The TMRCA of general genealogies in populations of variable size

Alejandro H. Wences[1], Lizbeth Peñaloza[2], Matthias Steinrücken[3], and Arno Siri-Jégousse [4]

[1]LAAS - CNRS, Université de Toulouse, France.
[2]Instituto de Investigación de Matemáticas y Actuaría, Universidad del Mar, campus Huatulco, México.
[3]Department of Ecology and Evolution, University of Chicago, USA.
[3]Department of Human Genetics, University of Chicago, USA.
[4]Instituto de Investigaciones en Matemáticas Aplicadas y Sistemas, Universidad Nacional Autónoma de México, México.

September 9, 2024

Abstract

We study the time to the most recent common ancestor of a sample of finite size in a wide class of genealogical models for populations with variable size. This is made possible by recently developed results on inhomogeneous phase-type random variables, allowing us to obtain the density and the moments of the TMRCA of time-dependent coalescent processes in terms of matrix formulas. We also provide matrix simplifications permitting a more straightforward calculation. With these results, the TMRCA provides an explicative variable to distinguish different evolutionary scenarios.

Keywords: Coalescent theory, Phase-type theory, Variable size population.

1 Introduction

The general theory of coalescent processes aims to provide a rigorous mathematical framework that can be used to model natural phenomena where a collection of particles fuse together as the system evolves over time. It has a variety of applications

in distinct disciplines, such as physics and biology. In biology, particularly in the field of population genetics, it is used to model the parental relationships of a given sample or population as we trace the ancestry of individuals backwards in time, thus constructing a genealogical tree. In this setting, the coalescence of particles occurs at the time when a group of individuals has a common ancestor in the past. Once we have a suitable coalescent model for the genealogy of a population, we can employ mathematical tools to tackle biological questions, such as determining the time needed to reach the most recent common ancestor of the sample or population (TMRCA), the expected genetic diversity for neutral positions of the genome, or whether natural selection has played an important role in the evolution of the population.

In mathematical population genetics, the study of coalescent processes focuses on two main questions. On the one hand, a lot of effort is made to establish a parity between coalescent processes and population models with varying biological assumptions, such as constant or varying population size, the presence of mutation, the strength of natural selection, the effect of genetic drift, spatial constraints, as well as dormancy/latency mechanisms as in virus populations. The motivation behind this effort is that once a coalescent model is inferred for the genealogy of a population from genetic data, this parity may allow for the inference of the evolutionary forces that significantly influenced the dynamics of the population. One of the first coalescent models, Kingman's coalescent, was established as the null model for the genealogies of populations evolving at equilibrium, i.e., of constant size, evolving under neutrality, and with low variance reproductive laws [31]. In more recent years, the Bolthausen-Sznitman coalescent has emerged as an alternative null model for the genealogy of constant-size populations that nonetheless are subject to the effect of natural selection [12, 38, 13, 39]. The genealogies of populations with stochastically varying population size, or evolving in a random environment, have also been addressed; for example, neutral populations undergoing recurrent (i.e. i.i.d. across generations) bottlenecks were studied in [29, 21, 20, 42] for both high and low-variance reproductive laws. Also, neutral populations evolving under deterministically-varying population size and with low-variance reproductive laws were studied in the seminal work of [24]; the coalescent that describes the genealogies of these populations is a time-inhomogeneous coalescent process that can be expressed as a deterministically-time-changed Kingman's coalescent.

On the other hand, the theoretical characterization of different functionals on coalescent processes such as the tree height, the total tree length, and/or the size of external and internal branches, provides inference tools for distinct aspects of the evolutionary past of a population, such as the forces at play throughout its history, the presence of bottlenecks, the TMRCA, etc. In this work, we are interested in the study of the density and moments of the TMRCA for time-inhomogeneous coalescent processes describing the genealogies of populations evolving under deterministically-varying population size. This functional, apart from being interesting as a mathematical object in its own right, is very useful as a step-variable in applications such as the inference of demographic history (see e.g. [28, 44]) or in the computation of the expected SFS [43]. This variable was previ-

ously analyzed for particular examples such as Kingman's coalescent with general time-change in [24] (first moment), but also in [46] (second moment), and in [15] (any moment). For general coalescent models with piecewise constant time-change, the first moment was established in [43]. All these methods are hard to generalize due to analytical difficulties caused by the time dependence and combinatorial issues when trying to consider more general models, higher moments, or their density function. Here, we provide a new technique based on inhomogeneous phase-type theory, developed in [2], to efficiently obtain any moment and the density of the TMRCA for general markovian genealogical models, under any sufficiently smooth time change.

2 The model

Time-changed coalescents. We consider the general class of deterministically time-changed Ξ-coalescents characterized by a finite measure Ξ on the simplex $\{(p_i)_{i>1} : \sum_{i=1}^{\infty} p_i \leq 1\}$, and a deterministic time change function $\zeta \colon \mathbb{R}_+ \to \mathbb{R}_+$. For a given $n > 0$, these are pure-jump non-homogeneous Markov processes with state space the set of partitions of $[n] := \{1, \ldots, n\}$. States represent configurations of ancestral lineages, starting with the partition into singletons $\{\{1\}, \ldots, \{n\}\}$, and being absorbed in the state $\{\{1, \ldots, n\}\}$, where one single lineage is ancestral to all particles. Specifically, at time $t \geq 0$ a jump directed by $\boldsymbol{p} = (p_i)_{i\geq 1}$ occurs at rate $\frac{1}{\zeta(t)} \frac{\Xi(d\boldsymbol{p})}{\sum_{i=1}^{\infty} p_i^2}$; at such an event, each active lineage (i.e. block of the current partition) is independently and uniformly placed in the interval $[0, 1]$, which is divided into subintervals of lengths $(p_i)_{i\geq 1}$. The new state of the system (a coarser partition of $[n]$) is constructed by merging (coalescing) all the lineages that fall into the same subinterval. Note that the lineages falling outside of any subinterval do not participate in any coagulation (this is Kingman's paintbox construction, see e.g. [4]). When the measure Ξ is supported on the set $\{(p_i)_{i\geq 1} : p_1 \in (0, 1], \text{ and } p_i = 0 \quad \forall i \geq 2\}$ the corresponding coalescent is called a multiple merger coalescent; in this case we denote by Λ the push-forward measure of Ξ on $[0, 1]$ after projecting into the p_1 coordinate. Finally, our results also apply to coalescent processes with Kingman's dynamics, in which every pair of blocks independently coalesce at an additional rate of $\frac{c}{\zeta(t)}$ for some $c > 0$.

Established coalescence measures/processes and their associated population dynamics are:

- Kingman's coalescent ($c > 0$ and $\Xi = 0$): For populations evolving at equilibrium.
- Beta coalescents ($\Lambda \sim \texttt{Beta}(2 - \alpha, \alpha)$, $1 \leq \alpha < 2$): For populations with skewed offspring distribution [40] or selection; this class includes as limit cases the Kingman coalescent ($\alpha \to 2$, neutral evolution) and the Bolthausen-Sznitman coalescent ($\alpha = 1$, strong selection).

- Psi coalescent: For populations with skewed offspring distribution or selection [14].
- Beta-Ξ coalescent: Modelling diploid reproduction [6].
- Symmetric or General Dirichlet coalescent: For populations with recurrent drastic bottlenecks [21, 20].
- Ξ^β coalescent: Modelling seed bank effects [23].

The time changes arise from distinct assumptions on the dynamics on the total population size, originally in the work of Griffiths and Tavaré [24] (Theoretical Biology) or of Möhle [36], Kaj and Krone [29] (Probability). The population size can increase or decrease, for more details, see also [18]. Common examples of time-changes functions ζ are

- Exponential growth [25]: $\zeta(t) = e^{-\rho t}$.
- Frequent bottlenecks [15]: $\zeta(t) = 1 + \varepsilon \sin(\omega t)$.
- Piecewise constant function ζ [33, 43].
- Piecewise exponential function ζ [5].

We can also consider variations of coalescent processes representing genealogies of populations with more complex evolutionary scenarios such as recombination graphs with a finite number of loci [41], seed bank coalescents [22], or multispecies coalescents [8, 9, 32]. Our method can be applied in any scenario where the process describing the number of lineages is a continuous time Markov chain in a finite state space.

3 Methods

Our approach will make use of phase-type distributions, which define a class of random variables including sums and mixtures of exponentials. They are commonly defined as the time to absorption of a continuous-time Markov chain. Methods based on phase-type distributions have been used in biology and medicine [1, 34, 16] and, more recently, in population genetics [27, 26], in particular, to examine balancing selection [45], or seed bank dynamics [10, 22].

Basics on phase-type theory. As shown in [27] for time-homogeneous Markovian genealogies, some of the most important statistics on coalescent processes such as the TMRCA and the total branch length can be cast into the phase-type framework, leading to explicit expressions for their density and their moments in terms of a suitable chosen rate matrix of a Markov process with absorption. In this work we extend these results to the non-time-homogeneous setting, and provide simple formulas for the density and the moments whenever possible.

Formally, consider a Markov chain with n states and time-inhomogeneous transition matrix of size $n \times n$ given by

$$\mathbf{Q}(t) = \begin{pmatrix} \mathbf{S}(t) & \boldsymbol{s}(t) \\ \mathbf{0} & 0 \end{pmatrix}. \tag{1}$$

The last row of zeros in this matrix corresponds to an absorbing state. The column vector $\boldsymbol{s}(t)$ of size $n-1$ gives the infinitesimal jump rates of the process at time t from any non-absorbing state to the absorbing state. The matrix $\mathbf{S}(t)$ of size $(n-1) \times (n-1)$ gives the infinitesimal jump rates of the process at time t from any non-absorbing state to another non-absorbing state, and the negative value of the total jump rates from non-absorbing states on the diagonal. The whole matrix $\mathbf{Q}(t)$ can thus be recovered from $\mathbf{S}(t)$.

With these definitions, the absorption time τ of the Markov process is *inhomogeneous phase-type distributed* with parameter $\mathbf{S}(t)$. The starting point plays in general an important role, but in our setting we will always consider that the chain starts almost surely at the first state, corresponding to the first row of the matrix $\mathbf{Q}(t)$, and omit it from the notation. Thus we will simply write $\tau \sim IPH(S(t))$ for the absorption time of the processes starting at the first state. The density, the Laplace transform and the moments of τ can be computed in terms of the matrix $\mathbf{S}(t)$ and the vector $\boldsymbol{s}(t)$. Note that in the time-homogeneous case, the distribution of τ corresponds to a sum of a random number of exponential random variables.

4 Results

In all time-changed coalescent models considered in this paper, the TMRCA corresponds to the time that the coalescent process reaches its absorbing state, the state where only a single lineage is left that is ancestral to all particles. The TMRCA thus has an inhomogeneous phase-type distribution associated with a transition matrix of the form

$$\mathbf{Q}_T(t) = \frac{1}{\zeta(t)} \begin{pmatrix} \mathbf{S} & \boldsymbol{s} \\ \mathbf{0} & 0 \end{pmatrix}, \tag{2}$$

where ζ is the time-change function. This special case is simpler than the general case and is treated in Theorems 2.8 and 2.9 of [2], which lead to the following.

Theorem 4.1. *The TMRCA is equal to $g(X)$ where $X \sim PH(\mathbf{S})$ (homogeneous) and the inverse of g is*

$$g^{-1}(x) = \int_0^x \frac{du}{\zeta(u)}. \tag{3}$$

In particular, its density function is given by

$$f(t) = \frac{1}{\zeta(t)} \boldsymbol{\alpha} \exp\left(\mathbf{S} \int_0^t \frac{du}{\zeta(u)}\right) \boldsymbol{s}, \tag{4}$$

where $\boldsymbol{\alpha} = (1, 0, \dots)$*. Also, the* k*-th moment of the TMRCA is given by*

$$m_k = \boldsymbol{\alpha} L_{g^k}(-\mathbf{S})\boldsymbol{s} = \boldsymbol{\alpha}\left(\int_0^\infty g^k(x)e^{\mathbf{S}x}dx\right)\boldsymbol{s} \tag{5}$$

where L_{g^k} *denotes the Laplace transform of* g^k*, parametrized by the matrix* $-\mathbf{S}$*.*

The Laplace transform applied to a matrix in equation (5) is difficult to implement. In Theorem 4.2 we provide a modification of (5) involving the Laplace transform applied to each eigenvalue, which significantly eases the computation. This works when the genealogical model considered is given by any time-inhomogeneous Ξ-coalescent (see also Remark 1).

Theorem 4.2. *For any* Ξ*-coalescent starting with* n *particles there exists a matrix* $\mathbf{P} \equiv \mathbf{P}_{\Xi,n}$ *and a vector* $\boldsymbol{s}_n \equiv \boldsymbol{s}_{\Xi,n}$ *such that for any deterministic time-change function* $\zeta(t)$ *with* $g(t)$ *as in* (3) *we have, for the density* f *of the TMRCA,*

$$f(t) = \frac{1}{\zeta(t)}\boldsymbol{\alpha}\mathbf{P}^{-1}\begin{pmatrix}\exp\{-q_n\int_0^t \frac{du}{\zeta(u)}\} & & 0\\ & \ddots & \\ 0 & & \exp\{-q_2\int_0^t \frac{du}{\zeta(u)}\}\end{pmatrix}\mathbf{P}\boldsymbol{s}_n, \tag{6}$$

where q_j *is the total jump rate of the* Ξ*-coalescent when there are* j *particles. Similarly, for the* k*-th moment of the TMRCA,*

$$m_k = \boldsymbol{\alpha}\mathbf{P}^{-1}\begin{pmatrix}L_{g^k}(q_n) & & 0\\ & \ddots & \\ 0 & & L_{g^k}(q_2)\end{pmatrix}\mathbf{P}\boldsymbol{s}_n. \tag{7}$$

Proof. Consider the Markov chain rate matrix $\mathbf{Q}$ where $\mathbf{Q}_{i,j}$ is the rate at which the (homogeneous) block-counting process jumps from $n-i+1$ to $n-j+1$, where $1 \le i < j \le n$. Note that the corresponding matrix $\mathbf{S}$ defined in equation (2) is upper triangular and diagonalizable. In particular,

$$\mathbf{S} = \mathbf{P}^{-1}\mathbf{D}\mathbf{P}, \tag{8}$$

where the rows of $\mathbf{P}$ are the (left) eigenvectors of $\mathbf{S}$ and $\mathbf{D}_{i,i} = \mathbf{S}_{i,i}$ for $1 \le i < n$, since the eigenvalues of $\mathbf{S}$ are $\{\mathbf{S}_{1,1}, \dots, \mathbf{S}_{n-1,n-1}\}$. Then, since $\boldsymbol{\alpha} = (1, 0, \dots, 0)$,

$$\begin{aligned} m_k &= \boldsymbol{\alpha}\mathbf{P}^{-1}\left(\int_0^\infty g^k(x)e^{\mathbf{D}x}dx\right)\mathbf{P}\boldsymbol{s}\\ &= \boldsymbol{\alpha}\mathbf{P}^{-1}\begin{pmatrix}L_{g^k}(-\mathbf{D}_{1,1}) & & 0\\ & \ddots & \\ 0 & & L_{g^k}(-\mathbf{D}_{n-1,n-1})\end{pmatrix}\mathbf{P}\boldsymbol{s}\\ &= \left(\mathbf{P}_{1,1}^{-1}L_{g^k}(-\mathbf{D}_{1,1}), \dots, \mathbf{P}_{1,n-1}^{-1}L_{g^k}(-\mathbf{D}_{n-1,n-1})\right)\mathbf{P}\boldsymbol{s}, \end{aligned}$$

and, similarly,

$$f(t) = \frac{1}{\zeta(t)} \left(\mathbf{P}_{1,1}^{-1} e^{\mathbf{D}_{1,1} g^{-1}(t)}, \ldots, \mathbf{P}_{1,n-1}^{-1} e^{\mathbf{D}_{n-1,n-1} g^{-1}(t)} \right) \mathbf{P} \boldsymbol{s}.$$

The statement of the theorem follows from the observation that $\mathbf{D}_{i,i} = \mathbf{S}_{i,i} = -q_{n-i+1}$. □

Remark 1. *The above proof rests solely on the (left) eigen-value decomposition of* $\mathbf{S}$ *as in* (8)*. Thus, equations* (6) *and* (7) *remain valid for any inhomogeneous phase-type distribution with transition matrix of the form* (2) *satisfying* (8)*, as long as the corresponding eigenvectors* $\mathbf{P}$ *and eigenvalues (which will replace* $(-q_1, \cdots, -q_n)$ *in* (6) *and* (7)*) are used.*

The following lemma aims at easing the computation of the vector $\boldsymbol{\alpha} \mathbf{P}_{\Xi,n}^{-1}$.

Lemma 4.3. *Let* $\mathbf{P}_{\Xi,n}$ *be as in Theorem 4.2.*

1) *The matrix* $\mathbf{P}_{\Xi,n}$ *can be obtained by removing the first row and the first column of* $\mathbf{P}_{\Xi,n+1}$.
2) *The* $(1,i)$*-th entry of* $\mathbf{P}^{-1} \equiv \mathbf{P}_{\Xi,n}^{-1}$ *is given by*

$$\mathbf{P}_{1,i}^{-1} = (-1)^{1+i} \frac{\det\left(\mathbf{P}_{\{1,\ldots,i-1\},\{2,\cdots,i\}}\right)}{\prod_{j=1}^{i} \mathbf{P}_{j,j}} \tag{9}$$

where, for $I, J \subset [n-1]$*, we have used the notation* $\mathbf{P}_{I,J}$ *to denote the matrix* $(\mathbf{P}_{i,j})_{i \in I, j \in J}$.

Proof. Item 1) follows from the fact that $\mathbf{S} \equiv \mathbf{S}_n$ is upper triangular for every n and that $\mathbf{S}_n$ can be obtained from $\mathbf{S}_{n+1}$ by removing its first row and column.

To prove item 2), note that if $x^{(i)} = (x_1^{(i)}, \cdots, x_{n-1}^{(i)})^T$ is the solution to

$$\mathbf{P} x^{(i)} = e_i,$$

where e_i is the i-th unit vector, then, by Cramer's rule,

$$\mathbf{P}_{1,i}^{-1} = x_1^{(i)} = \frac{\det(\mathbf{P}^{(i)})}{\det(\mathbf{P})}$$

where $\mathbf{P}^{(i)}$ is constructed from $\mathbf{P}$ by replacing the first column by the vector e_i. Computing the determinant of $\mathbf{P}^{(i)}$ along the first column using the Laplace expansion then gives

$$\det(\mathbf{P}^{(i)}) = (-1)^{1+i} \det\left(\mathbf{P}_{[n-1]\setminus\{i\},[n-1]\setminus\{1\}}\right), \tag{10}$$

since all elements of the first column are zero, except the i-th entry, which is equal to one.

Since $\mathbf{S}$ is upper-triangular, the matrix $\mathbf{P}$ of its left eigenvectors is upper-triangular, and so is $\mathbf{P}_{\{i+1,\cdots,n-1\},\{i+1,\cdots,n-1\}}$. This block structure for the determinant on the right-hand side of (10) implies

$$\det(\mathbf{P}^{(i)}) = (-1)^{1+i} \left(\prod_{j=i+1}^{n-1} \mathbf{P}_{j,j} \right) \det \left(\mathbf{P}_{\{1,\ldots,i-1\},\{2,\cdots,i\}} \right)$$

and

$$\mathbf{P}_{1,i}^{-1} = (-1)^{1+i} \frac{\det \left(\mathbf{P}_{\{1,\ldots,i-1\},\{2,\cdots,i\}} \right)}{\prod_{j=1}^{i} \mathbf{P}_{j,j}}.$$

□

Equation (9) can be evaluated for all $1 \leq i \leq n-1$ efficiently using a dynamic program that reuses computations for $i-1$ in the computations for i. To this end, define

$$G(i) := \det \left(\mathbf{P}_{\{1,\ldots,i\},\{2,\cdots,i+1\}} \right) \tag{11}$$

and

$$H(i,j) := \det \left(\mathbf{P}_{\{1,\ldots,i\},\{2,\cdots,i\}\cup\{j\}} \right) \tag{12}$$

for $1 \leq i < n-1$ and $i+1 < j \leq n-1$. Now note that for all i and j, the matrices in the definitions (11) and (12) only have two non-zero entries in the last row, specifically, in the last two columns. We can thus compute the determinants along the last row using the Laplace expansion to show that the recurrence relations

$$\begin{aligned} G(i) &= \det \left(\mathbf{P}_{\{1,\ldots,i\},\{2,\cdots,i+1\}} \right) \\ &= (-1)^{i+i-1} \mathbf{P}_{i,i} \det \left(\mathbf{P}_{\{1,\ldots,i-1\},\{2,\cdots,i-1\}\cup\{i+1\}} \right) \\ &\quad + (-1)^{i+i} \mathbf{P}_{i,i+1} \det \left(\mathbf{P}_{\{1,\ldots,i-1\},\{2,\cdots,i\}} \right) \\ &= -\mathbf{P}_{i,i} H(i-1,i+1) + \mathbf{P}_{i,i+1} G(i-1) \end{aligned} \tag{13}$$

and

$$\begin{aligned} H(i,j) &= \det \left(\mathbf{P}_{\{1,\ldots,i\},\{2,\cdots,i\}\cup\{j\}} \right) \\ &= (-1)^{i+i-1} \mathbf{P}_{i,i} \det \left(\mathbf{P}_{\{1,\ldots,i-1\},\{2,\cdots,i-1\}\cup\{j\}} \right) \\ &\quad + (-1)^{i+i} \mathbf{P}_{i,j} \det \left(\mathbf{P}_{\{1,\ldots,i-1\},\{2,\cdots,i\}} \right) \\ &= -\mathbf{P}_{i,i} H(i-1,j) + \mathbf{P}_{i,j} G(i-1) \end{aligned} \tag{14}$$

hold. The functions $G(i)$ and $H(i,j)$ for all j can then be computed iteratively, starting with $i=1$ and incrementing i by 1 until $i=n-2$, reusing the values of $G(\cdot)$ and $H(\cdot,\cdot)$ from the previous step. Once $G(i)$ is computed for all i, equation (9) can be readily evaluated for all i. Note that we did observe improved numerically stability when absorbing all but the highest factor of the product in the denominator of equation (9) into $\mathbf{P}$ in the numerator by dividing each row i with the respective value $\mathbf{P}_{i,i}$ on the diagonal.

Example (Populations with recurrent bottlenecks): In Figure 1 we compare two models for populations undergoing recurrent bottlenecks, one homogeneous and one inhomogeneous. The first one consists of the symmetric coalescent introduced in [21]. The latter is a special case of Ξ-coalescents that arise from Wright-Fisher models that suffer from drastic decays of the population size for one generation; these decays occur at a constant rate. The symmetric coalescent is characterized by a function F on $\mathbb{Z}$ such that $F(0) < \infty$ and $\sum_{k\geq 1} F(k)/k < \infty$; for convenience we also introduce a scalar parameter $A > 0$ modulating the overall coagulation rate. The dynamics are described as follows: when there are b blocks in the coalescent, at rate $AF(k)$, we distribute the b blocks into k boxes uniformly at random, and blocks falling in the same box merge. See also [42] for more general models of this type.

The second model for the genealogies of populations undergoing recurrent bottlenecks is the Kingman's coalescent with sinusoidal time change introduced in [15]. This model corresponds in our framework to setting $\zeta(t) = B(1 + \varepsilon \sin(\omega t))$ where the parameter $B > 0$ gives the rate at which pair-wise merges occur, and ε and ω relate to the size and frequency of the bottleneck events.

In Figure 1 we compare the density of the TMRCA of the above two models. For the symmetric coalescent we set F to be the density of a Poisson r.v. of parameter λ which we set to $\lambda = n\varepsilon$ where n is the initial number of blocks; whereas A is set to $A = 1/\omega$. For the sinusoidal Kingman coalescent we fix two distinct combinations of ε and ω, and then choose B so that the expectation $\mathbb{E}[TMRCA]$ is equal to that of the corresponding symmetric model (here, all the expectations $\mathbb{E}[TMRCA]$ are computed using (5) and they are matched numerically). The top of Figure 1 corresponds to the choice $\varepsilon = 0.8$ and $\omega = 0.5$, whereas the bottom corresponds to $\varepsilon = 0.5$ and $\omega = 1$. It is interesting to note that, depending on the choice of the parameters ε and ω, the density of the TMRCA of the corresponding sinusoidal Kingman's coalescent can be made to be multimodal, resembling the density of a discrete random variable. On the other hand, the density corresponding to the symmetric coalescent remains unimodal and appears as a continuous approximation of that of the sinusoidal Kingman's case.

Example (Multiple merger coalescents with exponential growth): For the special case of exponential growth, the time-scale function is given by $\zeta(t) = e^{-\rho t}$, and we obtain $g^{-1}(x) = \rho^{-1}(e^{\rho x} - 1)$ and $g(x) = \rho^{-1}\log(1 + \rho x)$. Thus

$$L_{g^k}(s) = \int_0^\infty \left(\rho^{-1}\log(1+\rho x)\right)^k e^{-sx} dx$$

and, for any Ξ-coalescent,

$$f(t) = e^{\rho t}\boldsymbol{\alpha}\mathbf{P}_{\Xi,n}^{-1}\begin{pmatrix} \exp\{-q_n\frac{e^{\rho t}-1}{\rho}\} & & 0 \\ & \ddots & \\ 0 & & \exp\{-q_2\frac{e^{\rho t}-1}{\rho}\}\end{pmatrix}\mathbf{P}_{\Xi,n}\boldsymbol{s}_{\Xi,n}$$

follows. In Figure 2, we show the density of the TMRCA for Kingman's coalescent and the Bolthausen-Sznitman coalescent with $n = 30$ for different values of ρ. In the

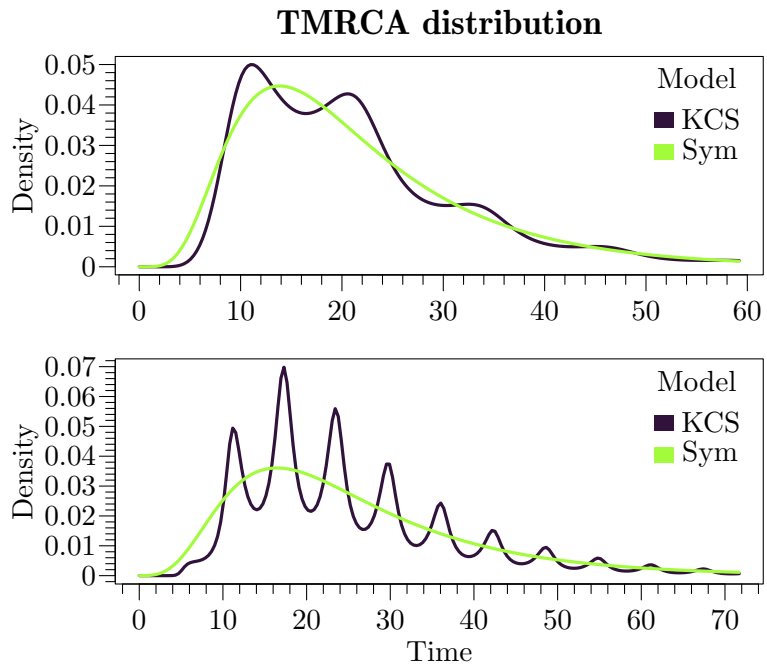

Figure 1: The density of the TMRCA for the two models with recurrent bottlenecks described in the main text, the sinusoidal Kingman's coalescent (KCS) and the symmetric coalescent (Sym) with $n = 30$. Upper panel: $\varepsilon = 0.8$, $\omega = 0.5$; lower panel: $\varepsilon = 0.5$, $\omega = 1$.

same scenarios, Figure 3 depicts the respective moments m_k for different values of k. To validate our analytic formulas for the densities, we compared them to values estimated from 40,000 simulated replicates of the underlying process. The results are shown in Figure 4 and we observe that the analytic formulas and simulations match well.

Example (Dormancy in a population with exponential growth): An example of a Markovian genealogical model that does not belong to the class of multiple merger coalescents is the seed bank coalescent. It models the limit genealogies of populations undergoing strong dormancy phenomena, i.e. individuals can remain inactive for a large amount of generations. The seed bank coalescent is a Kingman coalescent with lineages being active or inactive. Every pair of active lineages merges at rate 1. Moreover, each of its active lineages gets deactivated at rate $c_1 > 0$ and inactive lineages activate at rate $c_2 > 0$. This model has been introduced in [11] and studied for population genetics applications in [22]. It is known from the literature that its TMRCA behaves like $\log\log(n)$, which is also the limiting behavior under the Bolthausen-Sznitman coalescent. Since the latter is used to model genealogies of rapidly evolving populations, we examine in more detail if the

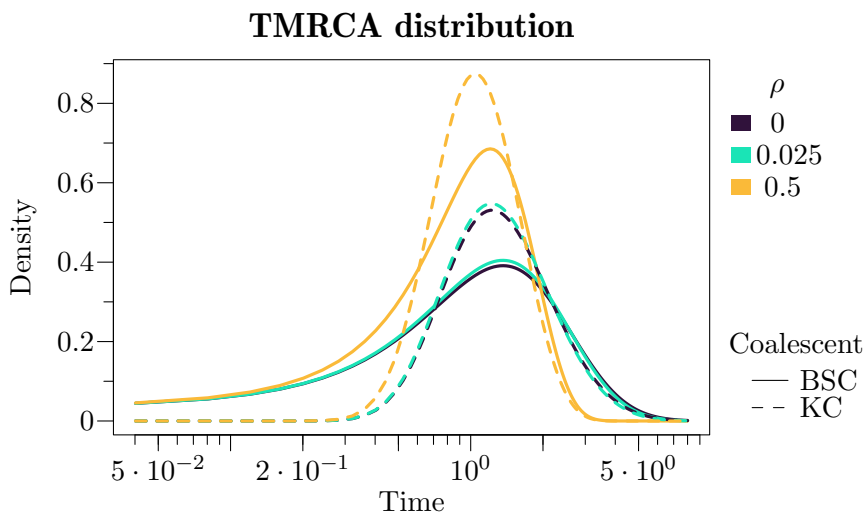

Figure 2: The density of the TMRCA for different choices of ρ in the exponential growth model, for Kingman's (KC) and the Bolthausen-Sznitman coalescent (BSC) with $n = 30$.

TMRCA can provide a statistic to discriminate dormancy versus rapid evolution, when the population size also varies.

Recently in [17], it was shown that the TMRCA of the seed bank coalescent for large sample size n behaves as

$$TMRCA \approx \frac{\log\log(n)}{c_2} + \frac{\log(2c_1)}{c_2} + G \tag{15}$$

where G is a standard Gumbel random variable. On the other hand, it behaves in the Bolthausen-Sznitman case as

$$TMRCA \approx \log\log(n) - \log(E)$$

where E is a standard exponential r.v. [19, 37, 30]. Figure 5 shows the differences between the densities in both models with exponential growth (in the case $c_1 = c_2 = 1$) computed using our approach. Here we can see that the difference between both densities is mainly explained by the choice of the parameters, the constant $\log(2c_1)/c_2$ appearing in (15) and the limit distributions. This is still the case in the exponential growth regime, though the growth rate seems to reduce the differences between densities tails.

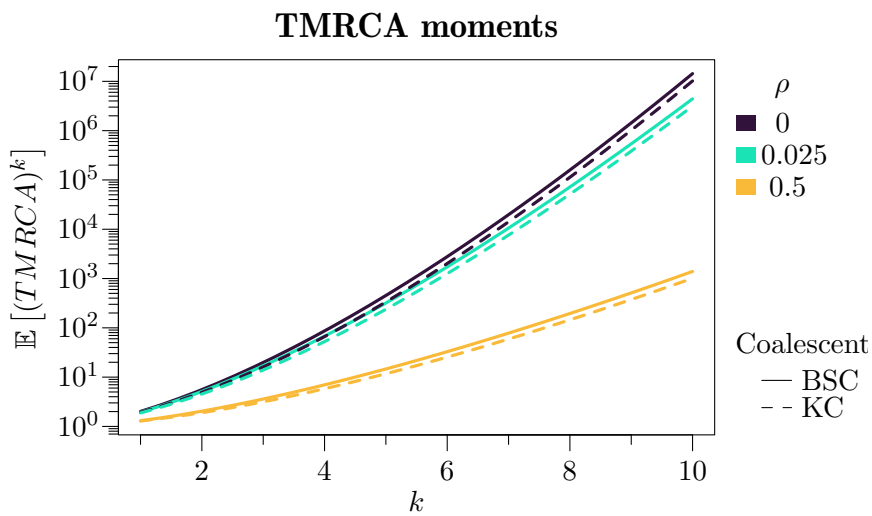

Figure 3: Moments of the TMRCA for different choices of order $k \in [1, 10]$ and exponential growth parameter ρ, for Kingman's (KC) and Bolthausen-Sznitman coalescent (BSC) with $n = 30$.

5 Discussion

In this manuscript, we exhibited a connection between general genealogical models of varying-sized populations and the inhomogeneous phase-type theory described in [2]. We enriched this theory with interesting applications in population genetics, where the IPH theory provides explicit formulas for the TMRCA. In particular, we obtained expressions for its density and moments in a wide class of time-inhomogeneous coalescent processes, improving previous results in the literature and also generalizing them to a much wider spectrum of models, including those that involve coalescents with simultaneous multiple mergers. This method is notably robust and can be applied to any Markovian genealogy starting with finitely many individuals. It also significantly eases the computational load present in inference applications by separating the effects of the time-inhomogeneity (the time change ζ) and the coalescent dynamics (i.e., the coagulation rates).

Unfortunately, this straightforward method does not readily generalize to other summary statistics, such as the total branch length or the site frequency spectrum (SFS). On the one hand, it can be easily shown that the total branch length is also IPH distributed; nonetheless, in this case, the corresponding transition matrix is not easily factorized into a time-inhomogeneity and a coalescent component. Indeed, its density and moments can be expressed as a product integral of a time-

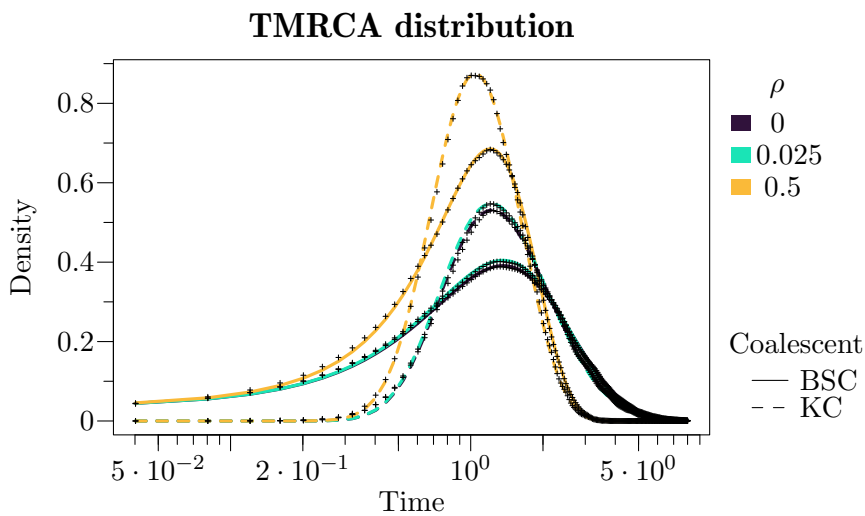

Figure 4: The density of the TMRCA for different choices of the exponential growth parameter ρ, for both Kingman's (KC) and the Bolthausen-Sznitman coalescent (BSC) with $n = 30$. We compare the values obtained from our analytical formulas against values estimated from 40,000 simulated replicates (indicated by plus signs).

dependent matrix for which computational methods must be developed. We note, for example, that the computation of this product integral can be recast in terms of PDE's by adapting techniques of [35] or [7, Ch. 8.1.3]; however, the complexity and the substantially different nature of this approach make it fall out of the scope of this article.

On the other hand, the study of the SFS motivates the development of a multivariate IPH theory. In addition, this could also provide interesting insights into the covariance of the TMRCA and/or the total branch length. For now, this multivariate setting can only be established when the respective IPH random variables are of the form (2) (see [3]). These advances could also be essential in studying multivariate genealogical models such as recombination trees.

Acknowledgement

This project was partially supported by DGAPA - PAPIIT grant IN102824, DGAPA - PASPA support, and by the National Institute of General Medical Sciences (NIGMS) of the National Institutes of Health under award R01GM146051 (MS).

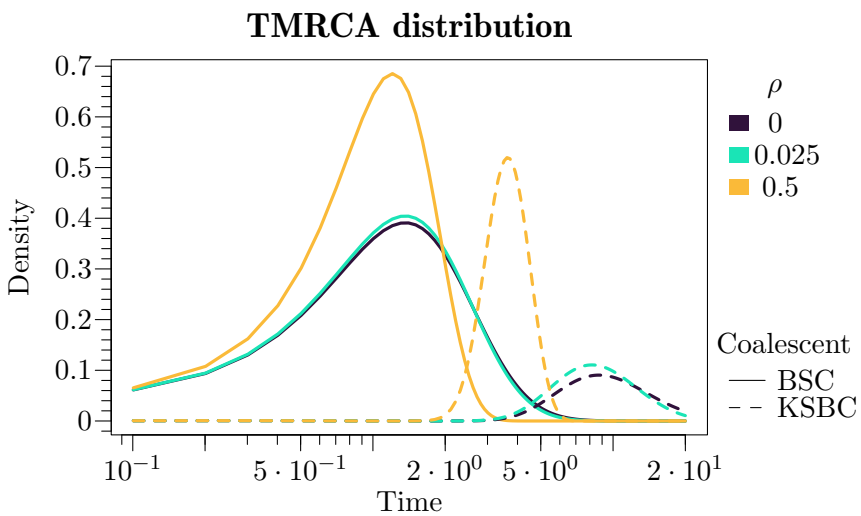

Figure 5: The density of the TMRCA with $n = 30$ for different choices of the exponential growth parameter ρ, for the seed bank (KCSB) with $c_1 = 1$ and $c_2 = 1$, and the Bolthausen-Sznitman coalescent (BSC).

References

[1] Aalen, O.O.: *Phase type distributions in survival analysis.* Scandinavian Journal of Statistics, 22(4):447–463, 1995. `https://www.jstor.org/stable/4616373`.

[2] Albrecher, H. and M. Bladt: *Inhomogeneous phase-type distributions and heavy tails.* Journal of Applied Probability, 56(4):1044–1064, 2019.

[3] Albrecher, H., M. Bladt, and J. Yslas: *Fitting inhomogeneous phase-type distributions to data: the univariate and the multivariate case.* Scandinavian Journal of Statistics, 49(1):44–77, 2022.

[4] Bertoin, J.: *Random fragmentation and coagulation processes.* Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2006.

[5] Bhaskar, Anand, Y.X. Rachel Wang, and Yun S. Song: *Efficient inference of population size histories and locus-specific mutation rates from large-sample genomic variation data.* Genome Research, 25(2):268–279, 2015.

[6] Birkner, M., H. Liu, and A. Sturm: *Coalescent results for diploid exchangeable population models.* Electronic Journal of Probability, 23(none), 2018.

[7] Bladt, M. and B.F. Nielsen: *Matrix-Exponential Distributions in Applied Probability*. Springer US, May 2017, ISBN 149397047X. https://www.ebook.de/de/product/28657748/mogens_bladt_bo_friis_nielsen_matrix_exponential_distributions_in_applied_probability.html.

[8] Blancas, A., J. J. Duchamps, A. Lambert, and A. Siri-Jégousse: *Trees within trees: simple nested coalescents*. Electronic Journal of Probability, 23(none), jan 2018.

[9] Blancas, A., T. Rogers, J. Schweinsberg, and A. Siri-Jégousse: *The nested kingman coalescent: Speed of coming down from infinity*. The Annals of Applied Probability, 29(3), jun 2019.

[10] Blath, J., E. Buzzoni, J. Koskela, and M. Wilke Berenguer: *Statistical tools for seed bank detection*. Theoretical Population Biology, 132:1–15, 2020.

[11] Blath, J., A. González Casanova, N. Kurt, and D. Spano: *The ancestral process of long-range seed bank models*. Journal of Applied Probability, 50(3):741–759, 2013.

[12] Brunet, E. and B. Derrida: *Shift in the velocity of a front due to a cutoff*. Physical Review E, 56(3):2597–2604, sep 1997.

[13] Cortines, A. and B. Mallein: *A n-branching random walk with random selection*. Latin American Journal of Probability and Mathematical Statistics, 14(1):117, 2017.

[14] Eldon, B. and J. Wakeley: *Coalescent processes when the distribution of offspring number among individuals is highly skewed*. Genetics, 172(4):2621–2633, 2006.

[15] Eriksson, A., B. Mehlig, M. Rafajlovic, and S. Sagitov: *The total branch length of sample genealogies in populations of variable size*. Genetics, 186(2):601–611, 2010.

[16] Fackrell, M.: *Modelling healthcare systems with phase-type distributions*. Health Care Management Science, 12(1):11–26, 2008.

[17] Fittipaldi, M.C., A. González Casanova, and J.E. Nava: *Lookdown construction for a moran seed-bank model*. arXiv preprint arXiv:2305.12489, 2023.

[18] Freund, F.: *Cannings models, population size changes and multiple-merger coalescents*. Journal of Mathematical Biology, 80(5):1497–1521, 2020.

[19] Goldschmidt, C. and J. Martin: *Random recursive trees and the Bolthausen-Sznitman coalesent*. Electronic Journal of Probability, 10(none):718 – 745, 2005. https://doi.org/10.1214/EJP.v10-265.

[20] González Casanova, A., V. Miró Pina, E. Schertzer, and A. Siri-Jégousse: *Asymptotics of the frequency spectrum for general dirichlet ξ-coalescents*. Electronic Journal of Probability, 29:1–35, 2024.

[21] González Casanova, A., V. Miró Pina, and A. Siri-Jégousse: *The symmetric coalescent and wright–fisher models with bottlenecks*. The Annals of Applied Probability, 32(1), feb 2022.

[22] González Casanova, A., L. Peñaloza, and A. Siri-Jégousse: *The shape of a seed bank tree*. Journal of Applied Probability, 59(3):631–651, 2022.

[23] González Casanova, A., L. Peñaloza, and A. Siri-Jégousse: *Seed bank cannings graphs: How dormancy smoothes random genetic drift*. ALEA, Latin American Journal of Probability and Mathematical Statistics, 20:1165–1186, 2023.

[24] Griffiths, R.C. and S. Tavaré: *Sampling theory for neutral alleles in a varying environment*. Philosophical Transactions: Biological Sciences, 344(1310):403–410, 1994, ISSN 09628436.

[25] Griffiths, R.C. and S. Tavaré: *The age of a mutation in a general coalescent tree*. Communications in Statistics. Stochastic Models, 14(1-2):273–295, 1998.

[26] Hobolth, A., I. Rivas-González, M. Bladt, and A. Futschik: *Phase-type distributions in mathematical population genetics: An emerging framework*. Theoretical Population Biology, 2024.

[27] Hobolth, A., A. Siri-Jégousse, and M. Bladt: *Phase-type distributions in population genetics*. Theoretical Population Biology, 127:16–32, 2019.

[28] Hoscheit, P. and O.G. Pybus: *The multifurcating skyline plot*. Virus evolution, 5(2):vez031, 2019.

[29] Kaj, I. and S.M. Krone: *The coalescent process in a population with stochastically varying size*. Journal of Applied Probability, 40(1):33–48, mar 2003.

[30] Kersting, G., A. Siri-Jégousse, and A. H. Wences: *Site frequency spectrum of the bolthausen-sznitman coalescent*. ALEA Lat. Am. J. Probab. Math. Stat., 18(1):1483, 2021. `https://doi.org/10.30757/alea.v18-53`.

[31] Kingman, J.F.C.: *On the genealogy of large populations*. Journal of Applied Probability, 19(A):27–43, 1982.

[32] Lambert, A. and E. Schertzer: *Coagulation-transport equations and the nested coalescents*. Probability Theory and Related Fields, 176(1-2):77–147, apr 2019.

[33] Li, H. and R. Durbin: *Inference of human population history from individual whole-genome sequences*. Nature, 475(7357):493–496, 2011. `https://doi.org/10.1038/nature10231`.

[34] Marshall, A.H. and S.I. McClean: *Using coxian phase-type distributions to identify patient characteristics for duration of stay in hospital*. Health Care Management Science, 7:285–289, 2004.

[35] Miroshnikov, A. and M. Steinrücken: *Computing the joint distribution of the total tree length across loci in populations with variable size*. Theoretical Population Biology, 118:1–19, dec 2017.

[36] Möhle, M.: *The coalescent in population models with time-inhomogeneous environment.* Stochastic processes and their applications, 97(2):199–227, 2002.

[37] Möhle, M. and H. Pitters: *A spectral decomposition for the block counting process of the Bolthausen-Sznitman coalescent.* Electronic Communications in Probability, 19(none):1 – 11, 2014. `https://doi.org/10.1214/ECP.v19-3464`.

[38] Neher, R.A. and O. Hallatschek: *Genealogies of rapidly adapting populations.* Proceedings of the National Academy of Sciences, 110(2):437–442, dec 2013.

[39] Schertzer, E. and A.H. Wences: *Relative vs absolute fitness in a population genetics model. how stronger selection may promote genetic diversity.* arXiv preprint arXiv:2301.07762, 2023.

[40] Schweinsberg, J.: *Coalescent processes obtained from supercritical galton–watson processes.* Stochastic Processes and their Applications, 106(1):107–139, 2003.

[41] Simonsen, K.L. and G.A. Churchill: *A markov chain model of coalescence with recombination.* Theoretical population biology, 52(1):43–59, 1997.

[42] Siri-Jégousse, A. and A. H. Wences: *Exchangeable coalescents beyond the cannings class.* Preprint arXiv:2212.02154 [math.PR], 2022. `https://arxiv.org/abs/2212.02154`.

[43] Spence, J.P., J.A. Kamm, and Y.S. Song: *The site frequency spectrum for general coalescents.* Genetics, 202(4):1549–1561, 2016.

[44] Upadhya, G. and M. Steinrücken: *Robust inference of population size histories from genomic sequencing data.* PLOS Computational Biology, 18(9):e1010419, 2022.

[45] Zeng, K., B. Charlesworth, and A. Hobolth: *Studying models of balancing selection using phase-type theory.* Genetics, 218(2), 2021.

[46] Živković, D. and T. Wiehe: *Second-order moments of segregating sites under variable population size.* Genetics, 180(1):341–357, 2008.