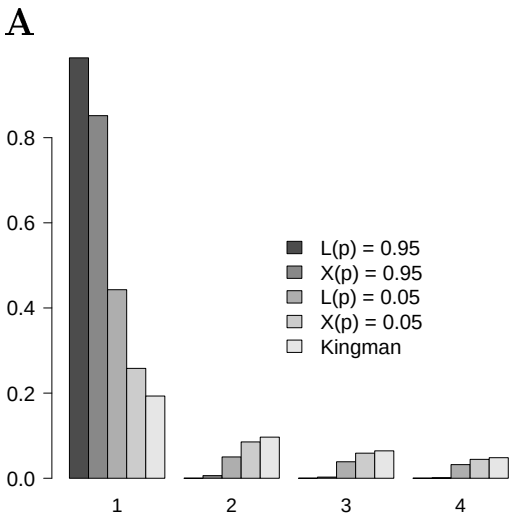
Figure 2: The relative expected lengths $\varphi_i^{(n,\Pi)}$ (32) for $n = 100$ and coalescent process Π as follows: `L(a)` denotes the Lambda-Beta coalescent with parameter value `a` (α) as shown, and `X(a)` denotes the Xi-Beta coalescent with parameter value `a` (α) as shown. In **A**, only the first 4 classes are shown; in **B**, the remaining classes.

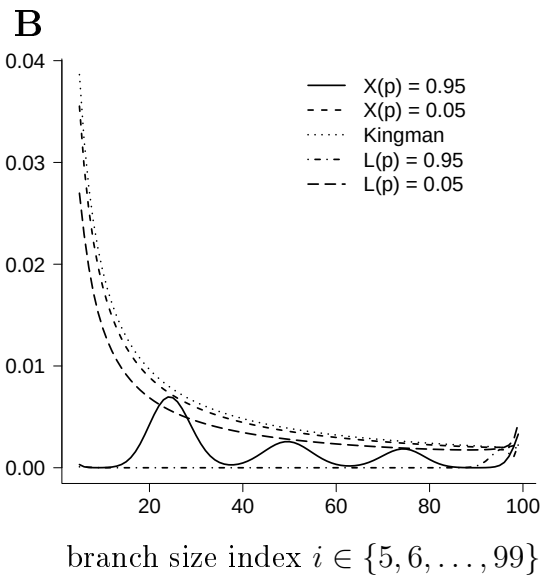




bioRxiv preprint doi: https://doi.org/10.1101/025684; this version posted February 6, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 3: The relative expected lengths $\varphi_i^{(n,\Pi)}$ (32) for $n = 100$ and coalescent process Π as follows: `L(p)` denotes the Lambda-Dirac coalescent with parameter value `p` (ψ) as shown, and `X(p)` denotes the Xi-Dirac coalescent with parameter value `p` (ψ) as shown. In **A**, only the first 4 classes are shown; in **B**, the remaining classes.





bioRxiv preprint doi: https://doi.org/10.1101/025684; this version posted February 6, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Coalescent parameter estimates

As Figures 2 and 3 indicate, Xi- and Lambda-coalescents predict different site-frequency spectra. In Table 2 we record parameter estimates obtained when the 'data' are branch lengths B_i simulated under either a 4-fold Xi-coalescent (14) or a Lambda-coalescent (5) with parameter values as shown. However, we consider an extensive comparison of different examples of the large class of Xi-coalescent beyond the scope of the current work. Let ϑ denote a generic coalescent parameter. If $0 \leq \vartheta < 1$, $\Xi(\vartheta)$ denotes a 4-fold Xi-Dirac coalescent, with F-measure $F(dx) = \delta_\vartheta(x)dx$ in (14), and $\Lambda(\vartheta)$ a Lambda-Dirac coalescent. If $\vartheta \in [1, 2)$, $\Xi(\vartheta)$ denotes a 4-fold Xi-Beta coalescent, and $\Lambda(\vartheta)$ a Lambda-Beta coalescent. Estimates of ϑ attributed to coalescent process Π_2 are obtained with an ℓ_2 norm applied to the normalised lengths $R_i := \frac{B_i}{B}$ (where we drop the superscript '(n)') drawn from coalescent process Π_1, and $\varphi_i^{(n,\Pi_2)}$,

$$\ell_2(\Pi_1, \Pi_2) = \sqrt{\sum_{i=1}^{n-1} \left(R_i - \varphi_i^{(n,\Pi_2)}\right)^2}. \tag{33}$$

If the B_i are drawn from a Xi-Dirac coalescent (Π_1), we estimate ϑ associated with a Lambda-Dirac coalescent (Π_2), and vice versa. If the B_i are drawn from a Xi-Beta coalescent (Π_1), we estimate ϑ associated with a Lambda-Beta coalescent (Π_2), and vice versa. The ℓ_2 norm (33) is appealing since it makes no assumptions about the distribution of R_i. In contrast, the Poisson Random Field (PRF) model [39] assumes Poisson distribution of mutation counts. Indeed, [16] prove that the law of the joint site-frequency spectrum converges, in the limit of infinite sample size, to that of independent Poissons, when associated with the Kingman coalescent. The asymptotic results of [4, 6, 5] show that multiple-merger coalescents belong to completely different asymptotic regimes than the PRF model assumes.

A more suitable distance statistic, similar to the G_ξ statistic suggested by [25, Eq. (8)], could be

$$d(\Pi_1, \Pi_2) = \sqrt{\sum_{i=1}^{n-1} \frac{\left(R_i - \mathbb{E}^{(n,\Pi_2)}[R_i]\right)^2}{\mathbb{V}^{(n,\Pi_2)}[R_i]}}, \tag{34}$$

where $\mathbb{V}^{(n,\Pi_2)}[R_i]$ denotes the variance of R_i computed with respect to Π_2. However, we can

bioRxiv preprint doi: https://doi.org/10.1101/025684; this version posted February 6, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

neither represent $\mathbb{V}^{(n,\Pi_2)}[R_i]$ nor $\mathbb{E}^{(n,\Pi_2)}[R_i]$ as simple functions of ϑ or n. In actual applications, one would replace R_i in (34) with $\zeta_i := \frac{\xi_i}{\xi}$, the normalised site-frequency spectrum (normalised by the total number of segregating sites ξ).

As Table 2 shows, a Lambda-Dirac coalescent underestimates ψ when the data are generated by a Xi-Dirac coalescent, and Lambda-Beta overestimates α when the data are generated by a Xi-Beta coalescent. When we switch the generation of data from Xi- to Lambda-coalescents, we reach the opposite conclusions. A Xi-Dirac coalescent overestimates the parameter (ψ) when the data are generated by a Lambda-Dirac coalescent, and the Xi-Beta coalescent underestimates α when the data are generated by a Lambda-Beta coalescent. The ℓ_2 norm (33) recovers the true parameter values reasonably well when the model is correctly specified (Table 4).

The difference between the corresponding Xi- and Lambda-coalescents are further illustrated in Figure 4, in which the distance between the normalised expected spectra $\varphi_i^{(n,\Pi)}$ (32), with Π as shown, is quantified by the ℓ_2 norm (33). The graphs in Figure 4 show clearly that even when the parameters associated with the corresponding Xi- and Lambda-coalescents are the same, the difference in $\varphi_i^{(n,\Pi)}$ can be substantial (except of course when the Lambda-coalescent is the Kingman coalescent; which happens when $\alpha = 2$ or $\psi = 0$).

The difference in estimates between corresponding Lambda- and Xi-coalescents may be understood from the way the Xi-coalescent process is constructed. Indeed, given k blocks drawn from the associated Lambda-coalescent, we see one k-merger with probability 4^{1-k}, which quickly becomes small as k increases. A much more likely outcome is for the blocks to become (evenly) distributed into four groups. The effect on the genealogy of drawing a large number k is thus reduced in a Xi-coalescent relative to the corresponding Lambda-coalescent.

C code written for the computations is available at `http://page.math.tu-berlin.de/~eldon/programs.html`.

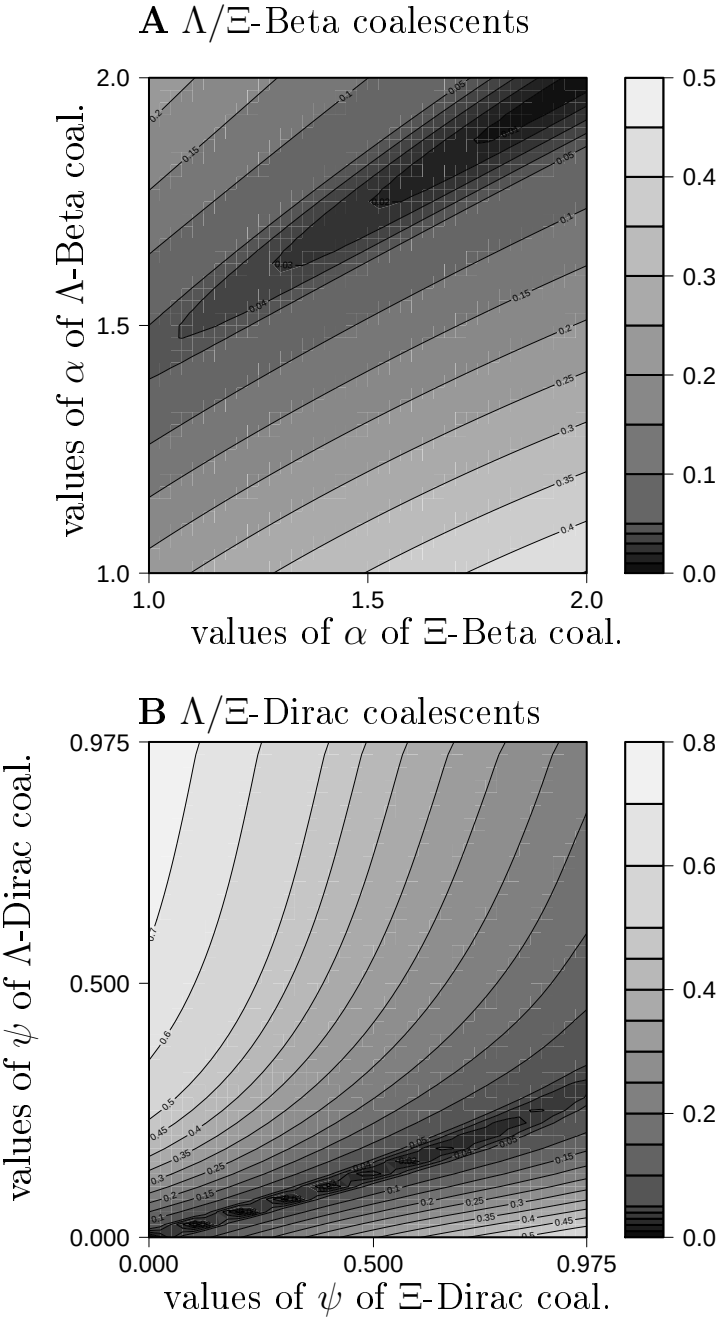
bioRxiv preprint doi: https://doi.org/10.1101/025684; this version posted February 6, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 2: Estimates of coalescent parameter ϑ, obtained by the use of the ℓ_2 norm (33), when the 'data' (branch lengths of a realized genealogy) are obtained by a Lambda- ($\Lambda(\vartheta)$) or a Xi-coalescent ($\Xi(\vartheta)$) as shown. In the top half, the 'data' is generated by a 4-fold Xi-coalescent with parameter value as shown, and a parameter estimate obtained for the corresponding Lambda-coalescent. In the bottom half, branch lengths are generated from a Lambda-coalescent with parameter values as shown, and parameter estimates obtained for the corresponding 4-fold Xi-coalescent. Parameter estimates (mean $\overline{\vartheta}$; standard deviation $\widehat{\vartheta}$) obtained for $n = 50$ leaves from 10^4 replicates. Estimates were obtained over the grids $\{0.0, 0.05, \ldots, 0.95\}$ and $\{1.0, 1.05, \ldots, 1.95\}$.

$\Pi(\vartheta)$	$\overline{\vartheta}$	$\widehat{\vartheta}$
$\Xi(0.05)$	0.02	0.029
$\Xi(0.95)$	0.30	0.225
$\Xi(1.0)$	1.44	0.257
$\Xi(1.5)$	1.70	0.208
$\Lambda(0.05)$	0.23	0.144
$\Lambda(0.95)$	0.95	0.035
$\Lambda(1.0)$	1.01	0.041
$\Lambda(1.5)$	1.20	0.256

bioRxiv preprint doi: https://doi.org/10.1101/025684; this version posted February 6, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 4: The distance between $\varphi_i^{(n,\,\Xi\text{-Beta})}$ (32) and $\varphi_i^{(n,\,\Lambda\text{-Beta})}$ (**A**) as quantified by the ℓ_2 norm (33); the distance between $\varphi_i^{(n,\,\Xi\text{-Dirac})}$ and $\varphi_i^{(n,\,\Lambda\text{-Dirac})}$ (**B**) as quantified by the ℓ_2 norm. The number of leaves $n = 50$. Values were computed over the grid $\{1.0, 1.025, \ldots, 2\}$ when associated with the Beta-coalescent (**A**); and $\{0, 0.025, \ldots, 0.975\}$ when associated with the Dirac-coalescent (**B**).



bioRxiv preprint doi: https://doi.org/10.1101/025684; this version posted February 6, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Application to Atlantic cod data

Atlantic cod is a diploid highly fecund marine organism, whose reproduction is potentially characterised by a skewed offspring distribution [1, 2]. Since Xi-coalescents can arise from diploid population models which admit skewed offspring distributions [34, 9], one should analyse population genetic data of autosomal loci in diploid highly fecund populations with Xi-coalescent models. Indeed, [2] obtain population genetic data at three autosomal loci from Atlantic cod. We use the ℓ_2-norm (33) to fit (see Table 3) the 4-fold Xi-Beta and 4-fold Xi-Dirac coalescents to the unfolded site-frequency spectrum (USFS) of *Ckma*, *Myg*, and *HbA2* genes obtained by [2]. The 4-fold Xi-coalescents correspond to a diploid population in which one successful pair of parents contributes offspring in each generation.

The USFS of the autosomal genes *Ckma*, *Myg*, and *HbA2* are all characterised by a high relative amount of singletons. Thus, singletons have the most weight in our estimate, in particular since we do not calibrate the difference between observed and expected values by the variance. The estimates of α associated with the Xi-Beta coalescent are therefore all at 1.0, which we attribute to the excessive amount of singletons. The excessive amount of singletons also increases the estimate of ψ associated with the Xi-Dirac coalescent. In particular, our Xi-based estimates of ψ are higher than the Lambda-based estimates (Table 3). Possibly the Xi-coalescent assigns less mass to the external branches than the corresponding Lambda-coalescent for a given parameter value, but the exact shift in mass may vary between different Xi-coalescents. The Xi-Dirac coalescent is able to predict the excessive amount of singletons, the Xi-Beta coalescent much less so (Figures 5–6).

Our estimates of ψ for the combined data on *Ckma* are smaller than for the partitioned data (into A and B alleles), and for the supposedly neutral loci *Myg* and *HbA2*. [2] also observe a similar pattern. Of the three loci, the Xi-coalescents give best fit to the *Ckma* data. In view of the modes in the right tail of the USFS for *Ckma*, [2] conclude that *Ckma* is under strong selection. Even though the Xi-Dirac coalescent does show multi-modal spectrum (Figure 3), the modes are small relative to the expected length of external branches, and do not quite explain the modes observed for *Ckma* (Figure 6).

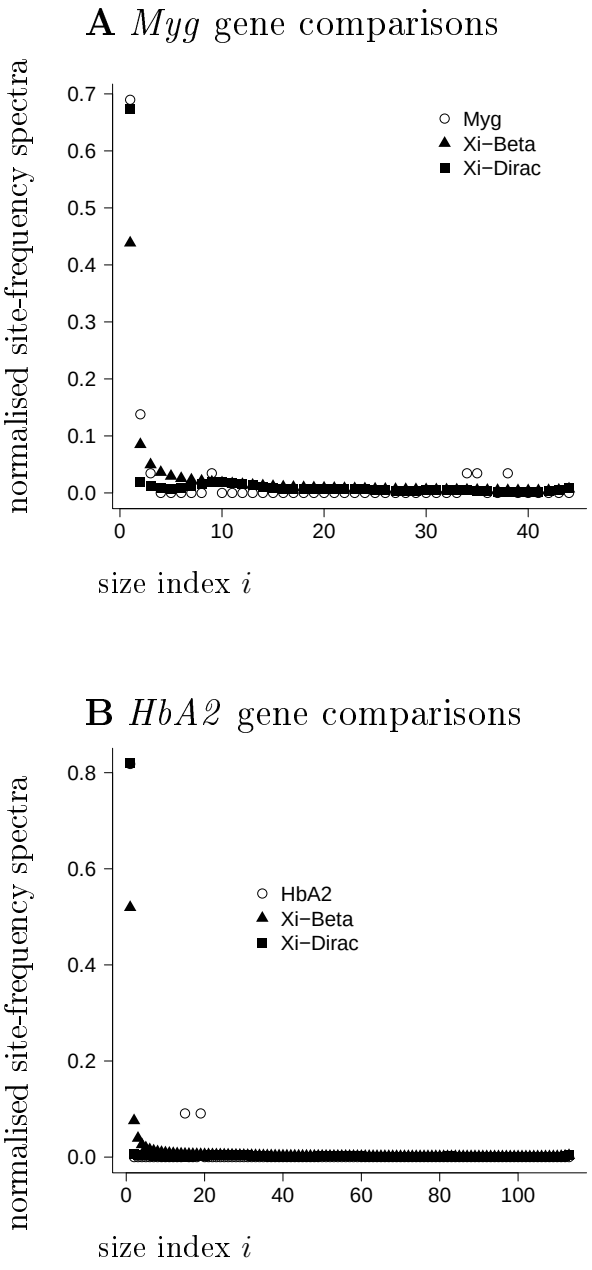
bioRxiv preprint doi: https://doi.org/10.1101/025684; this version posted February 6, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 3: Parameter estimates $(\widehat{\alpha}, \widehat{\psi})$ obtained by minimising the ℓ_2-norm (33) for the unfolded site-frequency spectrum of Atlantic cod data [2]. Parameter α associated with the Xi-Beta and the Lambda-Beta coalescent was estimated over the grid $\{1.0, 0.05, \ldots, 1.95\}$, and ψ associated with the Xi-Dirac and the Lambda-Dirac coalescent was estimated over the grid $\{0.0, 0.05, \ldots, 0.95\}$.

			Xi-coalescents			
locus	sample size	seg. sites	$\widehat{\alpha}$	$\ell_2(\widehat{\alpha})$	$\widehat{\psi}$	$\ell_2\left(\widehat{\psi}\right)$
Ckma	122	91	1.0	0.08	0.30	0.10
CkmaA	43	47	1.0	0.21	0.65	0.12
CkmaB	79	55	1.0	0.18	0.50	0.13
Myg	45	29	1.0	0.27	0.80	0.14
HbA2	114	11	1.0	0.34	0.80	0.13
			Lambda-coalescents			
Ckma	122	91	1.25	0.07	0.05	0.12
CkmaA	43	47	1.1	0.13	0.15	0.11
CkmaB	79	55	1.1	0.08	0.15	0.13
Myg	45	29	1.0	0.10	0.20	0.14
HbA2	114	11	1.0	0.19	0.25	0.13

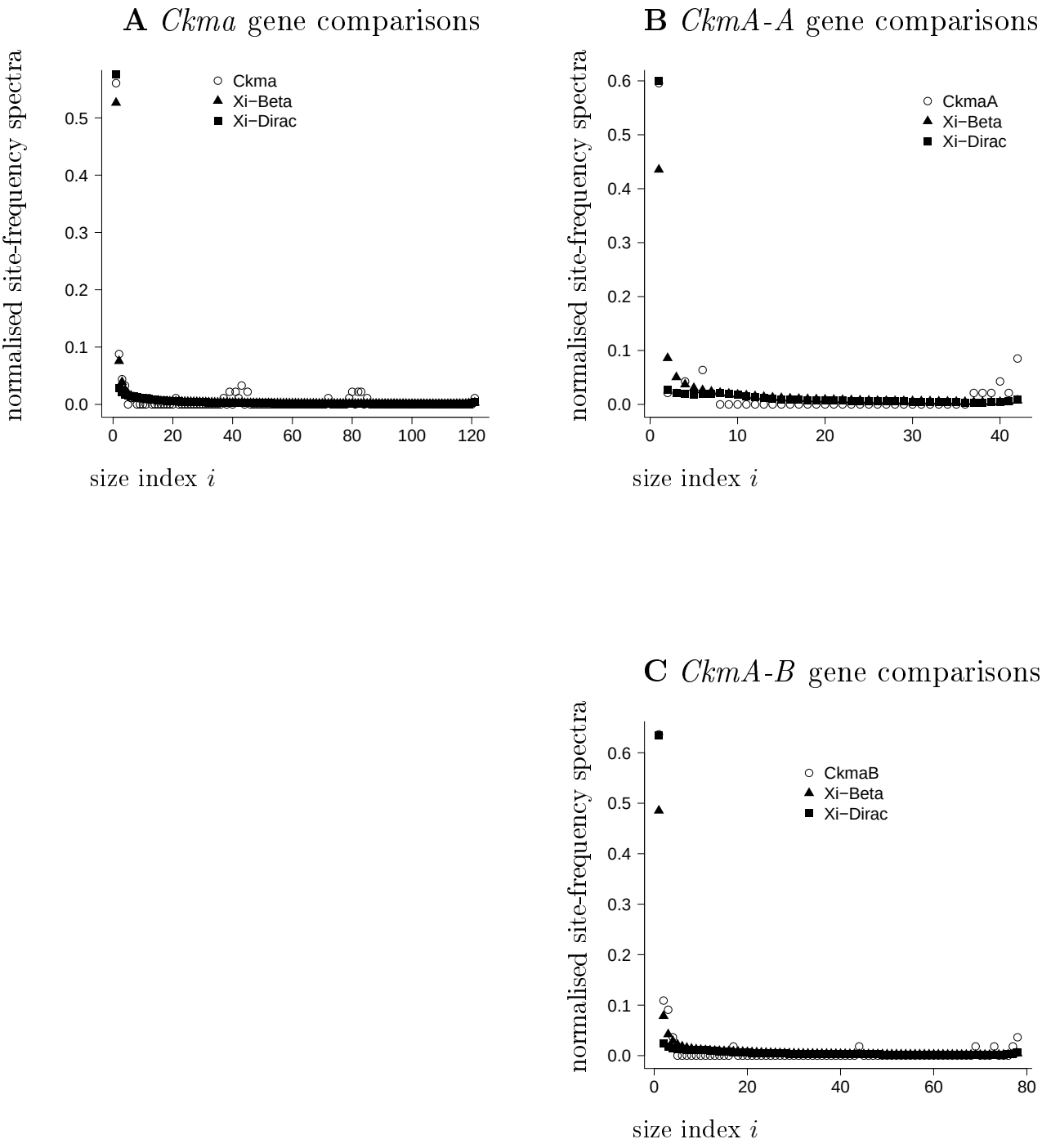
bioRxiv preprint doi: https://doi.org/10.1101/025684; this version posted February 6, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 5: Comparison of observed (χ_i; ○) and expected $\left(\varphi_i^{(n,\Pi)}; ■, ▲\right)$ (32) normalised site-frequency spectrum for *Myg* and *HbA2* [2] with 4-fold Xi-coalescent process as shown. The associated parameter values are given in Table 3. In **A**, the observed and expected values for the *Myg* gene are compared; in **B** for the *HbA2* gene.



bioRxiv preprint doi: https://doi.org/10.1101/025684; this version posted February 6, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 6: Comparison of observed $(\chi_i; \circ)$ and expected $\left(\varphi_i^{(n,\Pi)}; \blacksquare, \blacktriangle\right)$ (32) normalised site-frequency spectrum for *Ckma* [2] with 4-fold Xi-coalescent process as shown. The associated parameter values are given in Table 3. In **A**, all the *CkmA* sequences are included; in **B**, only the *A* allele is considered; in **C**, only the *B* allele is considered.



bioRxiv preprint doi: https://doi.org/10.1101/025684; this version posted February 6, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Discussion

We prove recursions for the expected site-frequency spectrum associated with Xi-coalescents which admit simultaneous multiple mergers of active ancestral lineages (blocks of the current partition). We give a class of Xi-coalescents which is 'driven' by a finite measure (a Lambda-measure) on the unit interval, which determines the law of the total number of active lineages which may merge each time. This class of Xi-coalescents can be applied to populations of arbitrary ploidy. We apply the recursions to compare estimates of coalescent parameters between Lambda- and Xi-coalescents. Finally, we estimate coalescent parameters associated with Xi-coalescents for Atlantic cod where the data are the unfolded site-frequency spectrum on autosomal loci.

The framework we develop will allow us to extend the recursions to more complicated scenarios, such as populations structured into discrete subpopulations [21, 31], or possibly by considering some sort of continuous distribution in space [3]. However, one would extend the recursions to such more structured frameworks at the high risk of increasing their computational complexity.

Computing the full expected site-frequency spectrum for a 4-fold Xi-coalescent with sample size $n \geq 100$ takes unfortunately a bit of time (on the order of hours). We do not provide detailed analysis, but the time can be shortened by considering the lumped site-frequency spectrum, where one would collect all classes of size larger than some number $m \ll n$ into one class. How such lumping would affect the inference remains to be seen. However, one can analyse small samples ($n \leq 100$) with our recursions, as we provide an example of. Exact likelihood methods in the spirit of [8] are yet to be developed for Xi-coalescents, and will likely be computationally intensive. In a very recent preprint, [43] introduce a different method to compute the expected SFS associated with Xi-coalescents. The method of [43] is claimed to be approximately n^2 faster than our method, where n is sample size.

Our simple method of minimising the deviation between observed and expected values of the unfolded site-frequency spectrum should not be applied in a formal test procedure, since we do not scale the deviations by the corresponding variance. We do not give recursions for

bioRxiv preprint doi: https://doi.org/10.1101/025684; this version posted February 6, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

the (co)-variances of the site-frequency spectrum associated with Xi-coalescents, as these will very likely be too computationally intensive to be useful. This has already been shown to be the case for the much simpler Lambda-coalescents [10]. Our recursions provide a way to distinguish between Xi-coalescents and other demographic effects such as population growth, with the use of approximate likelihoods [22].

We obtain estimates of coalescent parameters associated with Xi-coalescents for data on autosomal loci of Atlantic cod [2]. Our estimates differ from previous estimates obtained with the use of Lambda-coalescents [2], due to the simultaneous merger characteristic of Xi-coalescents. Regardless of exact estimates, our results, coupled with those of [2], suggest that multiple merger coalescents might be the proper null model with which to analyse population genetic data on Atlantic cod, as well as other populations which may exhibit high fecundity coupled with skewed offspring distribution (HFSOD). Our specific examples of Xi-coalescents may well represent an oversimplification of the actual mating schemes; ie. assuming only one successful couple contributes offspring in each generation. However, one could use the framework of [14] to develop new examples of Xi-coalescents for specific mating schemes, for example when one successful female produces offspring with many males.

Rigorous inference methods to distinguish the effects of HFSOD from selection are yet to be developed. The common notion is that selective sweeps lead to an excess of singletons. The main genetic signature of HFSOD is also an excess of singletons. The unfolded site-frequency spectrum of the *Ckma* gene [2] is trimodal, with an excess of singletons, and small modes of mutations of larger size. These smaller modes are not captured by the examples of Xi-coalescents that we apply. Durrett and Schweinsberg [42, 20] use a stick-breaking construction to obtain a good approximation ($\leq O(1/(\log(N))^2)$) to a selective sweep where N denotes the population size. [2] conclude that *Ckma* is under a form of balancing selection. Our examples of Xi-coalescents also give best fit to the *Ckma* gene of the three loci studied by [2], although our method does not constitute a formal test. We refer to [2] for more detailed discussion of the variation observed at *Ckma*, and the supposedly neutral nuclear loci *Myg* and *HbA2*. The congruence between our Xi-coalescent examples and *Ckma*, and between Xi-

bioRxiv preprint doi: https://doi.org/10.1101/025684; this version posted February 6, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

coalescents and selective sweeps studied by [42, 20], and the different site-frequency spectra predicted by Lambda- and Xi-coalescents, leads us to conclude that Xi-coalescents form an important class of mathematical objects.

ACKNOWLEDGMENTS

JB and BE acknowledge support by Deutsche Forschungsgemeinschaft (DFG) grant BL 1105/3-1 as part of SPP Priority Programme 1590 'Probabilistic Structures in Evolution'.

References

[1] E Árnason. Mitochondrial cytochrome *b* variation in the high-fecundity Atlantic cod: trans-Atlantic clines and shallow gene genealogy. *Genetics*, 166:1871–1885, 2004.

[2] Einar Árnason and Katrín Halldórsdóttir. Nucleotide variation and balancing selection at the *Ckma* gene in Atlantic cod: analysis with multiple merger coalescent models. *PeerJ*, 3:e786, 2015.

[3] NH Barton, AM Etheridge, and A Véber. A new model for evolution in a spatial continuum. *Electron J Probab*, 15:162–216, 2010.

[4] A-L Basdevant and C Goldschmidt. Asymptotics of the allele frequency spectrum associated with the Bolthausen-Sznitman coalescent. *Electron J Probab*, 13:486–512, 2008.

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

[6] J Berestycki, N Berestycki, and J Schweinsberg. Small-time behavior of beta coalescents. *Ann Inst H Poincaré Probab Statist*, 44:214–238, 2008.

[7] N Berestycki. Recent progress in coalescent theory. *Ensaios Mathématicos*, 16:1–193, 2009.

[8] M Birkner and J Blath. Computing likelihoods for coalescents with multiple collisions in the infinitely many sites model. *J Math Biol*, 57:435–465, 2008.

bioRxiv preprint doi: https://doi.org/10.1101/025684; this version posted February 6, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[9] M Birkner, J Blath, and B Eldon. An ancestral recombination graph for diploid populations with skewed offspring distribution. *Genetics*, 193:255–290, 2013.

[10] M Birkner, J Blath, and B Eldon. Statistical properties of the site-frequency spectrum associated with Λ-coalescents. *Genetics*, 195:1037–1053, 2013.

[11] M Birkner, J Blath, M Möhle, M Steinrücken, and J Tams. 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.

[12] M Birkner, J Blath, and M Steinrücken. Importance sampling for Lambda-coalescents in the infinitely many sites model. *Theor Popul Biol*, 79:155–173, 2011.

[13] M Birkner, J Blath, and M Steinrücken. Analysis of DNA sequence variation within marine species using Beta-coalescents. *Theor Popul Biol*, 87:15–24, 2013.

[14] M Birkner, H Liu, and A Sturm. A note on coalescent results for diploid exchangeable population models. *Preprint*, 2015.

[15] M C Cronjäger. On the expected site-frequency spectrum associated with Ξ-coalescents. Master's thesis, TU Berlin, Berlin, 2014.

[16] I Dahmer and G Kersting. The internal branch lengths of the Kingman coalescent. *Ann Applied Probab*, 25:1325–1348, 2015.

[17] P Donnelly and T G Kurtz. Particle representations for measure-valued population models. *Ann Probab*, 27:166–205, 1999.

[18] R Durrett. *Probability models for DNA sequence evolution*. Springer, New York, 2nd edition, 2008.

[19] R Durrett and J Schweinsberg. Approximating selective sweeps. *Theor Popul Biol*, 66:129–138, 2004.

bioRxiv preprint doi: https://doi.org/10.1101/025684; this version posted February 6, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[20] R Durrett and J Schweinsberg. A coalescent model for the effect of advantageous mutations on the genealogy of a population. *Stoch Proc Appl*, 115:1628–1657, 2005.

[21] B Eldon. Structured coalescent processes from a modified Moran model with large offspring numbers. *Theor Popul Biol*, 76:92–104, 2009.

[22] B Eldon, M Birkner, J Blath, and F Freund. Can the site-frequency spectrum distinguish exponential population growth from multiple-merger coalescents. *Genetics*, 199:841–856, 2015.

[23] B Eldon and J Wakeley. Coalescent processes when the distribution of offspring number among individuals is highly skewed. *Genetics*, 172:2621–2633, 2006.

[24] Y-X Fu. Statistical properties of segregating sites. *Theor Popul Biol*, 48:172–197, 1995.

[25] Y-X Fu. New statistical tests of neutrality for DNA samples from a population. *Genetics*, 143:557–570, 1996.

[26] D Hedgecock. Does variance in reproductive success limit effective population sizes of marine organisms? In A Beaumont, editor, *Genetics and evolution of Aquatic Organisms*, pages 1222–1344, London, 1994. Chapman and Hall.

[27] D Hedgecock and A I Pudovkin. Sweepstakes reproductive success in highly fecund marine fish and shellfish: a review and commentary. *Bull Marine Science*, 87:971–1002, 2011.

[28] J F C Kingman. The coalescent. *Stoch Proc Appl*, 13:235–248, 1982.

[29] J F C Kingman. Exchangeability and the evolution of large populations. In G Koch and F Spizzichino, editors, *Exchangeability in Probability and Statistics*, pages 97–112, Amsterdam, 1982. North-Holland.

[30] J F C Kingman. On the genealogy of large populations. *J App Probab*, 19A:27–43, 1982.

bioRxiv preprint doi: https://doi.org/10.1101/025684; this version posted February 6, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[31] V Limic and A Sturm. The spatial Λ-coalescent. *Electron J Probab*, 11:363–393, 2006.

[32] H McCombie, S Lapègue, F Cornette, C Ledu, and P Boudry. Chromosome loss in biparental progenies of tetraploid Pacific oysters *Crassostrea gigas*. *Aquaculture*, 247:97–105, 2005.

[33] M Möhle and S Sagitov. A classification of coalescent processes for haploid exchangeable population models. *Ann Probab*, 29:1547–1562, 2001.

[34] M Möhle and S Sagitov. Coalescent patterns in diploid exchangeable population models. *J Math Biol*, 47:337–352, 2003.

[35] J Pitman. Coalescents with multiple collisions. *Ann Probab*, 27:1870–1902, 1999.

[36] S Sagitov. The general coalescent with asynchronous mergers of ancestral lines. *J Appl Probab*, 36:1116–1125, 1999.

[37] S Sagitov. Convergence to the coalescent with simultaneous mergers. *J Appl Probab*, 40:839–854, 2003.

[38] O Sargsyan and J Wakeley. A coalescent process with simultaneous multiple mergers for approximating the gene genealogies of many marine organisms. *Theor Pop Biol*, 74:104–114, 2008.

[39] S A Sawyer and D L Hartl. Population genetics of polymorphism and divergence. *Genetics*, 132:1161–1176, 1992.

[40] J Schweinsberg. Coalescents with simultaneous multiple collisions. *Electron J Probab*, 5:1–50, 2000.

[41] J Schweinsberg. Coalescent processes obtained from supercritical Galton-Watson processes. *Stoch Proc Appl*, 106:107–139, 2003.

[42] J Schweinsberg and R Durrett. Random partitions approximating the coalescence of lineages during a selective sweep. *Ann Appl Probab*, 1591–1651, 2005.

bioRxiv preprint doi: https://doi.org/10.1101/025684; this version posted February 6, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[43] JP Spence, JA Kamm, and YS Song. The site-frequency spectrum for general coalescents. http://arxiv.org/abs/1510.05631.

[44] R P Stanley. *Enumerative Combinatorics*, volume 2. Cambridge University Press, New York, 2001.

[45] J Wakeley. *Coalescent theory*. Roberts & Co, 2007.

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

Appendix

Proof of recursion (26) for $p^{(n)}[k,i]$ (Thm. 2)

The requirement $p^{(n)}[n,1] = 1$ is obvious (therefore $p^{(n)}[n,k] = 0$ for all $k > 1$). From now on, we consider the case $1 < k < n$. By $\mathcal{P}_{\mathbb{N}}$ we denote the set of all partitions of $\mathbb{N} := \{1, 2, \ldots\}$. Let $\Pi \equiv \{\Pi_t, t \geq 0\}$ be a $\mathcal{P}_{\mathbb{N}}$-valued exchangeable coalescent, defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. By $\Pi^{(n)} \equiv \left\{\Pi_t^{(n)}, t \geq 0\right\}$ we denote the projection of Π onto $\mathcal{P}_n$, which will be associated with coalescent processes started from $n \geq 2$ leaves. Define

$$\tau_k^{(n)} := \inf\left(\left\{t > 0 | \#\Pi_t^{(n)} = k\right\} \cup \{\infty\}\right), \tag{35}$$

ie. the first time the block counting process associated with $\Pi^{(n)}$ hits state k, with $\tau_n^{(n)} = 0$. Let

$$\Omega_k^{(n)} := \{\tau_k^{(n)} < \infty\}.$$

Recall that we assume that $\mathbb{P}\left(\tau_k^{(n)} < \infty\right) > 0$ for each $k \in [n]$. Thus we can define the conditional law $\mathbb{P}_k^{(n)}(\cdot) := \mathbb{P}\left(\cdot | \Omega_k^{(n)}\right)$. On the conditional probability space $\left(\Omega_k^{(n)}, \mathcal{F}|_{\Omega_k^{(n)}}, \mathbb{P}_k^{(n)}\right)$, we sample uniformly at random a block π_0 from the blocks of $\Pi^{(n)}_{\tau_k^{(n)}}$, ie. $\pi_0(\omega) \in \Pi^{(n)}_{\tau_k^{(n)}}(\omega)$ for

bioRxiv preprint doi: https://doi.org/10.1101/025684; this version posted February 6, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

$\omega \in \Omega_k^{(n)}$. Then we can write

$$p^{(n)}[k,i] = \mathbb{P}_k^{(n)}\left(\#\pi_0 = i\right). \tag{36}$$

We define the first jump time τ of the block-counting process

$$\tau := \inf\left\{t > 0 : \#\Pi_t^{(n)} < n\right\}.$$

Now consider some $k \le m < n$, and partition $|\nu| \stackrel{m}{=} n$, ie. an integer partition ν of n into m elements. Define

$$\Omega_k^{(n,\nu)} := \left\{\left(\Pi_\tau^{(n)}\right)^\downarrow = \nu\right\} \cap \Omega_k^{(n)}$$

(recall that $\pi^\downarrow$ denotes the integer partition associated to $\pi \in \mathcal{P}_{[n]}$ obtained by listing the block sizes of π in decreasing order). Then it is clear that we have the decomposition

$$\Omega_k^{(n)} = \bigcup_{m=k}^{n-1} \bigcup_{|\nu| \stackrel{m}{=} n} \Omega_k^{(n,\nu)}. \tag{37}$$

We define the conditional law $\mathbb{P}_k^{(n,\nu)}(\cdot) := \mathbb{P}\left(\cdot | \Omega_k^{(n,\nu)}\right)$.

For each $\omega \in \Omega_k^{(n,\nu)}$, we define the set of blocks

$$\pi^{(\nu,0)}(\omega) := \left\{\pi_j \in \Pi_\tau^{(n)}(\omega) : \pi_j \subseteq \pi_0, j \in [m]\right\}, \qquad \omega \in \Omega_k^{(n,\nu)}. \tag{38}$$

The set of blocks $\pi^{(\nu,0)}(\omega)$ contains all blocks of the partition $\Pi_\tau^{(n)}(\omega)$ that will eventually merge into the block $\pi_0(\omega)$.

For any integer subpartition $\varrho = \langle \beta_1, \beta_2, \ldots \rangle \subset \nu = \langle \alpha_1, \alpha_2, \ldots \rangle$, we need to be able to compute the probability $\mathbb{P}_k^{(n,\nu)}\left(\left(\pi^{(\nu,0)}\right)^\downarrow = \varrho\right)$ on $\Omega_k^{(n,\nu)}$. Indeed,

$$\begin{aligned}\mathbb{P}_k^{(n,\nu)}\left(\left(\pi^{(\nu,0)}\right)^\downarrow = \varrho\right) &= \mathbb{P}_k^{(n,\nu)}\left(\left(\pi^{(\nu,0)}\right)^\downarrow = \varrho \,|\, \#\pi^{(\nu,0)} = \#\varrho\right) \cdot \mathbb{P}_k^{(n,\nu)}\left(\#\pi^{(\nu,0)} = \#\varrho\right) \\ &= \mathbb{1}_{(\#\varrho \le \#\nu - k + 1)} \cdot p^{(\#\nu)}[k, \#\varrho] \frac{\prod_\ell \binom{\alpha_\ell}{\beta_\ell}}{\binom{\#\nu}{\#\varrho}}.\end{aligned} \tag{39}$$

bioRxiv preprint doi: https://doi.org/10.1101/025684; this version posted February 6, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The event $\left\{\left(\pi^{(\nu,0)}\right)^{\downarrow} = \varrho\right\}$ states that the sizes of the blocks of $\pi^{(\nu,0)}$ are given by the integer subpartition ϱ. Equation (39) follows from the exchangeability of Π, and the form of the probability density function of the multivariate hypergeometric distribution, which applies to the event of sampling β_ℓ blocks from α_ℓ for each ℓ, for a total of $\#\varrho$ out of $\#\nu$. The condition $\#\varrho \leq \#\nu - k + 1$ is required since we condition on the block counting process to hit k blocks.

Given (39) we can compute $\mathbb{P}_k^{(n,\nu)}(\#\pi_0 = i)$. Indeed, the decomposition

$$\{\#\pi_0 = i\} \cap \Omega_k^{(n,\nu)} = \bigcup_{\substack{\varrho \subset \nu \\ |\varrho| = i}} \left\{\left(\pi^{(\nu,0)}\right)^{\downarrow} = \varrho\right\}$$

(where the union is disjoint) together with (39) gives

$$\begin{aligned}
\mathbb{P}_k^{(n,\nu)}(\#\pi_0 = i) &= \sum_{\substack{\varrho \subset \nu \\ |\varrho| = i}} \mathbb{P}_k^{(n,\nu)}\left((\pi^{(\nu,0)})^{\downarrow} = \varrho\right) \\
&= \sum_{\substack{\varrho \subset \nu \\ |\varrho| = i}} \mathbb{1}_{(\#\varrho \leq \#\nu - k + 1)} \cdot p^{(\#\nu)}[k, \#\varrho] \frac{\prod_\ell \binom{\alpha_\ell}{\beta_\ell}}{\binom{\#\nu}{\#\varrho}} \\
&= \sum_{j=1}^{i \wedge (\#\nu - k + 1)} \sum_{\substack{\varrho \subset \nu \\ |\varrho| \overset{j}{=} i}} p^{(\#\nu)}[k, j] \frac{\prod_\ell \binom{\alpha_\ell}{\beta_\ell}}{\binom{\#\nu}{j}}.
\end{aligned} \tag{40}$$

Given Equation (40) one can now compute $p^{(n)}[k, i]$: by the decomposition (37), we obtain

$$\begin{aligned}
p^{(n)}[k, i] &= \mathbb{P}_k^{(n)}(\#\pi_0 = i) \\
&= \sum_{m=k}^{n-1} \sum_{|\nu| \overset{m}{=} n} \mathbb{P}_k^{(n)}\left(\{\#\pi_0 = i\} \cap \Omega_k^{(n,\nu)}\right) \\
&= \sum_{m=k}^{n-1} \sum_{|\nu| \overset{m}{=} n} \mathbb{P}_k^{(n)}\left(\Omega_k^{(n,\nu)}\right) \mathbb{P}_k^{(n)}\left(\#\pi_0 = i \,|\, \Omega_k^{(n,\nu)}\right) \\
&= \sum_{m=k}^{n-1} \sum_{|\nu| \overset{m}{=} n} \mathbb{P}_k^{(n)}\left(\Omega_k^{(n,\nu)}\right) \mathbb{P}_k^{(n,\nu)}(\#\pi_0 = i).
\end{aligned} \tag{41}$$

bioRxiv preprint doi: https://doi.org/10.1101/025684; this version posted February 6, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

We apply Lemma 5 (stated and proved below) to $\mathbb{P}_k^{(n)}\left(\Omega_k^{(n,\nu)}\right)$ and Equation (40) to $\mathbb{P}_k^{(n,\nu)}(\#\pi_0 = i)$ to obtain

$$p^{(n)}[k,i] = \sum_{m=k}^{n-1} \sum_{|\nu| \stackrel{m}{=} n} p_\nu^{(n)} \frac{g(m,k)}{g(n,k)} \sum_{j=1}^{i\wedge(m-k+1)} \sum_{\substack{\varrho \subset \nu \\ |\varrho| \stackrel{j}{=} i}} p^{(m)}[k,j] \frac{\prod_\ell \binom{\alpha_\ell}{\beta_\ell}}{\binom{m}{j}}.$$

□

Proof of computation of $\mathbb{P}_k^{(n)}\left(\Omega_k^{(n,\nu)}\right)$ (Lemma 5)

Let $\Omega_k^{(n,\nu)}$, $\mathbb{P}_k^{(n)}(\cdot)$, $\tau_k^{(n)}$ (35) and τ be as defined in the proof for Thm. 2. Recall further that $p_\nu^{(n)} := \mathbb{P}\left((\Pi_\tau^{(n)})^\downarrow = \nu\right)$, see Equation (25).

By $\left\{Y_t^{(n)}; t \geq 0\right\}$, $Y_t^{(n)} := \#\Pi_t^{(n)}$ we denote the block-counting process associated with $\Pi^{(n)}$, starting from n. Let $g(n,k)$ denote the expected length of time that $Y_t^{(n)}$ spends in state $k \leq n$,

$$g(n,k) := \mathbb{E}\left[\int_0^\infty \mathbb{1}_{\left(Y_s^{(n)} = k\right)} ds\right]. \tag{42}$$

Clearly, $g(k,k) = \frac{1}{\lambda_k} = -q_{\pi,\pi}$ where $\#\pi = k$, see (12).

Lemma 5. *For $k \leq m < n$ and $|\nu| \stackrel{m}{=} n$, assuming $\mathbb{P}\left(\tau_k^{(n)} < \infty\right) > 0$, we have*

$$\mathbb{P}_k^{(n)}\left(\Omega_k^{(n,\nu)}\right) \equiv \mathbb{P}_k^{(n)}\left((\Pi_\tau^{(n)})^\downarrow = \nu\right) = p_\nu^{(n)} \frac{g(m,k)}{g(n,k)}, \tag{43}$$

where $g(\cdot\,,\cdot)$ is defined in (42).

bioRxiv preprint doi: https://doi.org/10.1101/025684; this version posted February 6, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Proof. Assuming $\mathbb{P}\left(\tau_k^{(n)} < \infty\right) > 0$, one obtains

$$\begin{aligned}
\mathbb{P}_k^{(n)}\left((\Pi_\tau^{(n)})^\downarrow = \nu\right) &= \frac{\mathbb{P}\left(\left\{\left(\Pi_\tau^{(n)}\right)^\downarrow = \nu\right\} \cap \left\{\tau_k^{(n)} < \infty\right\}\right)}{\mathbb{P}\left(\tau_k^{(n)} < \infty\right)} \\
&= \mathbb{P}\left((\Pi_\tau^{(n)})^\downarrow = \nu\right) \frac{\mathbb{P}\left(\tau_k^{(n)} < \infty \middle| \left(\Pi_\tau^{(n)}\right)^\downarrow = \nu\right)}{\mathbb{P}\left(\tau_k^{(n)} < \infty\right)} \\
&= p_\nu^{(n)} \frac{\mathbb{P}\left(\tau_k^{(m)} < \infty\right)}{\mathbb{P}\left(\tau_k^{(n)} < \infty\right)},
\end{aligned} \tag{44}$$

where we use $\#\nu = m$ and apply the strong Markov property of the process $\left(\Pi^{(n)}\right)^\downarrow$ at time τ to obtain the last equality.

Now we consider $\frac{g(m,k)}{g(n,k)}$, and obtain

$$\begin{aligned}
\frac{g(m,k)}{g(n,k)} &= \frac{\mathbb{E}\left[\int_0^\infty \mathbb{1}_{\left(Y_s^{(m)}=k\right)} ds \mid \tau_k^{(m)} < \infty\right] \mathbb{P}\left(\tau_k^{(m)} < \infty\right)}{\mathbb{E}\left[\int_0^\infty \mathbb{1}_{\left(Y_s^{(n)}=k\right)} ds \mid \tau_k^{(n)} < \infty\right] \mathbb{P}\left(\tau_k^{(n)} < \infty\right)} \\
&= \frac{\mathbb{E}\left[\int_{\tau_k^{(m)}}^\infty \mathbb{1}_{\left(Y_s^{(m)}=k\right)} ds \mid \tau_k^{(m)} < \infty\right] \mathbb{P}\left(\tau_k^{(m)} < \infty\right)}{\mathbb{E}\left[\int_{\tau_k^{(n)}}^\infty \mathbb{1}_{\left(Y_s^{(n)}=k\right)} ds \mid \tau_k^{(n)} < \infty\right] \mathbb{P}\left(\tau_k^{(n)} < \infty\right)} \\
&= \frac{g(k,k)}{g(k,k)} \frac{\mathbb{P}\left(\tau_k^{(m)} < \infty\right)}{\mathbb{P}\left(\tau_k^{(n)} < \infty\right)},
\end{aligned} \tag{45}$$

where we use the strong Markov property of the block-counting process in the last step. Now the statement follows from (44) and (45).

□

bioRxiv preprint doi: https://doi.org/10.1101/025684; this version posted February 6, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Proof of recursion (23) for $g(n,k)$ (Lemma 1)

Proof. For $n \geq 2$, let $\lambda_n := -q_{\pi,\pi}$ with $q_{\pi,\pi}$ given by (12), and $\#\pi = n$. Again write $Y_t^{(n)} := \#\Pi_t^{(n)}$ for the block-counting process starting from n.

Let $\tau := \inf\left\{t > 0 : Y_t^{(n)} < n\right\}$ denote the first jump time of the block-counting process, and let $p_{n,m} := \mathbb{P}\left(Y_\tau^{(n)} = m\right)$, $m \leq n-1$. For $k = n$, one obtains

$$g(n,n) = \mathbb{E}\left[\int_0^\infty \mathbb{1}_{\left(Y_s^{(n)}=n\right)} ds\right] = \mathbb{E}[\tau] = \lambda_n^{-1},$$

and (24) is established. For $k < n$, we decompose according to the value of the block-counting process after the first jump and obtain

$$\begin{aligned} g(n,k) = \mathbb{E}\left[\int_0^\infty \mathbb{1}_{\left(Y_s^{(n)}=k\right)} ds\right] &= \sum_{m=k}^{n-1} \mathbb{E}\left[\int_\tau^\infty \mathbb{1}_{\left(Y_s^{(n)}=k\right)} ds \middle| Y_\tau^{(n)} = m\right] \mathbb{P}\left(Y_\tau^{(n)} = m\right) \\ &= \sum_{m=k}^{n-1} \mathbb{E}\left[\int_0^\infty \mathbb{1}_{\left(Y_s^{(m)}=k\right)} ds\right] p_{n,m} \\ &= \sum_{m=k}^{n-1} p_{n,m}\, g(m,k), \end{aligned}$$

where we use the strong Markov property of Y in the second-to-last equality. Thus (23) is established.

Proof of Equation (14)

Fix $m \geq 2$, $\underline{k} := (k_1, \ldots, k_r)$ with $k_1 \geq \ldots \geq k_r \geq 2$, $r \in [M]$, and let $|\underline{k}| := k_1 + \cdots + k_r$, $s := m - |\underline{k}| = m - k_1 - \cdots - k_r$. Then we have to prove that

$$\lambda_{m,\underline{k}} = \sum_{\ell=0}^{s \wedge (M-r)} \binom{s}{\ell} (M)_{r+\ell} M^{-(|\underline{k}|+\ell)} \int_{[0,1]} x^{|\underline{k}|+\ell-2} (1-x)^{s-\ell} F(dx).$$

bioRxiv preprint doi: https://doi.org/10.1101/025684; this version posted February 6, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Let Δ denote the infinite simplex, and recall the function f from (7) used to describe the rates (12) of a Xi-coalescent. We define the map $h_{(m,\underline{k})} : \Delta \ni \xi \mapsto f(\xi, m, \underline{k}) \in \mathbb{R}$. We then have that the rate of a $\underline{k}$-merger is given by

$$\lambda_{m,\underline{k}} = \int_{\xi\in\Delta} h_{(m,\underline{k})}(\xi)\, \Xi(d\xi). \tag{46}$$

Since $\mathrm{supp}(\Xi) \subseteq \Delta_M := \{(\underbrace{\tfrac{x}{M}, \ldots, \tfrac{x}{M}}_{M \text{ times}}, 0, 0, \ldots) \in \Delta : x \in [0,1]\}$, we can rewrite the rate in (46) as

$$\lambda_{m,\underline{k}} = \frac{1}{M} \int_{x\in[0,1]} h_{(m,\underline{k})}(\underbrace{\tfrac{x}{M}, \ldots, \tfrac{x}{M}}_{M \text{ times}}, 0, 0, \ldots) F(dx). \tag{47}$$

Define $x_i := \frac{x}{M} \mathbb{1}_{(i \le M)}$. Then the integrand in (47) is given by

$$\frac{\sum_{\ell=0}^{s} \sum_{i_1 \neq \ldots \neq i_{r+\ell}} \binom{s}{\ell} x_{i_1}^{k_1} \cdots x_{i_r}^{k_r} x_{i_{r+1}} \cdots x_{i_{r+\ell}} \left(1 - \sum_j x_j\right)^{s-\ell}}{\sum_j x_j^2},$$

which is zero if $r + \ell > M$. For $r + \ell \le M$, one obtains

$$x_{i_1}^{k_1} \cdots x_{i_r}^{k_r} x_{i_{r+1}} \cdots x_{i_{r+\ell}} = \frac{x^{|\underline{k}|+\ell}}{M^{|\underline{k}|+\ell}}$$

for any choice of indices $i_1, \ldots, i_{r+\ell}$ which are all different and smaller or equal to M. Therefore, with $(a)_n := a(a-1)\cdots(a-n+1)$ for $n \in \mathbb{N}$, $(a)_0 := 1$ denoting the falling factorial, we have

$$\sum_{i_1 \neq \ldots \neq i_{r+\ell}} x_{i_1}^{k_1} \cdots x_{i_r}^{k_r} x_{i_{r+1}} \cdots x_{i_{r+\ell}} = \mathbb{1}_{(r+\ell \le M)} \frac{(M)_{r+\ell}}{M^{|\underline{k}|+\ell}} x^{|\underline{k}|+\ell}.$$

We note further that

$$1 - \sum_j x_j = 1 - x, \qquad \sum_j x_j^2 = M \cdot \left(\frac{x}{M}\right)^2 = \frac{x^2}{M}.$$

bioRxiv preprint doi: https://doi.org/10.1101/025684; this version posted February 6, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Thus we get

$$h_{(m,\underline{k})}(\underbrace{\tfrac{x}{M},\ldots,\tfrac{x}{M}}_{M \text{ times}},0,0,\ldots) = \sum_{\ell=0}^{s\wedge(M-r)} \binom{s}{\ell}(1-x)^{s-\ell}x^{|\underline{k}|+\ell-2}(M)_{r+\ell}M^{1-|\underline{k}|-\ell},$$

and we can represent the rate $\lambda_{m,\underline{k}}$ as

$$\begin{aligned}\lambda_{m,\underline{k}} &= \frac{1}{M}\int_{x\in[0,1]} h_{(m,\underline{k})}(\underbrace{\tfrac{x}{M},\ldots,\tfrac{x}{M}}_{M \text{ times}},0,0,\ldots)F(dx)\\ &= \sum_{\ell=0}^{s\wedge(M-r)} \binom{s}{\ell}(M)_{r+\ell}M^{-(|\underline{k}|+\ell)}\int_{x\in[0,1]} x^{|\underline{k}|+\ell-2}(1-x)^{s-\ell}F(dx).\end{aligned}$$

This completes our proof. □

A condition for the block-counting process to hit state k (Proposition 6)

We now give a necessary and sufficient condition for the block-counting process of a Ξ-n-coalescent to hit state $k \in [n]$ with positive probabilty, i.e. for $\mathbb{P}(\tau_k^{(n)} < \infty) > 0$, where $\tau_k^{(n)}$ is defined as in (35).

Proposition 6. *Let n, k be positive integers satisfying $n > k > 1$, and let $\{\Pi_t^{(\Xi,n)}, t \geq 0\}$ be a Ξ-n-coalescent. Let Δ denote the infinite simplex (6), and $\Delta_{\mathbf{0}} := \Delta \setminus \{(0,0,\ldots)\}$. Then $\mathbb{P}(\tau_k^{(n)} < \infty) > 0$ holds if and only if* $\text{supp}(\Xi) \not\subset \{\boldsymbol{\zeta} \in \Delta_{\mathbf{0}} : \sum_{i<k}\zeta_i = 1\}$.

Coalescents that fail to satisfy the condition $\text{supp}(\Xi) \not\subset \{\boldsymbol{\zeta} \in \Delta_{\mathbf{0}} : \sum_{i<k}\zeta_i = 1\}$ must have the property that the *first* merger-event will almost surely result in a state of less than k blocks, regardless of the value of n. Hence it follows that any m-coalescent which has a 'Kingman part' (i.e. Ξ has mass at $(0,0,\ldots)$), or which admits at least k simultaneous mergers, must necessarily satisfy $\mathbb{P}(\tau_k^{(n)} < \infty) > 0$.

Proof. We consider the decomposition $\Xi = a \cdot \delta_{(0,0,\ldots)} + \Xi_0$ into a Kingman part and a non-Kingman part, as used earlier in the paper.

bioRxiv preprint doi: https://doi.org/10.1101/025684; this version posted February 6, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

In the case $a > 0$, the condition $\mathrm{supp}(\Xi) \not\subset \{\boldsymbol{\zeta} \in \Delta_{\mathbf{0}} : \sum_{i<k} \zeta_i = 1\}$ is satisfied since $(0, 0, \ldots) \notin \Delta_{\mathbf{0}}$. Since there is a non-zero probability of the first $m - k$ merger-events all being single binary mergers, it follows that $\mathbb{P}(\tau_k^{(n)} < \infty) > 0$ must hold in this case.

In the remainder of this proof, we shall concern ourselves with the case $a = 0$. We begin by considering the Poisson process construction of the Ξ-coalescent as outlined in [11, section 1.4]. This corresponds to letting $\{\Pi_t^{(\Xi,n)}, t \geq 0\}$ be driven by a Poisson point process $\mathcal{N}$ on $[0, \infty) \times \Delta \times [0, 1]^n$ with intensity-measure

$$dt \otimes \frac{\Xi_0(d\zeta)}{\sum_i \zeta_i^2} \otimes \left(\mathbb{1}_{[0,1]}(t)\, dt\right)^{\otimes n}.$$

From a realization of the point process $\mathcal{N}$, we then construct $\{\Pi_t^{(\Xi,n)}, t \geq 0\}$ by considering the atoms of $\mathcal{N}(\omega)$ in ascending order of the first coordinate. Given $(t, (\zeta_i)_{i=1,2,\ldots}, (u_i)_{i=1,\ldots,n})$ and $\Pi_{t-}^{(\Xi,n)} = \pi \in \mathcal{P}_n$, we construct $\Pi_t^{(\Xi,n)} = \pi'$ as follows:

- For $i = 1, \ldots, n$ let $j_i \in \mathbb{N} \cup \{\infty\}$ be defined by

$$j_i := \min\left\{j \,:\, \sum_{k<j} \zeta_k \leq u_i < \sum_{k \leq j} \zeta_k\right\}.$$

- For $j = 1, 2, \ldots$ (∞ *not* included), let

$$I_j := \{i : j_i = j\}.$$

  Let π_i denote the ith block of π in order of least elements, and for all $j = 1, 2, \ldots$ merge all blocks with an index belonging to I_j, in order to obtain π'.

In the following, we let

$$\tau := \inf\{t > 0 : \exists\, (\boldsymbol{\zeta}, \underline{u}) \in \Delta \times [0,1]^n \text{ s.t. } (t, \boldsymbol{\zeta}, \underline{u}) \in \mathrm{supp}(\mathcal{N})\}$$

be an $\mathcal{N}$-measurable stopping time corresponding to the time until the first event of $\mathcal{N}$.[1]

[1]The definition of τ is in terms of $\mathcal{N}$, and *not* in terms of $\Pi^{(\Xi,n)}$. As a consequence, we may have

bioRxiv preprint doi: https://doi.org/10.1101/025684; this version posted February 6, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Since Ξ is a non-trivial measure, it follows that $\tau < \infty$ holds almost surely. Furthermore, it follows by construction that $\tau \leq \tau_k^{(n)}$.

We first consider the case $\mathrm{supp}(\Xi) \subset \{\boldsymbol{\zeta} \in \Delta_{\mathbf{0}} : \sum_{i<k} \zeta_i = 1\}$. It follows immediately from the Poisson point process construction that $\#\Pi_\tau^{(\Xi,n)} < k$ must hold (and by extension $\tau_k^{(n)} = \infty$), since every block of $\Pi_0^{(\Xi,n)}$ must merge into one out of strictly less than k groups of blocks at time τ. This proves the 'only if'-part in Proposition 6.

In order to show the 'if'-part in Proposition 6, suppose now that $\mathrm{supp}(\Xi) \not\subset \{\boldsymbol{\zeta} \in \Delta_{\mathbf{0}} : \sum_{i<k} \zeta_i = 1\}$. We consider the two cases $\mathrm{supp}(\Xi) \cap \{\boldsymbol{\zeta} \in \Delta_{\mathbf{0}} : \zeta_k > 0\} \neq \emptyset$ and $\mathrm{supp}(\Xi) \cap \{\boldsymbol{\zeta} \in \Delta_{\mathbf{0}} : \sum_i \zeta_i < 1\} \neq \emptyset$.

In the case $\mathrm{supp}(\Xi) \cap \{\boldsymbol{\zeta} \in \Delta_{\mathbf{0}} : \zeta_k > 0\} \neq \emptyset$, let $\{\|\boldsymbol{\zeta}\|_0 \geq k\}$ be shorthand for the event

$$(\tau, \boldsymbol{\zeta}, \underline{u}) \in [0,\infty) \times \{\boldsymbol{\zeta} \in \Delta_{\mathbf{0}} : \zeta_k > 0\} \times [0,1]^n.$$

We now aim at giving a non-trivial lower bound on $\mathbb{P}(\tau_k^{(n)} < \infty)$ by bounding the right side of the following inequality from below:

$$\begin{aligned}
\mathbb{P}(\tau_k^{(n)} < \infty) &\geq \mathbb{P}(\tau_k^{(n)} = \tau, \|\boldsymbol{\zeta}\|_0 \geq k) \\
&= \int_{\boldsymbol{x} \in \{\boldsymbol{\zeta} \in \Delta_{\mathbf{0}} : \zeta_k > 0\}} \mathbb{P}(\tau_k^{(n)} = \tau, \boldsymbol{\zeta} \in d\boldsymbol{x}) \\
&= \int_{\boldsymbol{x} \in \{\boldsymbol{\zeta} \in \Delta_{\mathbf{0}} : \zeta_k > 0\}} \mathbb{P}(\tau_k^{(n)} = \tau | \boldsymbol{\zeta} = \boldsymbol{x}) \mathbb{P}(\boldsymbol{\zeta} \in d\boldsymbol{x}).
\end{aligned}$$

The inequality follows since the event $\tau_k^{(n)} < \infty$ is a consequence of the event $\tau_k^{(n)} = \tau$, and $\tau < \infty$ holds almost surely. The equalities on the other hand are an immediate consequence of the Poisson point process construction. Hence we may now devote ourselves to proving that the last integral is positive:

$\tau \neq \inf\{t > 0 : \Pi_t^{(\Xi,n)} \neq \Pi_0^{(\Xi,n)}\}$.

bioRxiv preprint doi: https://doi.org/10.1101/025684; this version posted February 6, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Let $\boldsymbol{x} = (x_1, x_2, \ldots) \in \Delta_{\mathbf{0}}$, and let $x_k > 0$. Since

$$\begin{aligned}\mathbb{P}(\tau_k^{(n)} = \tau | \boldsymbol{\zeta} = \boldsymbol{x}) &\geq \mathbb{P}\left(\exists j_1 \ldots j_k \text{ distinct } : I_{j_1} \cup I_{j_2} \cup \ldots \cup I_{j_k} = [n] | \boldsymbol{\zeta} = \boldsymbol{x}\right) \\ &\geq \mathbb{P}(I_1 \cup I_2 \cup \ldots \cup I_k = [n] | \boldsymbol{\zeta} = \boldsymbol{x})\end{aligned}$$

is true by construction, and

$$\mathbb{P}(I_1 \cup I_2 \cup \ldots \cup I_k = [n] | \boldsymbol{\zeta} = \boldsymbol{x}) = \sum_{\substack{l_1, \ldots, l_k \geq 1 \\ l_1 + \cdots + l_k = n}} \frac{\binom{n}{l_1\, l_2 \ldots l_k}}{{x_1}^{l_1} \cdot {x_2}^{l_2} \cdot \ldots \cdot {x_k}^{l_k}} > 0$$

holds, it follows that

$$\begin{aligned}&\int_{\boldsymbol{x} \in \{\boldsymbol{\zeta} \in \Delta_{\mathbf{0}} : \zeta_k > 0\}} \mathbb{P}(\tau_k^{(n)} = \tau | \boldsymbol{\zeta} = \boldsymbol{x}) \mathbb{P}(\boldsymbol{\zeta} \in d\boldsymbol{x}) \\ \geq &\int_{\boldsymbol{x} \in \{\boldsymbol{\zeta} \in \Delta_{\mathbf{0}} : \zeta_k > 0\}} \sum_{\substack{l_1, \ldots, l_k \geq 1 \\ l_1 + \cdots + l_k = n}} \frac{\binom{n}{l_1\, l_2 \ldots l_k}}{{x_1}^{l_1} \cdot {x_2}^{l_2} \cdot \ldots \cdot {x_k}^{l_k}} \mathbb{P}(\boldsymbol{\zeta} \in d\boldsymbol{x}) \\ = &\int_{\boldsymbol{x} \in \{\boldsymbol{\zeta} \in \Delta_{\mathbf{0}} : \zeta_k > 0\}} \sum_{\substack{l_1, \ldots, l_k \geq 1 \\ l_1 + \cdots + l_k = n}} \frac{\binom{n}{l_1\, l_2 \ldots l_k}}{{x_1}^{l_1} \cdot {x_2}^{l_2} \cdot \ldots \cdot {x_k}^{l_k}} \frac{\Xi_0(d\boldsymbol{x})}{\sum_i x_i^2}\end{aligned}$$

must also hold. Since the map

$$\begin{aligned}&\{\boldsymbol{\zeta} \in \Delta_{\mathbf{0}} : \zeta_k > 0\} \to [0, \infty), \\ &\qquad \boldsymbol{x} \mapsto \sum_{\substack{l_1, \ldots, l_k \geq 1 \\ l_1 + \cdots + l_k = n}} \frac{\binom{n}{l_1\, l_2 \ldots l_k}}{{x_1}^{l_1} \cdot {x_2}^{l_2} \cdot \ldots \cdot {x_k}^{l_k} \cdot \sum_i x_i^2}\end{aligned}$$

is strictly positive and $\Xi_0(\{\boldsymbol{\zeta} \in \Delta_{\mathbf{0}} : \zeta_k > 0\}) > 0$ holds by assumption, it follows that

$$\int_{\boldsymbol{x} \in \{\boldsymbol{\zeta} \in \Delta_{\mathbf{0}} : \zeta_k > 0\}} \sum_{\substack{l_1, \ldots, l_k \geq 1 \\ l_1 + \cdots + l_k = n}} \frac{\binom{n}{l_1\, l_2 \ldots l_k}}{{x_1}^{l_1} \cdot {x_2}^{l_2} \cdot \ldots \cdot {x_k}^{l_k}} \frac{\Xi_0(d\boldsymbol{x})}{\sum_i x_i^2} > 0$$

must hold. This in turn establishes a non-trivial lower bound on $\mathbb{P}(\tau_k^{(n)} < \infty)$.

bioRxiv preprint doi: https://doi.org/10.1101/025684; this version posted February 6, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

In the case $\mathrm{supp}(\Xi) \cap \{\boldsymbol{\zeta} \in \Delta_{\mathbf{0}} : \sum_i \zeta_i < 1\} \neq \emptyset$, we again rely on the Poisson point process construction. For $\boldsymbol{x} \in \{\boldsymbol{\zeta} \in \Delta_{\mathbf{0}} : \sum_i \zeta_i < 1\}$, denote $\|\boldsymbol{x}\|_1 := \sum_i x_i < 1$. Let $\{\|\boldsymbol{\zeta}\|_1 < 1\}$ be shorthand for the event

$$(\tau, \boldsymbol{\zeta}, \underline{u}) \in [0, \infty) \times \{\boldsymbol{\zeta} \in \Delta_{\mathbf{0}} : \|\boldsymbol{\zeta}\|_1 < 1\} \times [0, 1]^n.$$

In a manner analogous to the previous case, we can establish that

$$\mathbb{P}(\tau_k^{(n)} < \infty) \geq \int_{\boldsymbol{x} \in \{\boldsymbol{\zeta} \in \Delta_{\mathbf{0}} : \|\boldsymbol{\zeta}\|_1 < 1\}} \mathbb{P}(\tau_k^{(n)} = \tau | \boldsymbol{\zeta} = \boldsymbol{x}) \mathbb{P}(\boldsymbol{\zeta} \in d\boldsymbol{x})$$

holds. Again, the integrand is strictly positive. This time it is bounded from below by

$$\boldsymbol{x} \mapsto \frac{\binom{n}{n-k+1}}{{x_1}^{n-k+1}(1 - \|\boldsymbol{x}\|)^{k-1}} = \mathbb{P}(\#I_1 = n - k + 1, \#I_2 = 0, \ldots | \boldsymbol{\zeta} = \boldsymbol{x}).$$

Since $\mathbb{P}(\{\boldsymbol{\zeta} \in \Delta_{\mathbf{0}} : \|\boldsymbol{\zeta}\|_1 < 1\}) > 0$ holds by assumption, this again establishes a non-trivial lower bound on the right side of the above inequality.

bioRxiv preprint doi: https://doi.org/10.1101/025684; this version posted February 6, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Verification of the ℓ_2 norm (33)

Table 4: Verification of the ℓ_2 norm (33) for estimation of coalescent parameters (ϑ) associated with Λ- or Ξ-coalescents (Table 2). The 'data' are branch lengths associated with given coalescent, and parameter estimates (mean $\overline{\vartheta}$; standard deviation $\widehat{\vartheta}$) are given for the same coalescent using the ℓ_2 norm. Based on number of leaves $n = 50$, and 10^5 replicates.

$\Pi(\vartheta)$	$\overline{\vartheta}$	$\widehat{\vartheta}$
$\Xi(0.05)$	0.06	0.041
$\Xi(0.95)$	0.83	0.191
$\Xi(1.0)$	1.08	0.132
$\Xi(1.5)$	1.48	0.228
$\Lambda(0.05)$	0.06	0.041
$\Lambda(0.95)$	0.83	0.192
$\Lambda(1.0)$	1.08	0.131
$\Lambda(1.5)$	1.47	0.228