

Phase-type distributions in mathematical population genetics: An emerging framework

Asger Hobolth^{a,*}, Iker Rivas-González^b, Mogens Bladt^c, Andreas Futschik^d

^a Department of Mathematics, Aarhus University, Denmark

^b Bioinformatics Research Center, Aarhus University, Denmark

^c Department of Mathematical Sciences, University of Copenhagen, Denmark

^d Institute of Applied Statistics, Johannes Kepler University, Austria

ARTICLE INFO

Keywords:

Coalescent
Laplace transform
Likelihood inference
Phase-type theory
Population genetics
Reward transformation

ABSTRACT

A phase-type distribution is the time to absorption in a continuous- or discrete-time Markov chain. Phase-type distributions can be used as a general framework to calculate key properties of the standard coalescent model and many of its extensions. Here, the ‘phases’ in the phase-type distribution correspond to states in the ancestral process. For example, the time to the most recent common ancestor and the total branch length are phase-type distributed. Furthermore, the site frequency spectrum follows a multivariate discrete phase-type distribution and the joint distribution of total branch lengths in the two-locus coalescent-with-recombination model is multivariate phase-type distributed. In general, phase-type distributions provide a powerful mathematical framework for coalescent theory because they are analytically tractable using matrix manipulations. The purpose of this review is to explain the phase-type theory and demonstrate how the theory can be applied to derive basic properties of coalescent models. These properties can then be used to obtain insight into the ancestral process, or they can be applied for statistical inference. In particular, we show the relation between classical first-step analysis of coalescent models and phase-type calculations. We also show how reward transformations in phase-type theory lead to easy calculation of covariances and correlation coefficients between e.g. tree height, tree length, external branch length, and internal branch length. Furthermore, we discuss how these quantities can be used for statistical inference based on estimating equations. Providing an alternative to previous work based on the Laplace transform, we derive likelihoods for small-size coalescent trees based on phase-type theory. Overall, our main aim is to demonstrate that phase-type distributions provide a convenient general set of tools to understand aspects of coalescent models that are otherwise difficult to derive. Throughout the review, we emphasize the versatility of the phase-type framework, which is also illustrated by our accompanying R-code. All our analyses and figures can be reproduced from code available on GitHub.

1. Introduction

Phase-type distributions have successfully been applied in the actuarial sciences and queueing theory for more than 30 years. The pioneering work of A. K. Erlang at a Copenhagen telephone company, where he developed the method of *stages* for the duration of calls, is among the first systematic applications of phases in stochastic modeling. A fascinating recount of these developments can be found in Brockmeyer et al. (1948). Later, Jensen (1953) formalized the notion of stages, and defined a (univariate) phase-type distribution as it is known today. It would take another two decades before Neuts (1975) and co-workers provided a systematic development of the theory, with main examples and contributions in queueing theory. We also note that John

Kingman followed and contributed to queueing theory throughout his entire scientific career (Kingman, 2009) and that his knowledge about queueing theory was essential for developing coalescent theory (see the comments by Warren Ewens in the supplement of Rosenberg (2020)).

In the area of insurance risk, phase-type distributions have found a number of important applications (see e.g. Asmussen and Albrecher (2010), Bladt and Nielsen (2017), and references therein). The popularity of phase-type distributions in the aforementioned areas is mainly due to their tractability and generality: they allow for explicit expressions of distributions and summary statistics even in complex stochastic models having phase-type components. Furthermore, any non-negative distribution can be approximated arbitrarily closely by a phase-type

* Corresponding author.

E-mail addresses: asger@math.au.dk (A. Hobolth), irg@birc.au.dk (I. Rivas-González), bladt@math.ku.dk (M. Bladt), andreas.futschik@jku.at (A. Futschik).

distribution. Additionally, reward structures in the context of phase-type distributions were introduced by Kulkarni (1989), which allowed for the definition of multivariate phase-type distributions.

Much of the theory and methods for phase-type distributions have been developed for queuing and risk theory, but phase-type distributions are increasingly being applied in other disciplines (Hurtado and Richards, 2021). In particular, phase-type distributions are an emerging framework in mathematical population genetics, but it is still in its infancy and not yet fully explored. While most applications in queuing or risk theory employ phase-type distributions out of convenience, with phases being unobserved and without a physical interpretation, the situation in population genetics is somewhat unique in the sense that many models can be identified directly as being of phase-type form.

In queueing theory and insurance risk, the typical data are inter-arrival times, service times, and claim sizes. In general, there is no reason for these quantities to be phase-type distributed, but the assumption allows for expressing phase-type important properties like waiting-time distributions, queue lengths, or ruin probabilities by explicit formulas. From a statistical point of view, we are most often in the situation of incomplete data, since data are only times until absorption (modeling e.g. claim sizes) and not the whole underlying history of a Markov process. Estimation can be achieved by invoking an EM-algorithm (Asmussen et al., 1996) or a Markov chain Monte Carlo approach (Bladt et al., 2003). In population genetics, the data is even more incomplete: we only observe the present-day DNA sequences, and information about the underlying evolutionary tree is missing.

The usage of stochastic simulation is increasingly popular, easy, and efficient for statistical inference and model selection in population genetics (e.g. Baumdicker et al., 2022; Freund and Siri-Jégousse, 2021 and Schrider and Kern, 2018). An advantage of the simulation-based likelihood-free approaches is that they are very flexible and general; the analyst ‘just’ needs to be able to simulate from the model. A disadvantage is the (often high) computational cost and lack of optimality for the estimation procedure (see e.g. Chapter 6 in Bijma et al. (2017) for an introduction to optimality theory in mathematical statistics and the motivation for using maximum likelihood estimators).

In this review, we show that the class of population genetic models that allows a detailed mathematical treatment is larger than one might perhaps expect, and we point to references for extending the class of analytically tractable models even further. We have developed PhaseTypeR, a software package in R that facilitates straightforward implementation and simulation of coalescent models. In particular, all analyses in this paper can easily be reproduced and are available on GitHub. The package is available on CRAN and is described in Rivas-González et al. (2023). Software for fitting both time-homogeneous and time-inhomogeneous phase-type distributions with unidentified phases, as in Asmussen et al. (1996) and Albrecher et al. (2022), is available through the R package matrixdist (Bladt and Yslas, 2021).

Applications of phase-type methods to population genetic data is still in its infancy. Statistical inference procedures and analysis of large sample sizes based on phase-type theory remains a challenge. In this review we also discuss a number of limitations of the current phase-type framework. The limitations include a fast increase in the size of the state space when the sample size increases, and lack of tractable expressions for reward transformations for inhomogeneous coalescent models.

The review is organized as follows: Sections 2 and 3 constitute a basic introduction to the advantages of using the phase-type framework in population genetics. In Section 2, we consider univariate statistics such as tree height, total tree length and external branch length, and show that they are phase-type distributed. In Section 3, we consider the joint distribution of phase-type distributed variables. Section 4 is concerned with the discrete phase-type distribution, and we show that the number of segregating sites follows such a distribution. The time to fixation in the Wright–Fisher model is also discrete phase-type distributed. In Section 5, we establish the connection between likelihood inference

for coalescent models via the Laplace transform originally suggested by Lohse et al. (2011) and the phase-type framework. In particular, the Laplace transform is analytically tractable for multivariate phase-type distributed variables, and this feature facilitates likelihood inference. Finally, in Section 6, we discuss further applications, extensions, limitations, and perspectives on the use of phase-type theory in population genetics. All our analyses and figures are available at https://github.com/rivasiker/phasetype_review.

2. Phase-type distribution with a view towards coalescent theory

In this section, we describe why phase-type distributions are useful in population genetics. We first provide simple examples of key quantities in population genetics that are phase-type distributed. A phase-type distribution is the time to absorption in a continuous-time Markov chain and examples from population genetics include the tree height, total tree length, and external branch length in a coalescent tree. We then describe the general theory and most important formulae from continuous phase-type distributions and relate them to the classical first-step analysis of continuous-time Markov chains. Finally, we introduce reward-transformed phase-type distributions. Reward-transformed phase-type distributions allow us to weigh the time spent in a state before absorption. For example, the total tree length is the weighted tree height where the weights correspond to the number of lineages in the different states, and the external tree length is the weighted tree height where the weights correspond to the number of lineages that are ancestral to exactly one leaf (see Fig. 1). The joint distribution of reward-transformed phase-type distributions is multivariate phase-type. This is the topic for Section 3 and allows us to jointly consider e.g. total tree length, external tree length, and internal tree length.

2.1. Tree height, total tree length, and external branch length: Classical approach

In the standard (Kingman) coalescent model, the time it takes for two sequences to coalesce is exponentially distributed with rate 1, i.e., $H_2 \sim \exp(1)$. Generalizing this, the time H_i between coalescent events from i sequences to $(i-1)$ sequences is exponentially distributed, i.e. $H_i \sim \exp(\lambda_i)$, with rate $\lambda_i = \binom{i}{2} = \frac{i(i-1)}{2}$. Since these exponential distributions are independent, we can calculate the tree height H for a given number of starting sequences $n \geq 2$ as $H = \sum_{i=2}^n H_i$. In other words, the time H until the most recent ancestor of n sequences is the sum of independent exponential distributions (Fig. 1).

The mean and variance of H are well-known, and so is the probability density function $f_H(s)$. The calculation of $f_H(s)$ requires the use of convolutions of exponential distributions. In the simple case when $n = 3$, we get the convolution of two exponential distributions, namely $H_2 \sim \exp(\lambda_2)$ and $H_3 \sim \exp(\lambda_3)$. Since $\lambda_2 \neq \lambda_3$ we get the probability density function

$$f_H(s) = f_{H_2+H_3}(s) = \frac{\lambda_3}{(\lambda_3 - \lambda_2)} \lambda_2 e^{-\lambda_2 s} + \frac{\lambda_2}{(\lambda_2 - \lambda_3)} \lambda_3 e^{-\lambda_3 s}, \quad s \geq 0, \quad (1)$$

(Wakeley, 2008, equation 2.63). By substituting $\lambda_2 = 1$ and $\lambda_3 = 3$, we get the density function of H when $n = 3$. For $n > 3$, $f_H(s)$ requires a longer series of convolutions of exponential distributions, and then substituting the corresponding exponential rates.

Many other quantities in population genomics are also sums of exponential distributions like, e.g., the total tree length $L = \sum_{i=2}^n L_i = \sum_{i=2}^n i H_i$, which is similar to H but weighted by the number of ancestral branches in each state of the coalescent process. Recall that positive scalar multiples of exponential distributions remain exponential: if X is exponential with rate λ and $a > 0$, then aX is exponential with rate λ/a . We can then apply the same approach as described above to get $f_L(s)$, by using that $L_i \sim \exp(\lambda_i)$, where now

$$\lambda_i = \frac{i(i-1)}{2} \cdot \frac{1}{i} = \frac{i-1}{2}. \quad (2)$$

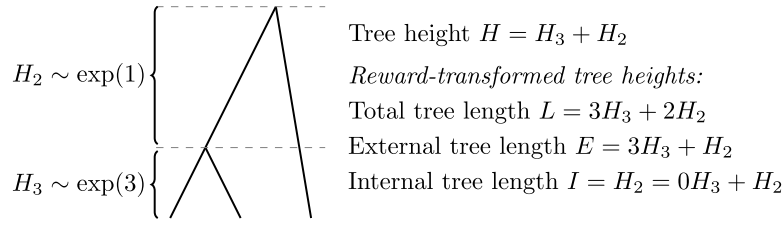


Fig. 1. Coalescent tree for a sample of size $n = 3$. The basic tree height $H = H_2 + H_3$ as well as any non-negative reward-transformed tree length $aH_3 + bH_2$, $a \geq 0$, $b \geq 0$, are phase-type distributed.

Now consider the singleton branch length or – equivalently – the external branch length E . In the case $n = 3$ we have $E = 3H_3 + H_2$, and since $3H_3 \sim \exp(1)$ and $H_2 \sim \exp(1)$ we now have to consider the sum of two exponential distributions with the same rate. The distribution is a so-called Erlang distribution, and we have that the sum of two exponential distributions with rate λ is

$$f_E(s) = f_{3H_3+H_2}(s) = \lambda^2 s e^{-\lambda s}, \quad s \geq 0. \quad (3)$$

By substituting $\lambda = 1$ we get the density for the singleton branch length when $n = 3$.

The above approach, while feasible, relies on daunting analytical formulations. Moreover, the convolution of exponential distributions as described above cannot be used in general for non-standard coalescent models, in which, unlike the Kingman coalescent, trees are not necessarily bifurcating, and the number of branches might be reduced by two or more after a coalescent event. This happens, for example, in multiple merger coalescent models such as the Λ -coalescent (Pitman, 1999; Sagitov, 1999), the beta coalescent (Schweinsberg, 2003), the psi coalescent (Eldon and Wakeley, 2006), and the seed-bank coalescent (Lambert and Ma, 2015; Blath et al., 2016), where $n \geq 2$ branches might coalesce to $n-1, n-2, \dots, 1$ branches. Additionally, when performing convolutions of exponential distributions, it is important to distinguish between the case when the rates are the same (as in Eq. (3)) and the case when they are different (as in Eq. (1)). Fortunately, the desired convolutions can easily be computed by formulating the coalescent as an absorbing Markov jump process and using phase-type distributions.

2.2. Properties of the coalescent tree: Phase-type distributions

Consider the standard coalescent when $n = 3$ as a Markov jump process $\{X_t\}_{t \geq 0}$, with $p = 2$ transient states and a single absorbing state. The transient states correspond to the stages in the coalescent tree where 3 or 2 lineages are present, while the absorbing state corresponds to the state when all of the sequences have coalesced. The coalescent can then be summarized using a transition rate matrix A , such that

$$A = \begin{pmatrix} T & t \\ \mathbf{0} & 0 \end{pmatrix}, \quad (4)$$

where

$$T = \begin{pmatrix} -\lambda_3 & \lambda_3 \\ 0 & -\lambda_2 \end{pmatrix} \text{ and } t = \begin{pmatrix} 0 \\ \lambda_2 \end{pmatrix}. \quad (5)$$

For H we have $(\lambda_3, \lambda_2) = (3, 1)$, for L we have $(\lambda_3, \lambda_2) = (1, 1/2)$, and for E we have $(\lambda_3, \lambda_2) = (1, 1)$.

In general, a phase-type distributed random variable is defined as the time to absorption in a continuous-time Markov chain with an absorbing state. Let the first p states of the Markov chain be the transient states and state $p+1$ be the absorbing state. We can then write the transition rate matrix for the Markov chain as in (4) where T is now a sub-intensity matrix of size $p \times p$ which holds the transition rates between the transient states, t is a column vector of size p with the exit rates from the transient states into the absorbing state, and $\mathbf{0}$ is a row vector of zeros of size p . By convention, the rows of A sum

to zero, i.e. $Ae = \mathbf{0}$ with the column vector $e = (1, \dots, 1)'$, where the superscript denotes transposition. Therefore, the transition rate matrix for the absorbing Markov jump process is fully determined by T , since the exit rate vector is given by $t = -Te$. We also need to provide an initial state distribution α of starting in each of the p transient states. We write $\tau \sim \text{PH}(\alpha, T)$ for a phase-type distributed random variable with parameters (α, T) , and the probability density function is given by

$$f_\tau(s) = \alpha e^{Ts} t = \sum_{i=0}^{\infty} \frac{s^i}{i!} \alpha T^i t, \quad s \geq 0. \quad (6)$$

A main advantage of continuous phase-type distributions (PHs) is that they have well-defined formulas in matrix notation for calculating their basic properties. In (6) we provided the probability density function. Similarly, the mean, variance, moments, and cumulative distribution function are also available in analytically tractable expressions. The mathematical expressions are general, defined solely by α and T , and available in our R package PhaseTypeR.

In those cases where phase-type distributions correspond to convolutions of exponential distributions, the phase-type formulas using matrix notation naturally yield the same result. We consider the two examples of tree height and external branch length to demonstrate this agreement.

2.2.1. Example: Tree height distribution

We return to $f_H(s)$ when $n = 3$ for an illustrative example. The tree height H is the time until absorption, $H = \inf\{t \geq 0 : X_t = p+1\}$, of a continuous-time Markov jump process $\{X_t\}_{t \geq 0}$ with $p = 2$ transient states. The initial distribution is given by the row vector $\alpha = (1, 0)$ and the $p \times p = 2 \times 2$ sub-intensity matrix

$$T = \begin{pmatrix} -\lambda_3 & \lambda_3 \\ 0 & -\lambda_2 \end{pmatrix} \quad (7)$$

where $\lambda_3 = 3$ and $\lambda_2 = 1$. When the tree height H is described as a phase-type distribution we write $H \sim \text{PH}(\alpha, T)$. In other words, H is the time until absorption of the sum of two sequentially occurring exponential distributions defined by a probability vector α of starting in a certain state, and a sub-intensity matrix T with the transition rates between the $p = 2$ transient states given by (7) with $(\lambda_3, \lambda_2) = (3, 1)$.

We now calculate the density function (6) for $\alpha = (1, 0)$, T given by (7) with $\lambda_2 \neq \lambda_3$, and $t = -Te = (0, \lambda_2)'$. The first term in (6) with $i = 0$ is zero because $\alpha T^0 t = \alpha t = 0$. The remaining terms with $i \geq 1$ are determined by

$$\begin{aligned} \alpha T^i t &= (-1)^{i+1} \sum_{k=1}^i \lambda_3^k \lambda_2^{i-k+1} = \frac{1}{(\lambda_3 - \lambda_2)} (-1)^{i+1} (\lambda_2 \lambda_3^{i+1} - \lambda_3 \lambda_2^{i+1}) \\ &= \frac{\lambda_3 \lambda_2}{(\lambda_3 - \lambda_2)} (-1)^{i+1} (\lambda_3^i - \lambda_2^i), \end{aligned} \quad (8)$$

where in order to obtain the second equal sign we used that we have an alternating sum when multiplying by $(\lambda_3 - \lambda_2)$. In the end, we get

$$\begin{aligned} f_H(s) &= \frac{\lambda_3 \lambda_2}{(\lambda_3 - \lambda_2)} \sum_{i=1}^{\infty} \frac{s^i}{i!} (-1)^{i+1} (\lambda_3^i - \lambda_2^i) = \frac{\lambda_2}{(\lambda_2 - \lambda_3)} \lambda_3 e^{-\lambda_3 s} \\ &\quad + \frac{\lambda_3}{(\lambda_3 - \lambda_2)} \lambda_2 e^{-\lambda_2 s}, \quad s \geq 0, \end{aligned} \quad (9)$$

which is the same result as in (1).

2.2.2. Example: External and total length distribution

For the external length distribution E we have a phase-type representation given by initial distribution $\alpha = (1, 0)$ and sub-intensity matrix T given by (7) with $\lambda_3 = \lambda_2 = 1$. In the following we let $\lambda_3 = \lambda_2 = \lambda$ to establish the connection with (3). Note that $t = -Te = (0, \lambda)'$. Let us again consider the density function (6). The first term with $i = 0$ is zero because $\alpha T^0 t = \alpha I t = \alpha t = 0$. The remaining terms with $i \geq 1$ are determined by

$$\alpha T^i t = (-1)^{i+1} i \lambda^{i+1}, \quad (10)$$

and we get

$$f_E(s) = \sum_{i=1}^{\infty} \frac{s^i}{i!} (-1)^{i+1} i \lambda^{i+1} = \lambda^2 s \sum_{i=1}^{\infty} (-1)^{i-1} \frac{s^{i-1}}{(i-1)!} \lambda^{i-1} = \lambda^2 s e^{-\lambda s}, \quad s \geq 0, \quad (11)$$

which is the same result as in (3).

For an arbitrary number of leaves n , the tree height H is phase-type distributed with initial distribution $\alpha = (1, 0, \dots, 0)$ and sub-intensity matrix

$$T = \begin{pmatrix} -\lambda_n & \lambda_n & 0 & \dots & 0 \\ 0 & -\lambda_{n-1} & \lambda_{n-1} & \dots & 0 \\ 0 & 0 & -\lambda_{n-2} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & -\lambda_2 \end{pmatrix}, \quad (12)$$

where $\lambda_i = \binom{i}{2}$. The sub-intensity matrix for the total tree length L has the same structure, but with $\lambda_i = (i-1)/2$ (recall Eq. (2)). Phase-type distributions with rate matrices given by (12) are called generalized Erlang distributions (Bladt and Nielsen, 2017), and if the starting state is $\alpha = (1, 0, \dots, 0)$ then this distribution corresponds to the distribution of the sum of exponential random variables with rates $\lambda_n, \lambda_{n-1}, \dots, \lambda_2$. In the special case when all the parameters λ_i are distinct, i.e. $\lambda_i \neq \lambda_j$ for all $i \neq j$, we have the closed form solution for the density

$$f(s) = \sum_{i=2}^n \lambda_i e^{-\lambda_i s} \prod_{j=2, j \neq i}^n \frac{\lambda_j}{(\lambda_j - \lambda_i)}$$

(see e.g. equation (2.64) in Wakeley (2008)). This expression is a linear combination of exponential distributions with positive and negative coefficients.

Calculation of the mean, variance, density function, and cumulative distribution function for H, L or E for any sample size $n \geq 2$ is straightforward because the phase-type representation with parameters α and T is available. In particular, the computational methods for calculating matrix exponentials (e.g. Moler and Van Loan (2003)) are stable and reliable; for instance, we have never encountered problems with the expm package in R (Goulet et al., 2021).

2.2.3. Example: Probability of a number of lineages at a fixed time in the past

The distribution of the number of lineages at a past time point t can also be obtained by using phase-type results. Tavaré (2004) (p. 19) derived the distribution of the number of lineages $A_n(t)$ at time t for a coalescent tree that starts with n lineages at time 0. More specifically, he showed that

$$\mathbb{P}(A_n(t) = j) = \sum_{k=j}^n e^{-\lambda_k t} b(n, k, j),$$

where

$$\lambda_k = \binom{k}{2}, \quad b(n, k, j) = \frac{(2k-1)(-1)^{k-j} j_{(k-1)} n_{[k]}}{j!(k-j)! n_{(k)}},$$

$$n_{(k)} = n(n+1) \cdots (n+k-1), \quad n_{[k]} = n(n-1) \cdots (n-k-1),$$

and $n_{[0]} = n_{(0)} = 1$.

Subsequently, Blum and Rosenberg (2007) used this distribution to obtain the distribution of intercoalescence times conditional on $A_n(t)$.

We briefly explain how the distribution of $A_n(t)$ can be obtained using a phase type approach. Indeed, let $H^{(l)} = \sum_{i=l+1}^n H_i$ be the time until the coalescent reaches a state with l lineages. Obviously, the event $A_n(t) \leq l$ is equivalent to the event $H^{(l)} \leq t$ which can be computed by choosing “ l lineages” as the absorbing state. Thus, the computation of the distribution of $H^{(l)}$ is analogous to the computation for the tree height in the previous example, but with the reduced sub-intensity matrix T taken as the $\{1, \dots, (n-l)\} \times \{1, \dots, (n-l)\}$ sub-matrix of (12).

2.3. Mean and variance for phase-type distributions from first-step analysis

We now consider the mean, variance, and higher-order moments for a phase-type distribution with the sub-intensity matrix T , where $T_{ij} = t_{ij}$. As described above, a phase-type distribution corresponds to the first passage time into the absorbing state, and the traditional procedure for determining the mean and variance uses a first-step analysis (e.g. Section 5.4 in Ibe (2013)). An alternative procedure for deriving the higher-order moments is to differentiate the Laplace transform and evaluate at zero. We describe both procedures in this subsection.

Let M_i be the expected time to absorption given the initial state is i . For any transient state i , we have (see e.g. Section 5.4 in Ibe (2013) or Section 5.1.3 in Wakeley (2008))

$$M_i = \frac{1}{-t_{ii}} + \sum_{k \neq i} \frac{t_{ik}}{(-t_{ii})} M_k, \quad (13)$$

because the time before leaving state i is exponential with rate $-t_{ii}$, and the probability of jumping from state i to another transient state k is $t_{ik}/(-t_{ii})$. We can rearrange the equation to get

$$-t_{ii} M_i - \sum_{k \neq i} t_{ik} M_k = 1,$$

or

$$-\sum_{k=1}^p t_{ik} M_k = 1. \quad (14)$$

Let $\mathbf{M} = (M_1, \dots, M_p)'$ be the vector of means. In matrix format, we write (14) as

$$-T\mathbf{M} = \mathbf{e},$$

and we get $\mathbf{M} = (-T)^{-1}\mathbf{e}$. We call $\mathbf{U} = (-T)^{-1}$ the Green matrix (Bladt and Nielsen, 2017). For $\tau \sim \text{PH}(\alpha, T)$, the initial distribution is α , and we get the mean

$$\mathbb{E}[\tau] = \alpha \mathbf{M} = \alpha (-T)^{-1} \mathbf{e}. \quad (15)$$

Now let us derive an expression for the variance. Let

$$Q_i = \int_0^\tau s^2 f_i(s) ds$$

be the second moment for the time to absorption given initial state i . Wakeley (2008, Section 5.1.3, page 145) derives the equation

$$Q_i = \mathbb{E}[\tau_i^2] + 2\mathbb{E}[\tau_i] \sum_{k \neq i} \frac{t_{ik}}{(-t_{ii})} M_k + \sum_{k \neq i} \frac{t_{ik}}{(-t_{ii})} Q_k.$$

Here, τ_i is the exponentially distributed time before the jump from state i (with rate $-t_{ii}$), and we have $\mathbb{E}[\tau_i] = 1/(-t_{ii})$ and $\mathbb{E}[\tau_i^2] = 2/(-t_{ii})^2$. Substituting, we get

$$Q_i = \frac{2}{(-t_{ii})^2} + 2 \frac{1}{(-t_{ii})} \sum_{k \neq i} \frac{t_{ik}}{(-t_{ii})} M_k + \sum_{k \neq i} \frac{t_{ik}}{(-t_{ii})} Q_k,$$

and re-arranging we obtain

$$t_{ii} Q_i + \sum_{k \neq i} t_{ik} Q_k = \frac{2}{t_{ii}} + 2 \sum_{k \neq i} \frac{t_{ik}}{t_{ii}} M_k.$$

Using Eq. (13), this formula amounts to

$$\sum_{k=1}^p t_{ik} Q_k = 2 \left(\frac{1}{t_{ii}} + \sum_{k \neq i} \frac{t_{ik}}{t_{ii}} M_k \right) = -2M_i.$$

Let $\mathbf{Q} = (Q_1, \dots, Q_p)'$ be the column vector with entries Q_i , $i = 1, \dots, p$. In matrix format, we have

$$\mathbf{T}\mathbf{Q} = -2\mathbf{M} = 2\mathbf{T}^{-1}\mathbf{e},$$

or

$$\mathbf{Q} = 2\mathbf{T}^{-2}\mathbf{e}.$$

We conclude that $\mathbb{E}[\tau^2] = 2\alpha(-\mathbf{T})^{-2}\mathbf{e}$

In general, the higher-order moments for a $\text{PH}(\alpha, \mathbf{T})$ distributed random variable are given by

$$\mathbb{E}[\tau^n] = n!\alpha(-\mathbf{T})^{-n}\mathbf{e}, \quad (16)$$

as stated in, e.g., Corollary 3.1.18 in [Bladt and Nielsen \(2017\)](#). This formula is general, has a simple implementation, and avoids calculating the moments using first-step analysis. Perhaps the easiest derivation of the general formula is to first calculate the Laplace transform

$$\mathcal{L}_\tau(u) = \int_0^\infty e^{-us} \alpha e^{\mathbf{T}s} ds = \alpha \int_0^\infty e^{-(u\mathbf{I} - \mathbf{T})s} ds = \alpha(u\mathbf{I} - \mathbf{T})^{-1}\mathbf{e}, \quad u \geq 0.$$

Second, the higher-order moments for the $\text{PH}(\alpha, \mathbf{T})$ distribution can be derived by differentiating the Laplace transform with respect to u an appropriate number of times, and evaluate at zero.

We end this subsection with a brief discussion of the evaluation of the matrix formulas in the spirit of [Røikjer et al. \(2022\)](#). Suppose we want to evaluate the first moment, i.e. $\alpha(-\mathbf{T})^{-1}\mathbf{e}$. If we let $\mathbf{z} = (-\mathbf{T})^{-1}\mathbf{e}$ then we need to solve the linear equation system $-\mathbf{T}\mathbf{z} = \mathbf{e}$, and multiply the solution \mathbf{z} by α from the left. In population genetics, the rate matrices are often upper triangular, and in this case the linear equation system can be solved using the back solve algorithm (e.g. [Arnold et al. \(2019\)](#), Section 2.5). To illustrate how the back solve algorithm works we consider the total tree height with sample size $n = 4$. In this case the linear equation system $-\mathbf{T}\mathbf{z} = \mathbf{e}$ becomes

$$-\begin{pmatrix} -6 & 6 & 0 \\ 0 & -3 & 3 \\ 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

Solving backwards we get $z_3 = 1$, $3z_2 - 3z_3 = 1$ or $z_2 = 1 + 1/3$, and finally $6z_1 - 6z_2 = 1$ or $z_1 = 1 + 1/3 + 1/6$. We have $\alpha = (1, 0, 0)$, and the mean is therefore $z_1 = 1 + 1/3 + 1/6$. The back solve algorithm for solving a linear equation system where the coefficient matrix is upper triangular is implemented in R in the `backsolve` command. Furthermore, the back solve methodology also works if the rate matrix is an upper block triangular matrix instead of a strictly upper triangular matrix; see [Røikjer et al. \(2022\)](#) for more information.

Now consider the problem of evaluating the second moment $2\alpha(-\mathbf{T})^{-2}\mathbf{e}$. We first find $\mathbf{z}_1 = (-\mathbf{T})^{-1}\mathbf{e}$ as the solution to the linear equation system $-\mathbf{T}\mathbf{z}_1 = \mathbf{e}$ using the back solve algorithm as described above. Second we solve the linear equation system $-\mathbf{T}\mathbf{z}_2 = \mathbf{z}_1$ with respect to \mathbf{z}_2 (again using the back solve algorithm), and finally we compute the second moment from $2\alpha(-\mathbf{T})^{-2}\mathbf{e} = 2\alpha(-\mathbf{T})^{-1}\mathbf{z}_1 = 2\alpha\mathbf{z}_2$.

As a final remark we note that the rate matrices in population genetics are often sparse and large. Using R, the matrices could be stored in the compressed sparse row (CSR) format for efficient evaluation. Efficient implementation of the phase-type matrix formulas is an important future research area.

2.4. Reward-transformed phase-type distributions

Recall from [Fig. 1](#) that quantities such as the total tree length, external tree length, or internal tree length are weighted versions of the time spent in the states of the ancestral process. In this subsection, we consider these so-called reward-transformed variables. The class of phase-type distributions is closed under reward-transformations, meaning that a reward-transformed phase-type distribution is, again, phase-type distributed. This is a huge advantage in order to obtain

the probability distribution and moments of the reward-transformed variable.

Let $\tau \sim \text{PH}(\alpha, \mathbf{T})$ and let $\{X_i\}_{i \geq 0}$ be the underlying Markov jump process. We then define a vector $\mathbf{r} = (r(1), r(2), \dots, r(p))$ consisting of non-negative rewards, so that a new reward-transformed random variable $\tilde{\tau}$ is given by

$$\tilde{\tau} = \int_0^\tau r(X_t) dt. \quad (17)$$

If $r(i) = 1$ for all i , then $\tilde{\tau} = \tau$. If $r(i) = 1$ and $r(j) = 0$ for $j \neq i$, then $\tilde{\tau}$ is the time spent in state i . If instead any of the entries in \mathbf{r} is a non-negative number different from 1, then $\tilde{\tau}$ corresponds to the total reward accumulated until absorption.

An alternative but equivalent definition of the reward-transformed variable is

$$\tilde{\tau} = \sum_{i=1}^p r(i)Z_i,$$

where Z_i is the total time spent in state i before absorption.

If all of the values in \mathbf{r} are positive, then it can be shown that $\tilde{\tau} \sim \text{PH}(\alpha, \tilde{\mathbf{T}})$, where $\tilde{\mathbf{T}} = \Delta(\mathbf{r})^{-1}\mathbf{T}$ ([Bladt and Nielsen, 2017](#)). Here, $\Delta(\mathbf{r})$ denotes the diagonal matrix of size $p \times p$ with \mathbf{r} in the diagonal. Alternatively, if one or more of the rewards in \mathbf{r} are equal to 0, then $\tilde{\tau}$ will also follow a PH, but the number of states in the new PH is p minus the number of zero-reward states. This reflects that even though transitions through a zero-reward state are possible, they should not be accounted for (they can be removed) in the distribution for $\tilde{\tau}$. Theorem 3.1.33 in [Bladt and Nielsen \(2017\)](#) provide general formulas for the initial distribution and sub-intensity matrix for a reward-transformed phase-type distribution, and these formulas are implemented in `PhaseTypeR`.

The ability to reward transform is extremely useful in population genetics, since, as explained above, many quantities are, in fact, weighted sums of exponential distributions. This is the case for the total tree length L , which is a version of the tree height H weighted by the number of branches in each time interval between coalescent events. Since $H \sim \text{PH}(\alpha_H, \mathbf{T}_H)$, as explained in Sections 2.1 and 2.2, we can define a reward vector $\mathbf{r}_L = (n, n-1, \dots, 2)$, so that $L \sim \text{PH}(\alpha_L, \mathbf{T}_L)$, where $\mathbf{T}_L = \Delta(\mathbf{r}_L)^{-1}\mathbf{T}_H$ and $\alpha_L = \alpha_H = \mathbf{e}_1$. The sub-intensity matrix \mathbf{T}_L will be the same as when calculated using Eq. (12).

Many other quantities in population genomics are linear combinations of the times spent in a state before absorption. This is the case for the total branch length leading to each of the elements of the site frequency spectrum (singletons, doubletons, etc.), the total length of the external branches, and the total length of the internal branches. Consider for example the total branch length B_1 leading to singletons when $n = 4$. A naive first step would be to represent the total tree height H using phase-type theory. We could use Eq. (12) to calculate the sub-intensity matrix

$$\mathbf{T}_H = \begin{pmatrix} -6 & 6 & 0 \\ 0 & -3 & 3 \\ 0 & 0 & -1 \end{pmatrix}, \quad (18)$$

so $H \sim \text{PH}(\alpha_H, \mathbf{T}_H)$, where $\alpha_H = (1, 0, 0)$. However, we cannot directly reward transform \mathbf{T}_H to get a PH-representation of the total singleton branch length. The reason for this is that we are now interested in modeling a specific type of branch, so we need to consider all possible trees for the coalescent process. For $n = 4$, there are two types of trees, depicted in [Fig. 2A](#). Both trees start with a coalescence between two singleton branches forming a doubleton branch. From the remaining sequences, any of the two singleton branches will coalesce with the doubleton branch to form a tripleton branch with probability 2/3 (left tree), or the two singleton branches will coalesce with probability 1/3 to form a second doubleton branch (right tree). Finally, in both trees, the two remaining branches will find common ancestry and enter the absorbing state in the next coalescent.

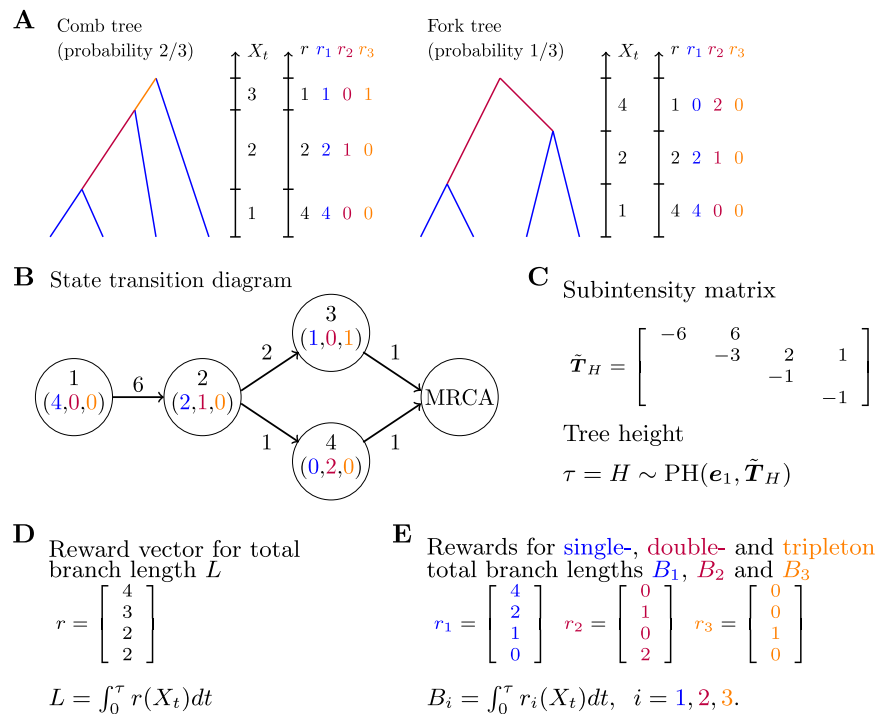


Fig. 2. **A.** The two possible trees with a sample size of $n = 4$. The coloring of the branches is according to the number of descendant leaves. **B.** The states and state transition diagram. **C.** The corresponding sub-intensity matrix and phase-type distribution for the tree height H . **D.** Rewards for total branch length L . **E.** Rewards for total singleton branch length B_1 , total doubleton branch length B_2 , and total tripleton branch length B_3 .

Following this tree pattern, we can re-formulate the PH representation of the tree height H by splitting the state in T_H with two branches into a state with one singleton branch and one tripton branch and a state with two doubleton branches

$$\tilde{T}_H = \begin{pmatrix} -6 & 6 & 0 & 0 \\ 0 & -3 & 2 & 1 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}. \quad (19)$$

See Fig. 2B for the state transition diagram and Fig. 2C for the new phase-type formulation of the tree height distribution. Thus, the tree height can also be modeled as $H \sim \text{PH}(\tilde{\alpha}_H, \tilde{T}_H)$, where $\tilde{\alpha}_H = (1, 0, 0, 0)$. This is an equivalent representation of H when using T_H and α_H , but with the added advantage of modeling all branch types explicitly.

We can now define reward vectors to obtain the distribution of total branch length L (Fig. 2D) and total singleton, doubleton and triplet branch lengths (Fig. 2E). The reward vector $\mathbf{r}_1 = (4, 2, 1, 0)$ corresponds to the number of singleton branches in each of the states. After reward transforming \tilde{T}_H with \mathbf{r}_1 to obtain T_1 , we get that $B_1 \sim \text{PH}(\alpha_1, T_1)$, where $\alpha_1 = (1, 0, 0)$. Note that the new sub-intensity matrix T_1 is of size 3 since state 4 in \tilde{T}_H is discarded due to it having a reward of 0 (no singleton branches). In a similar manner, we can get PH representations of the branch length leading to doubletons (B_2) and that leading to tripletons (B_3) by reward transforming \tilde{T}_H by $\mathbf{r}_2 = (0, 1, 0, 2)$ and $\mathbf{r}_3 = (0, 0, 1, 0)$, respectively. Reward transformation is implemented in PhaseTypeR, and for $n = 4$ we get (see the accompanying R code)

$$B_i \sim \text{PH}(\alpha_i, T_i), \quad i = 1, 2, 3,$$

with

$$\text{Singletons : } \alpha_1 = (1, 0, 0) \text{ and } T_1 = \begin{pmatrix} -3/2 & 3/2 & 0 \\ 0 & -3/2 & 1 \\ 0 & 0 & -1 \end{pmatrix}$$

Doubletons : $\alpha_2 = (1, 0)$ and $T_2 = \begin{pmatrix} -3 & 1 \\ 0 & -1/2 \end{pmatrix}$

Triplettons : $\alpha_3 = (2/3)$ and $T_3 = (-1)$.

The fork tree does not contain any tripleton branches and occurs with probability $1/3$. This is reflected in the initial distribution α_3 that only adds to $2/3$. We say that the tripleton branch length distribution is defective; with probability $1/3$ the tree does not contain tripleton branches.

For any sample size n , the reward vectors for all of these quantities can be automatically computed by careful consideration of the block counting process of the coalescent as described in Algorithm 4.2 in [Hobolth et al. \(2019\)](#) and Example 1 in [Rivas-González et al. \(2023\)](#). Recall from [Fig. 2B](#) that if the number of samples is 4, then the state space is of size 4, and the states are (4, 0, 0), (2, 1, 0), (1, 0, 1) and (0, 2, 0), where the first entry in each triplet is the number of singleton branches, the second entry is the number of doubleton branches, and the last entry is the number of tripton branches. We now note that $4 = 1 + 1 + 1 + 1 = 1 + 1 + 2 = 1 + 3 = 2 + 2$, and these are exactly the five possible ways of partitioning the number 4 into a sum of positive integers. In general, the size of the state space for the block counting process is given by the so-called partition function from number theory.

The partition function is sequence A00041 in the Online Encyclopedia Of Integer Sequences (OEIS) and can be found at <https://oeis.org/A000041>. In the second row in Table 1 we provide a selected number of values for the partition function. For a sample size n larger than 45, the size of the state space for the block counting process increases to more than 100,000 states. Fortunately, all hope is not lost. In Table 1 we show different measures of sparsity for the block counting process. In the third row we show the number of non-zero entries for the sub-intensity matrix. For example, the sub-intensity matrix for sample size $n = 4$ in Fig. 2C has seven non-zero entries. The number of out-going edges in Fig. 2B is five and the number of nodes is the partition number $p(4) = 5$, so the average number of out-going edges is 1. Finally, the sparsity is calculated as one subtracted by the fraction between the number of non-zero entries and the size of the subintensity matrix $(p(n) - 1)^2$. For $n = 4$, this amounts to $\{1 - [7/(5 - 1)^2]\} = 0.5625$. It is encouraging that the amount of sparsity grows with the sample size. For a sample size of $n = 50$ the average number of out-going edges for each node is smaller than 15 despite a total number of 204,226 nodes.

Table 1

Partition function $p(n)$ (second row), number of non-zero entries in the subintensity matrix of dimension $(p(n) - 1) \times (p(n) - 1)$ (third row), average number of out-going edges (fourth row), and sparsity for the block counting process (fifth row), with varying sample size n (first row). The average number of out-going edges is the accumulated total number of out-going edges for each node (number of states that have a non-zero rate for each state) divided by the total number of states (including the MRCA). See the main text for examples.

Sample size n	4	5	10	20	30	40	50
Partition function $p(n)$	5	7	42	627	5,604	37,338	204,226
Non-zero entries	7	13	150	4,161	53,885	470,254	3,181,424
Average out-going edges	1.00	1.29	2.71	5.65	8.62	11.60	14.58
Sparsity	0.56250	0.63889	0.91077	0.98938	0.99828	0.99966	0.99992

2.5. Moments for reward-transformed phase-type distributions

We now consider a first-step analysis for the mean reward before absorption. The reward transformation is given by $\mathbf{r} = (r(1), \dots, r(p)) = (r_1, \dots, r_p)$ and the accumulated reward is $Y = \int_0^\tau r(X_t)dt$. Let

$$A_i = \mathbb{E} \left[\int_0^\tau r(X_t)dt \mid X_0 = i \right]$$

be the mean accumulated reward given the initial state is i , and let τ_1 be the time until the first jump. We always jump from i to some other state $k \neq i$ and therefore

$$\begin{aligned} \int_0^\tau r(X_t)1(X_0 = i)dt &= \int_0^{\tau_1} r(X_t)1(X_0 = i)dt + \sum_{k \neq i} \int_{\tau_1}^\tau r(X_t)1(X_{\tau_1} = k)dt \\ &= r_i \tau_1 + \sum_{k \neq i} 1(X_{\tau_1} = k) \int_{\tau_1}^\tau r(X_t)dt. \end{aligned} \quad (20)$$

By taking the mean on both sides of the previous equation we get

$$A_i = \frac{r_i}{-t_{ii}} + \sum_{k \neq i} \frac{t_{ik}}{-t_{ii}} A_k. \quad (21)$$

We can rearrange this equation as

$$-t_{ii} A_i - \sum_{k \neq i} t_{ik} A_k = r_i.$$

Letting $\mathbf{A} = (A_1, \dots, A_p)'$ be the column vector of A_i , $i = 1, \dots, p$, we obtain

$$(-T)\mathbf{A} = \mathbf{r} = \Delta(\mathbf{r})\mathbf{e}, \quad \text{or} \quad \mathbf{A} = (-T)^{-1} \Delta(\mathbf{r})\mathbf{e}, \quad (22)$$

where $\Delta(\mathbf{r})$ denotes the diagonal matrix with \mathbf{r} as diagonal. If the initial probability vector is α , we get

$$\mathbb{E}[Y] = \alpha \mathbf{A} = \alpha (-T)^{-1} \Delta(\mathbf{r})\mathbf{e}.$$

In general, it holds that

$$\mathbb{E}[Y^n] = n! \alpha \left((-T)^{-1} \Delta(\mathbf{r}) \right)^n \mathbf{e}. \quad (23)$$

Note that in case of strictly positive rewards $r_i > 0$ for all $i = 1, \dots, p$, this result is consistent with (16) and Y being phase-type distributed with initial distribution α and sub-intensity matrix $\Delta(\mathbf{r})^{-1}T$.

3. Multivariate phase-type distribution and joint branch lengths

A multivariate phase-type distribution is the joint distribution of two or more reward-transformed phase-type distributions. Let $\tau \sim \text{PH}(\alpha, T)$ and let $\{X_t\}_{t \geq 0}$ be the corresponding Markov jump process with p transient states. Consider m non-negative reward functions

$$r_j : \{1, \dots, p\} \rightarrow \mathbb{R}_+, \quad j = 1, \dots, m, \quad (24)$$

and let $\mathbf{R} = \{R_{ij}\}$ be the $p \times m$ matrix with entries $R_{ij} = r_j(i)$. Hence, the j 'th column of \mathbf{R} , $\mathbf{R}_{\cdot j}$, consists of $(r_j(1), \dots, r_j(p))$. Let

$$Y_j = \int_0^\tau r_j(X_t)dt = \int_0^\tau R_{X_t, j} dt, \quad j = 1, \dots, m, \quad (25)$$

be the accumulated reward for reward function $r_j(\cdot)$. Then the random vector $\mathbf{Y} = (Y_1, \dots, Y_m)$ is said to be *multivariate phase-type distributed* with parameters α , T , and \mathbf{R} , and we write $\mathbf{Y} \sim \text{MPH}(\alpha, T, \mathbf{R})$.

Recall the tree height H , total tree length L , external tree length E , and internal tree length I from Fig. 1. The joint distribution of these four stochastic variables is multivariate phase-type distributed

$$(H, L, E, I) \sim \text{MPH}(e_1, T, \mathbf{R})$$

with sub-intensity matrix T and reward-matrix \mathbf{R} given by

$$T = \begin{pmatrix} -3 & 3 \\ 0 & -1 \end{pmatrix} \quad \text{and} \quad \mathbf{R} = \begin{pmatrix} 1 & 3 & 3 & 0 \\ 1 & 2 & 1 & 1 \end{pmatrix}.$$

In Section 3.1, we show how to calculate the covariance between the variables in a multivariate phase-type distribution, and in Section 3.2 we describe how to obtain the covariance between the branch lengths (H, L, E, I) in the standard coalescent model with sample size n .

The joint distribution of the total single-, double- and tripleton branch length (B_1, B_2, B_3) from Section 2.4 is another example of a multivariate phase-type distribution. We have $\mathbf{B} = (B_1, B_2, B_3) \sim \text{MPH}(\alpha_B, T_B, \mathbf{R}_B)$, where $\alpha_B = (1, 0, 0, 0)$, T_B is given in Eq. (19), and the reward matrix \mathbf{R}_B is given in Fig. 2E:

$$T_B = \begin{pmatrix} -6 & 6 & 0 & 0 \\ 0 & -3 & 2 & 1 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix} \quad \text{and} \quad \mathbf{R}_B = \begin{pmatrix} 4 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 2 & 0 \end{pmatrix}. \quad (26)$$

3.1. Covariances in the multivariate phase-type distribution

In this subsection, we use first-step analysis to derive the mixed moment of two reward-transformed variables. Bladt and Nielsen (2017) derive the mixed moment using the Laplace transform of the multivariate phase-type distribution.

Consider the two reward-transformed variables $Y_1 = \int_0^\tau r_1(X_t)dt$ and $Y_2 = \int_0^\tau r_2(X_t)dt$. Let

$$C_i = \mathbb{E}[Y_1 Y_2 \mid X_0 = i] = \mathbb{E} \left[\int_0^\tau r_1(X_t)dt \int_0^\tau r_2(X_t)dt \mid X_0 = i \right]$$

be the mixture moment of the accumulated rewards given that the initial state is i . Let τ_1 be the time until the first jump. We have

$$\begin{aligned} \int_0^\tau r_1(X_t)dt \int_0^\tau r_2(X_t)dt &= r_1(i)r_2(i)\tau_1^2 + r_1(i)\tau_1 \left(\int_0^\tau r_2(X_t)dt - r_2(i)\tau_1 \right) + \\ &\quad \left(\int_0^\tau r_1(X_t)dt - r_1(i)\tau_1 \right) r_2(i)\tau_1 + \\ &\quad \left(\int_0^\tau r_1(X_t)dt - r_1(i)\tau_1 \right) \left(\int_0^\tau r_2(X_t)dt - r_2(i)\tau_1 \right), \end{aligned}$$

and we get

$$\begin{aligned} C_i &= r_1(i)r_2(i) \frac{2}{(-t_{ii})^2} + r_1(i) \frac{1}{-t_{ii}} \sum_{k \neq i} \frac{t_{ik}}{-t_{ii}} A_2(k) \\ &\quad + r_2(i) \frac{1}{-t_{ii}} \sum_{k \neq i} \frac{t_{ik}}{-t_{ii}} A_1(k) + \sum_{k \neq i} \frac{t_{ik}}{-t_{ii}} C_k, \end{aligned}$$

where $A_1(k)$ and $A_2(k)$ are the mean accumulated rewards for each of the two variables given the initial state is k . Multiplying by $-t_{ii}$ on both sides of the equation and applying Eq. (21) we get

$$-t_{ii} C_i = r_1(i) A_2(i) + r_2(i) A_1(i) + \sum_{k \neq i} t_{ik} C_k.$$

We rearrange the terms to get

$$-t_{ii}C_i - \sum_{k \neq i} t_{ik}C_k = r_1(i)A_2(i) + r_2(i)A_1(i).$$

Letting $C = (C_1, \dots, C_p)'$ be the column vector of C_i , $i = 1, \dots, p$, and using Eq. (22) we get

$$(-T)C = \Delta(r_1)(-T)^{-1} \Delta(r_2)e + \Delta(r_2)(-T)^{-1} \Delta(r_1)e,$$

or

$$C = U \Delta(r_1)U \Delta(r_2)e + U \Delta(r_2)U \Delta(r_1)e,$$

where $U = (-T)^{-1}$. If the initial probability vector is α we get

$$\mathbb{E}[Y_1 Y_2] = \alpha U \Delta(r_1)U \Delta(r_2)e + \alpha U \Delta(r_2)U \Delta(r_1)e, \quad (27)$$

which is in accordance with Bladt and Nielsen (2017, page 440).

3.2. Covariances of branch lengths in the n -coalescent

A main advantage of using MPHs is that we can now easily determine the covariance between the different reward-transformed phase-type variables. For example, we can straightforwardly obtain the variance–covariance matrix of the i -ton branch lengths (B_1, \dots, B_{n-1}) for a sample of size n , given that we know the initial state vector, sub-intensity matrix and reward matrix for the block counting process.

Using MPHs, we can also calculate the joint moments of different summary statistics of coalescent trees. For example, for a sample size of n , we can define reward vectors for the total tree height H , the total tree length L , the total length of external branches E , and the total length of internal branches I . These rewards can be computed using the block-counting process of the standard coalescent model (Hobolth et al., 2021). As an example, for $n = 4$, recall Fig. 2E and the corresponding MPH-formulation in Eq. (26). The reward vector for L is given by $r_L = r_1 + r_2 + r_3 = (4, 3, 2, 2)$, which corresponds to the sum of all branches. The reward vector for E is given by $r_E = r_1 = (4, 2, 1, 0)$, because the external branches correspond to the sum of all branches leading to singletons. The reward vector for I is $r_I = r_L - r_E = (0, 1, 1, 2)$, because the total internal branch length is the sum of all branches not leading to singletons. Finally, $r_H = (1, 1, 1, 1)$, corresponding to weighting each state equally. Collected into a reward matrix, these reward vectors can be used to build a MPH with the base sub-intensity matrix in (26). A similar procedure can be used for an arbitrary sample size n .

After having defined the MPH, calculating the covariance between any pair of variables is straightforward using Eqs. (23) and (27), given that $\text{Cov}(Y_1, Y_2) = \mathbb{E}[Y_1 Y_2] - \mathbb{E}[Y_1]\mathbb{E}[Y_2]$. In Fig. 3, we show the covariance and correlation between tree height H , tree length L , external branch length E , and internal branch length I for varying sample sizes in the standard coalescent model. Our Fig. 3 is in accordance with Figure 2 of Alimpiev and Rosenberg (2022), where the authors provide a compendium for covariances and correlation coefficients between pairs of coalescent tree properties. The formulas in Alimpiev and Rosenberg (2022) are derived using careful bookkeeping. Phase-type theory, in contrast, offers an attractive alternative that operates on the matrix level, which avoids cumbersome mathematical derivations.

3.3. Beyond the n -coalescent

The application of the phase-type framework in mathematical population genetics extends well beyond the standard coalescent. The *multiple merger coalescent models* can also be defined within the phase-type framework (Hobolth et al., 2019), as well as the structured coalescent, the coalescent with recombination, and the seed bank coalescent.

For example, in the Λ -coalescent, i sequences coalesce to j lineages ($j = 1, \dots, i-1$) with a rate of

$$\lambda_{i,j} = \int_{[0,1]} x^{j-2}(1-x)^{i-j} \Lambda(dx), \quad (28)$$

where Λ is a probability measure on $[0, 1]$. The Λ -coalescent corresponds to the Kingman coalescent when $\Lambda = \delta_0$, that is, the unit mass at zero. For any other Λ , the sub-intensity matrix for the multiple merger coalescent can be calculated using

$$T = \begin{pmatrix} -g_n & g_{n,2} & g_{n,3} & \cdots & g_{n,n-1} \\ 0 & -g_{n-1} & g_{n-1,2} & \cdots & g_{n-1,n-2} \\ 0 & 0 & -g_{n-2} & \cdots & g_{n-2,n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & -g_2 \end{pmatrix}, \quad (29)$$

where $g_{i,j} = \binom{i}{j} \lambda_{i,j}$ for $i = 2, \dots, k$ and $j = 2, \dots, i$, and $g_i = \sum_{j=2}^i g_{i,j}$. The PH formulation of the tree height under the Λ -coalescent is now straightforward, since $H \sim \text{PH}(\alpha, T)$, where $\alpha = (1, 0, \dots, 0)$ and T is calculated from (29). We can now apply standard phase-type formulas, without the need for specific formulas for the Λ -coalescent. For the definition of the sub-intensity matrix for the Bolthausen–Sznitman coalescent, we refer to Kersting et al. (2021), and for other multiple merger models, we refer to Birkner and Blath (2021).

Blath et al. (2020) use phase-type theory to obtain the expected site frequency spectrum for the seed bank coalescent and the two-island structured coalescent. The correlation structure between tree heights in the coalescent with recombination with two samples and two loci (e.g. Wakeley (2008), Chapter 7.2) is described in Rivas-González et al. (2023).

4. The number of segregating sites and the discrete phase-type distribution

4.1. The classical approach

The number of segregating sites S is the total number of mutations that have happened along the coalescent process. Using the infinite-sites model, the mutation rate when i branches are present is $v_i = i\theta/2$. For $i = 2$, the mutation rate is $v_2 = \theta$. Mutations are allowed to occur until the two sequences find common ancestry, which happens with a coalescence rate of $\lambda_i = \binom{i}{2} = i(i-1)/2$. The number of segregating sites S when $i = 2$ is therefore geometrically distributed $S \sim \text{Geo}(p_2)$, with probability p_2 that a mutation occurs before a coalescent event. Here, we define the geometric distribution $X \sim \text{Geo}(p)$ as $\mathbb{P}(X = x) = p^x(1-p)$ for $x = 0, 1, \dots$, and $0 < p < 1$. To calculate p_2 , we can view the situation as a competition between two independent events with exponential waiting times, namely a mutation event and a coalescent event. In such cases, the probability that a particular event happens before the other can be computed as the relative rate of the events of interest. Thus,

$$p_2 = \mathbb{P}(\text{mutation} \mid \text{coalescent or mutation}) = \frac{v_2}{v_2 + \lambda_2} = \frac{\theta}{\theta + 1}.$$

The geometric distribution also applies for all other stages of the coalescent process, i.e., when i branches are present, the number of mutations that happen before coalescence into $i-1$ sequences is $S_i \sim \text{Geo}(p_i)$, where

$$p_i = \frac{v_i}{v_i + \lambda_i} = \frac{i\theta/2}{i\theta/2 + i(i-1)/2} = \frac{\theta}{\theta + i - 1}. \quad (30)$$

To obtain the full distribution of the total number of segregating sites with an initial sample size of n , we thus need to sum the independent geometric distributions, so that $S = \sum_{i=2}^n S_i$. Similar to how the distribution for the tree height H is calculated, the probability density function for S , $\mathbb{P}(S = s)$, can be calculated using convolutions of geometric distributions.

Another classical way of formulating the probability density function for S is by realizing that mutations happen following a Poisson process sprinkled over the evolutionary tree, i.e., $S \mid L \sim \text{Poisson}(\theta L/2)$, where L is the total tree length. By knowing the distribution of L , we can use formulas for calculating the marginal distribution of S given

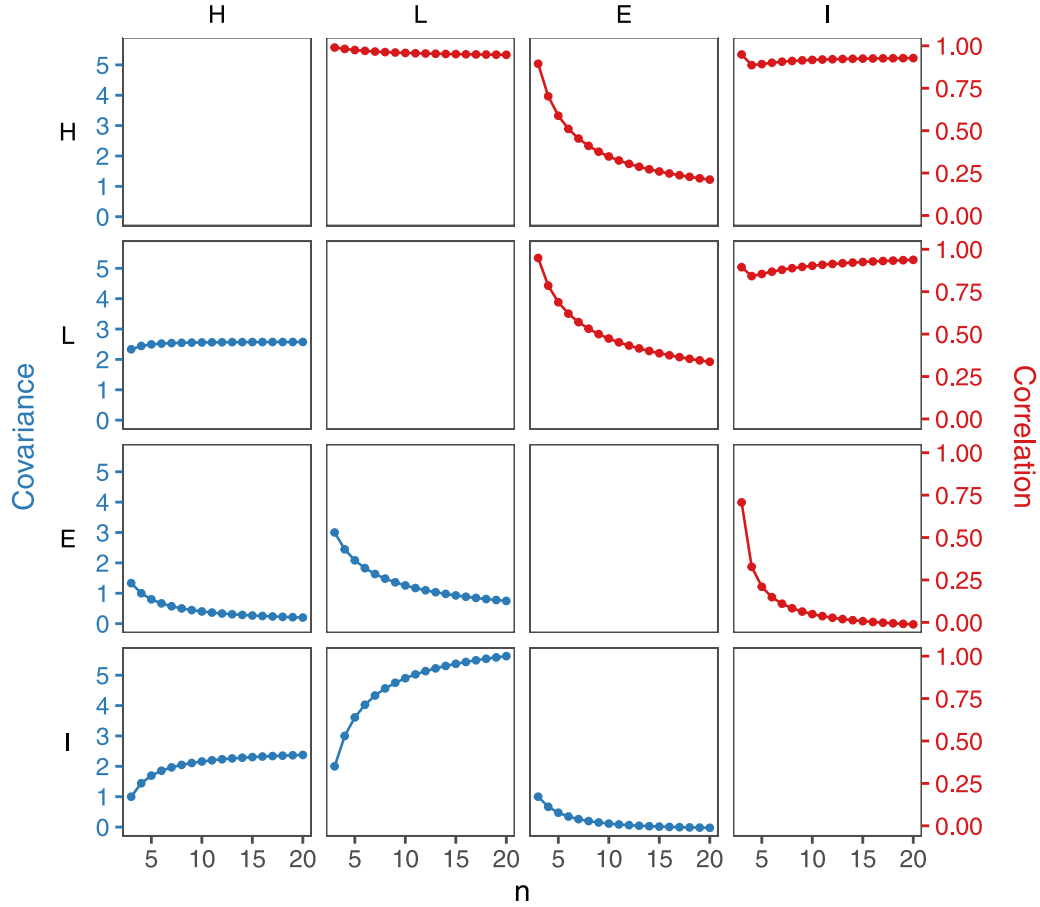


Fig. 3. Covariance (left axis, blue) and correlation (right axis, red) between tree height H , tree length L , external branch length E , and internal branch length I in the Kingman coalescent for varying sample size n .

the total tree length, and then integrate over all possible tree lengths such that

$$\mathbb{P}(S = s) = \int_0^\infty \mathbb{P}(S = s \mid t) f_L(t) dt. \quad (31)$$

We refer to Wakeley (2008, equation 4.3) for further details.

The issue with these approaches is that they both require convolutions of either geometric distributions to directly calculate $\mathbb{P}(S = s)$, or exponential distributions to calculate $f_L(t)$. This might become a problem, for example, when trying to calculate the distribution of S for non-standard coalescent models such as multiple merger models. An attractive alternative is to use phase-type distributions.

4.2. Using phase-type distributions

We can circumvent convolutions by embedding the number of segregating sites S into a discrete-time Markov chain $\{X_t\}_{t \in \mathbb{N}}$. For an initial sample size of $n = 3$, there are three possible states. The $p = 2$ transient states correspond to 3 branches in the ancestral process or 2 branches in the ancestral process, and the last state is the most recent common ancestor (the absorbing state). The probability of jumping from one state to another can be defined using a transition probability matrix A . Because $\{X_t\}_{t \in \mathbb{N}}$ only has a single absorbing state, A can be partitioned into

$$A = \begin{pmatrix} T & t \\ \mathbf{0} & 1 \end{pmatrix}, \quad (32)$$

where

$$T = \begin{pmatrix} p_3 & (1-p_3)p_2 \\ 0 & p_2 \end{pmatrix} \quad \text{and} \quad t = \begin{pmatrix} (1-p_3)(1-p_2) \\ 1-p_2 \end{pmatrix} \quad (33)$$

where p_i is given by (30). Here, T is a sub-transition matrix of size $p \times p$ which holds the transition probabilities among the transient states, t is a column vector of size p with the exit probabilities from the transient states into the absorbing state, and $\mathbf{0}$ is a row vector of zeros of size p . Since $A_{ij} = \mathbb{P}(X_{t+1} = j \mid X_t = i)$, each row in A sums to 1, i.e. $A\mathbf{e} = \mathbf{e}$. Thus, the exit probability vector is given by $t = \mathbf{e} - T\mathbf{e} = (\mathbf{I} - T)\mathbf{e}$, so the Markov chain can be defined solely by the sub-transition probability matrix T .

Let S be the number of jumps (or mutations) until absorption of the Markov chain so that $S = \inf\{t \geq 0 : X_t = p + 1\}$. If we define α , a vector of initial probabilities, such that $\mathbb{P}(X_0 = i) = \alpha_i$ for $i = 1, \dots, p$, then the total number of jumps until absorption follows a discrete phase-type distribution, i.e. $S \sim \text{DPH}(\alpha, T)$. In other words, S is the total number of jumps (or mutations) of a series of sequentially occurring geometric distributions, which is a Markov chain defined by a starting probability vector α and a sub-transition matrix T holding the transition probabilities among the p transient states. For S , when $n = 3$, $\alpha = (p_3, (1-p_3)p_2)$. Note that, similarly as in the continuous case, α does not necessarily need to sum to 1, since the Markov chain could also start directly in the absorbing state. In this case, this would correspond to all sequences having coalesced without any mutations. The probability of this happening is called the defect, which can be calculated as $\mathbb{P}(X_0 = p + 1) = 1 - \sum_{i=1}^p \alpha_i$.

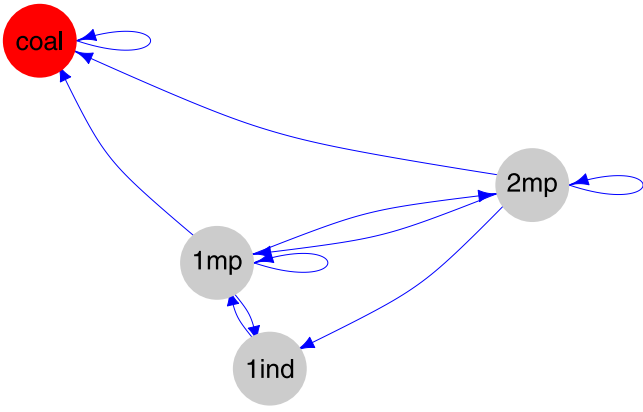


Fig. 4. Adjacency graph for the sib-mating model for two lineages. The states are “lineages within 1 individual” (1ind), “lineages within one mating pair” (1mp), “lineages in two mating pairs” (2mp), and coalescence (“coal”). The discrete time Markov chain may start in either of the transient states. Possible transitions are represented by arrows. The transition rates can be found in the sub-transition matrix T . PhaseTypeR was used to create the graph.

An alternative procedure of arguing for the discrete phase-type distribution for S is by using Theorem 3.5 in Hobolth et al. (2021). The theorem states that overlaying a Poisson process on a continuous PH distribution results in a DPH distribution. Mathematically, if $S \mid L \sim \text{Po}(\theta L)$ and $L \sim \text{PH}(\alpha_L, T_L)$, then $S + 1 \sim \text{DPH}(\pi, P)$, where

$$P = \left(I - \frac{2}{\theta} T_L\right)^{-1}, \text{ and } \pi = \alpha_L. \quad (34)$$

An equivalent formulation is $S \sim \text{DPH}(\pi', P)$, where P is as in (34), but now $\pi' = \alpha_L P$. Applying this theorem avoids the need for keeping track of individual geometric distributions, and it allows for easily obtaining a discrete phase-type representation of S for an arbitrary number of initial sequences n or for non-standard coalescent models, given that a continuous phase-type representation of L is available. For example, obtaining a phase-type representation of the number of segregating sites for the Λ -coalescent is straightforward by applying Eq. (34) to the sub-intensity matrix calculated from Eq. (29).

As for the continuous case, the main advantage of discrete phase-type distributions is that they have closed formulas for the mean, variance, moments, probability density function, and cumulative distribution function. These formulas are in matrix notation, so these properties can be calculated based solely on α and T .

4.3. Coalescence times for diploid consanguineous populations

Campbell (2015) considered the ancestral history of two lineages, in a model involving N diploid mating pairs resulting in a total of $2N$ individuals. In each generation cN of the mating pairs are randomly chosen to be consanguineous (siblings). Two lineages are followed backwards in time, and the expected time until their common ancestor is derived. This model can be phrased in terms of a Markov chain consisting of three transient and one absorbing state. The absorbing state is the coalescence of the two lineages. The transient states are – in this order – two lineages being in the same individual (1ind), two lineages in two individuals of the same mating pair (1mp), and two lineages being in two different mating pairs (2mp).

The adjacency graph can be found in Fig. 4, and the sub-transition probability matrix between the states is given by

$$T = \begin{pmatrix} 0 & 1 & 0 \\ \frac{c}{4} & \frac{c}{2} & 1-c \\ \frac{1}{4N} & \frac{1}{2N} & 1 - \frac{1}{N} \end{pmatrix}$$

The time until absorption has a discrete phase-type distribution. We may therefore use Corollary 1.2.64 in Bladt and Nielsen (2017) to obtain the expected time until absorption as $E(\tau) = \alpha(I - T)^{-1}e$ with starting distribution α and the vector of ones $e = (1, \dots, 1)'$. Using that

$$(I - T)^{-1} = \begin{pmatrix} 2 & 4 & 4(1-c)N \\ 1 & 4 & 4(1-c)N \\ 1 & 3 & 4\left(1 - \frac{3c}{4}\right)N \end{pmatrix},$$

we get

$$(I - T)^{-1}e = \begin{pmatrix} 4(1-c)N + 6 \\ 4(1-c)N + 5 \\ 4\left(1 - \frac{3c}{4}\right)N + 4 \end{pmatrix}.$$

The three entries in this vector correspond to the three possible starting states (1ind, 1mp, 2mp). This result reproduces findings in Campbell (2015), and can also be found as equations (4)–(6) in Severson et al. (2019).

Severson et al. (2021) also derive the variance of τ . Using a phase-type approach, the variance may be alternatively obtained using the expression for factorial moments provided by Theorem 1.2.69 in Bladt and Nielsen (2017). For the second factorial moments, this formula translates into

$$E[\tau(\tau - 1)] = 2\alpha T(I - T)^{-2}e.$$

In our case we get

$$2T(I - T)^{-2}e = \begin{pmatrix} 8(1-c)(4-3c)N^2 + 72(1-c)N + 52 \\ 8(3c^2 - 7c + 4)N^2 + 64(1-c)N + 42 \\ 2(4-3c)^2N^2 + (56-50c)N + 34 \end{pmatrix},$$

with the components representing the three possible starting states (1ind, 1mp, 2mp).

The variance is given by $E(\tau(\tau - 1)) + E(\tau) - [E(\tau)]^2$, and for the three different initial states we get

$$\begin{pmatrix} 8(c^2 - 3c + 2)N^2 + 28(1-c)N + 22 \\ 8(c^2 - 3c + 2)N^2 + 28(1-c)N + 22 \\ (4-3c)^2N^2 + (28-29c)N + 22 \end{pmatrix}.$$

Higher order moments can be obtained using similar computations.

4.4. The wright–Fisher process

Other classical quantities in population genetics apart from the number of segregating sites also follow discrete phase-type distributions. The time to fixation (in the number of generations) of the Wright–Fisher process is one example. Kruk et al. (2016) used classical absorbing Markov chain theory to determine the probability of fixation or extinction of an allele, the expected time to fixation or extinction, and other key properties of the Wright–Fisher process. Kruk et al. (2016) describes a very efficient and scalable solution for linear systems. Here, we provide the solutions of the linear systems using phase-type theory.

Given a panmictic, haploid population with a finite population size N , let a_t denote the number of individuals carrying a certain allele in generation t , and $f_t = a_t/N$ the frequency of the allele. Because the population has a finite size, there are only $N + 1$ possible allele frequencies ranging from lost ($f_t = 0$) to fixed ($f_t = 1$). The frequency of an allele in the population depends on the frequency of that allele in the previous generation. This way, the probability of observing a certain frequency can be calculated through binomial sampling from the previous generation, by fixing the number of trials to N and the probability of success to f_t . The dynamics are thus given by

$$a_{t+1} | a_t \sim \text{Binom}(N, f_t), \text{ where } f_t = a_t/N. \quad (35)$$

Following the probability mass function of the binomial distribution, we can calculate the probability of observing any of the intermediate frequencies f_{t+1} in the next generation for a certain f_t .

We can then define τ as the number of generations until the allele becomes either fixed or lost. This way, τ can be described in the phase-type framework as $\tau \sim \text{DPH}(\alpha_{\text{WF}}, T_{\text{WF}})$ by letting the transient states p be the $(N - 1)$ possible frequencies f_t between fixation or loss. The probabilities calculated through binomial sampling are the probabilities of jumping from a certain frequency f_t in generation t to a frequency f_{t+1} in generation $t + 1$. Thus, these probabilities can be summarized in the sub-transition matrix T_{WF} with entries

$$T_{\text{WF}}[i, j] = b\left(\frac{i}{N}, j\right), \quad 1 \leq i, j \leq N - 1, \quad (36)$$

where $b(p, k) = \binom{N}{k} p^k (1 - p)^{N-k}$. Additionally, all the starting probabilities in α_{WF} are set to 0 except the entry corresponding to the starting frequency f_0 , which is set to 1.

Instead of an allele drifting neutrally, the Wright–Fisher model can also be used to model alleles under selection. The time to fixation or loss is then $\tau' \sim \text{DPH}(\alpha'_{\text{WF}}, T'_{\text{WF}})$, where

$$T'_{\text{WF}}[i, j] = b\left(\frac{i(1+s)}{i(1+s) + N - i}, j\right), \quad 1 \leq i, j \leq N - 1, \quad (37)$$

where s is the selection coefficient for the focal allele (Etheridge, 2011).

By plotting the density function of τ' under different selective forces, we can observe a bimodal distribution for intermediate s values (see Fig. 5, black lines). This happens because the absorbing state of T'_{WF} includes both instances when an allele is either lost or fixed. In order to distinguish these two processes, i.e., in order to know the distribution of the time to loss and the time to fixation separately, we must transform the underlying phase-type distribution based on the exit vectors of each of the two processes.

Let A be a transition probability matrix with p transient states and q absorbing states, such that

$$A = \begin{pmatrix} T & t_1 & t_2 & \cdots & t_q \\ \mathbf{0} & 1 & 0 & \cdots & 0 \\ \mathbf{0} & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & 0 & 0 & 0 & 1 \end{pmatrix}, \quad (38)$$

where T is a sub-intensity matrix of size p and t_i is the exit rate vector into absorbing state $p + i$. If we define $\tau = \inf\{n \geq 0 : X_n \in \{p + 1, \dots, p + q\}\}$, then $\tau \sim \text{DPH}(\pi, T)$. The probability of exiting into absorbing state $p + i$ is then given by

$$\mathbb{P}(X_\tau = p + i) = \sum_{t=1}^{\infty} \pi T^{t-1} t_i = \pi(I - T)^{-1} t_i. \quad (39)$$

Additionally, it has been shown through time reversal that a discrete phase-type distributed variable with more than one absorbing state is still discrete phase-type distributed conditioned on being absorbed in a certain absorbing state (Gardner et al., 2021), so $\tau \mid (X_\tau = p + i) \sim \text{DPH}(\pi_i, T_i)$. The conditional phase-type distribution can be obtained by modifying the original sub-probability matrix T and initial probability vector π using the exit probability vector t_i of the corresponding absorbing state i to obtain a new sub-probability matrix T_i and initial probability vector π_i for each of the absorbing states (Gardner et al., 2021, Section 3.2).

Thus, one can obtain phase-type representations of the Wright–Fisher model with selection for the time until absorption conditional on loss or fixation. Let $t_1 = t_{\text{loss}}$ and $t_2 = t_{\text{fix}}$ be the exit vector rates for the loss and the fixation, respectively. We can define these to be

$$t_{\text{loss}}[i] = b\left(\frac{i(1+s)}{i(1+s) + N - i}, 0\right), \quad 1 \leq i \leq N - 1, \\ t_{\text{fix}}[i] = b\left(\frac{i(1+s)}{i(1+s) + N - i}, N\right), \quad 1 \leq i \leq N - 1.$$

It is now straightforward to obtain DPH representations conditional on these two absorbing states. Fig. 5 shows the conditional density for loss and fixation weighted by the probability of loss and fixation (see Eq. (39)), respectively, on top of the original phase-type distribution for different selection coefficients.

Apart from the discrete case, continuous phase-type distributions with multiple absorbing states can also be conditioned on a certain absorbing state, and the resulting distribution will also be a continuous phase-type (Andersen et al., 2000).

4.5. Reward transformation in discrete phase-type distributions

In addition to the continuous case, DPHs can also be reward-transformed (Campillo Navarro, 2018). Let $\tau \sim \text{DPH}(\alpha, T)$, where $\{X_t\}_{t \in \mathbb{N}}$ is the underlying Markov chain. We can define a vector $r = (r(1), r(2), \dots, r(p))$ consisting of non-negative integer rewards so that a new random variable τ' satisfies

$$\tau' = \sum_{t=0}^{\tau} r(X_t). \quad (40)$$

It can be shown that τ' is also DPH (Campillo Navarro, 2018). This is particularly useful to study the elements of the site frequency spectrum, since singletons, doubletons, etc. are all reward-transformed versions of the total number of segregating sites (see Hobolth et al. (2021) for further details).

5. The probability of a configuration and statistical inference

Phase-type theory also provides a systematic approach to computing probabilities for observed mutational patterns. These probabilities may then be used for subsequent likelihood-based statistical inference. Due to a large number of possible tree topologies, likelihood computations are notoriously difficult for coalescent models. Therefore, statistical inference often relies on summary statistics. In this section, we explain how phase-type theory provides a principled approach to obtain likelihoods for small sample sizes, as well as distributions for summary statistics such as the total number of segregating sites and features of the site frequency spectrum (SFS).

As a first example, consider the estimation of the scaled mutation parameter θ . Different estimators of this parameter have been proposed, such as Watterson's estimator, Tajima's π , or the estimate $\hat{\theta}_H$ in Fay and Wu (2000). In Hobolth et al. (2021), it is explained how the distributions of these estimates can be obtained using a phase-type approach. Again, this provides a flexible and general tool to obtain both univariate and multivariate distributions. It can be applied to several features of the site frequency spectrum, such as linear combinations used with test statistics such as Tajima's D . Indeed, Hobolth et al. (2021, page 15) explains how phase-type distributions are constructed by using a block-counting approach. Their Figures 3 and 4 provide illustrative examples for samples of size $n = 4$ and $n = 5$. The obtained distributions are useful when constructing cut-off values for hypotheses testing, and calculating confidence intervals. Although these distributions can also be obtained directly, see for instance equation (4) in Griffiths and Tavaré (2018) or (Tavaré, 2004) for the probability generating function of the number of segregating sites, the phase-type approach is quite convenient in terms of its generality and flexibility.

In this section, we illustrate that phase-type theory also provides tools to compute full small sample likelihoods for subsequent statistical inference. We will also explore the connection to the generating function approach by Lohse et al. (2011).

5.1. Likelihoods via Laplace transforms

We consider full likelihood inference in the framework of the infinite sites model where data consists of homologous DNA sequences. In the standard coalescent model, the data can be used to infer the scaled

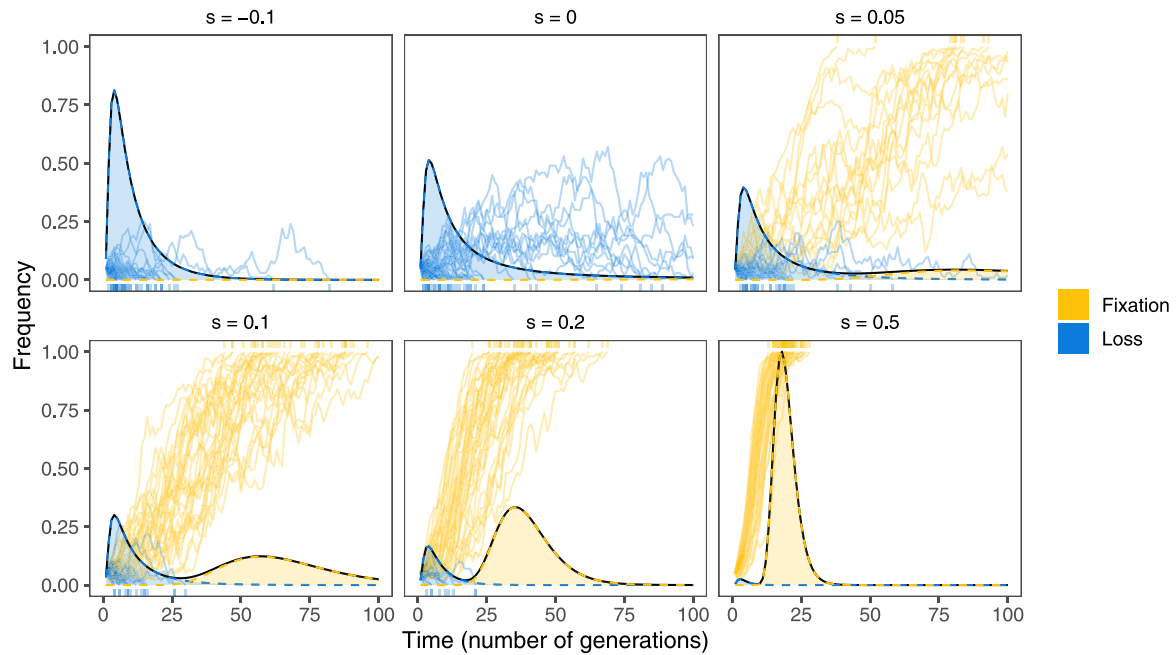


Fig. 5. Wright–Fisher model with selection, for $N = 100$ and an initial frequency for the selected allele of 0.05. The selection coefficient s is shown on top of every panel. The (re-scaled) phase-type density of the number of generations to either loss or fixation is shown as a black line. The full path of the phase-type distribution was simulated 50 times for each s , and each simulation is plotted as an allelic trajectory colored by whether the end state was loss (blue) or fixation (yellow). Blue and yellow lines and areas correspond to the conditional densities of the time to loss or fixation, respectively. These conditional densities are weighted by the probability of exiting into each absorbing state in order to match the phase-type density.

mutation rate θ . Due to a large number of possible coalescent tree topologies, full likelihood inference eventually becomes computationally infeasible with an increasing number of sampled sequences. Therefore, importance sampling methods such as *genetree* have been proposed (see Griffiths (1989), Griffiths and Tavaré (1994)) to approximate the likelihood under an infinite sites model.

Nevertheless, progress has been made, and for small samples Lohse et al. (2011), Uyenoyama et al. (2019) and Uyenoyama et al. (2020) suggested full likelihood methods for inference also for demographic parameters.

5.1.1. Joint likelihood for mutation counts on coalescent branches

In Lohse et al. (2011), the authors use generating functions (Laplace transforms) to obtain joint probabilities for the number of segregating sites on the branches of a small genealogy. This work has subsequently been extended; see, for instance, Lohse et al. (2016), and Bisschop (2022) for an efficient evaluation of the Laplace transform using a graph traversal algorithm.

The basic idea of their approach may be summarized as follows. Suppose there is a branch s in a standard coalescent tree and the mutation rate is $\lambda = \theta/2$ per unit length. A standard assumption is that the number of observed mutations M_s given the length t_s of lineage s is Poisson(λt_s). Then the probability of observing m_s mutations on branch s is

$$\begin{aligned} \mathbb{P}(M_s = m_s) &= \mathbb{E}_{T_s}[\mathbb{P}(M_s = m_s | T_s)] = \mathbb{E}_{T_s}\left[e^{-\lambda T_s} \frac{(\lambda T_s)^{m_s}}{m_s!}\right] \\ &= \frac{\lambda^{m_s}}{m_s!} \frac{\partial^{m_s}}{(\partial \omega)^{m_s}} \psi_s(\omega) \Big|_{\omega=\lambda}, \end{aligned}$$

where $\psi_s(\omega) = \mathbb{E}[e^{-\omega T_s}]$. If T_s is exponentially distributed with rate parameter γ , then $\psi_s(\omega) = \frac{\gamma}{\gamma + \omega}$. Applying this observation to a standard coalescent with two samples (denoted by a and b), we have $\gamma = 1$ (see e.g. page 76 in Wakeley (2008)). Thus, for lineage a

$$\mathbb{P}(M_a = m_a) = \frac{(-\lambda)^{m_a}}{m_a!} \frac{\partial^{m_a}}{(\partial \omega)^{m_a}} \left(\frac{1}{1 + \omega} \right) \Big|_{\omega=\lambda} = \frac{\lambda^{m_a}}{(1 + \lambda)^{m_a+1}}.$$

By using the multivariate Laplace transform

$$\psi(\omega_a, \omega_b) = \mathbb{E}[e^{-\omega_a T_a - \omega_b T_b}] = \mathbb{E}[e^{-(\omega_a + \omega_b) T_a}] = \frac{1}{1 + \omega_a + \omega_b},$$

two lineages may be considered simultaneously. Notice that here $T_a = T_b$, since the two lineages share the time until their coalescence. Similar arguments as above may now be used to obtain the joint probability of observing m_a mutations on branch a , and m_b mutations on b . This gives

$$\mathbb{P}(M_a = m_a, M_b = m_b) = \binom{m_a + m_b}{m_a} \frac{\lambda^{m_a + m_b}}{(1 + 2\lambda)^{m_a + m_b + 1}}. \quad (41)$$

This approach may be extended also to a few more lineages and to some basic demographic models. It also provides us with the likelihood when the resulting probabilities are viewed as a function of λ . We now explain how analogous results can be obtained via phase-type distributions.

5.1.2. General phase-type approach

To illustrate how results such as the one from the previous subsection can be obtained via a phase-type approach, we will use Theorem 2.7 in Hobolth et al. (2021). As explained above in Section 4.2, the theorem describes how to add Poisson-distributed mutations onto a (multivariate) phase-type distribution and provide the joint probability generating function of mutation counts on disjoint branches of the coalescent tree. More specifically, the result assumes multivariate phase-type distributed random variables (Y_1, \dots, Y_d) that are defined using rewards. Conditional on (Y_1, \dots, Y_d) , the random variables (Z_1, \dots, Z_d) are then taken as independent Poisson distributed with rates λY which implies $Z_j | Y_j \sim \text{Poisson}(\lambda Y_j)$. In this setup, the probability generating function of (Z_1, \dots, Z_d) is given as

$$\phi(\mathbf{z}) = \alpha (\lambda \Delta (\mathbf{R}[\mathbf{e} - \mathbf{z}]) - \mathbf{T})^{-1} \mathbf{t},$$

where $\Delta(\cdot)$ gives a diagonal matrix with entries specified by the argument.

We now explain this formula in a coalescent context. Assuming that the underlying Markov process has p states besides the absorbing state,

the p -dimensional vector α provides the starting distribution. In our context, it assigns probability one to the starting configuration before the first coalescence event. The reward matrix $\mathbf{R} \in \mathbb{R}_{+}^{p \times d}$ represents branches or sets of branches on which independent mutation counts (Z_1, \dots, Z_d) occur. Due to the independence assumption, the same mutations should not contribute to different Z_i 's. Each column of \mathbf{R} represents one mutation count. The $p \times p$ sub-intensity matrix \mathbf{T} contains the transition rates between the transient states. Finally, the vector $\mathbf{t} = -\mathbf{T}\mathbf{e}$ provides the absorption rate for each transient state. The vector $\mathbf{z} = (z_1, \dots, z_d)'$ contains the variables of the generating function for the mutation counts.

The joint distribution of the mutation counts of interest Z_i ($1 \leq i \leq d$) can now be obtained by taking suitable partial derivatives of the probability-generating function. More specifically,

$$\mathbb{P}(Z_i = m_i, 1 \leq i \leq d) = \frac{1}{m_1! \dots m_d!} \frac{\partial^{\sum_i m_i}}{(\partial z_1)^{m_1} \dots (\partial z_d)^{m_d}} \phi(z_1, \dots, z_d)|_{(0, \dots, 0)}. \quad (42)$$

5.1.3. Phase-type for two samples

With two samples, there are two possible states: the initial state (with two lineages), and the absorbing state after coalescence. The transition rate to the absorbing state is 1. Therefore, $p = 1$ and \mathbf{T} becomes the scalar -1 . Furthermore, the reward matrix $\mathbf{R} = (1, 1)$ uses the same reward twice to capture both branches of equal length. Additionally, the initial state vector is $\alpha = 1$ and the exit rate is $\mathbf{t} = 1$, since there is only one transient state. Together this leads to

$$\phi(z) = 1 \cdot (\lambda(z_1 + z_2 - 2) - 1)^{-1} \cdot 1 = \frac{1}{1 - \lambda(z_1 + z_2 - 2)}.$$

Therefore, the probability of no mutations on either branch is

$$\mathbb{P}(Z_1 = 0, Z_2 = 0) = \phi(0, 0) = \frac{1}{1 + 2\lambda}.$$

According to Eq. (42), $\mathbb{P}(Z_i = m_i, 1 \leq i \leq 2)$ can be obtained by taking partial derivatives. Indeed, with m_1 derivatives with respect to z_1 , and m_2 derivatives with respect to z_2 , an evaluation at $(0, 0)$ reproduces formula (41).

As numerical examples, the probability of two mutations on branch a , and one at branch b is given by

$$\mathbb{P}(Z_1 = 2, Z_2 = 1) = \frac{1}{2!1!} \frac{\partial^3}{(\partial z_1)^2 (\partial z_2)} \phi(z_1, z_2)|_{(0,0)} = \frac{3\lambda^3}{(2\lambda + 1)^4},$$

and, for $\lambda = 20$,

$$\mathbb{P}(Z_1 = 24, Z_2 = 17) = \frac{151584480450\lambda^{41}}{(2\lambda + 1)^{42}} \approx 0.000611.$$

It is now easy to obtain the maximum likelihood estimate (MLE) of λ by setting the first derivative of (41) to zero. This leads to $\hat{\lambda} = (m_1 + m_2)/2$ as our MLE.

5.1.4. Phase-type for three samples

We next extend to a coalescent model for three samples. Our model (see Fig. 6) may be represented by a Markov chain with two transient states $(3, 0, 0)$ and $(1, 1, 0)$ that represent the initial configuration with three singleton lineages, and the configuration with one singleton and one doubleton lineage after the first coalescence event. The absorbing state is denoted by $(0, 0, 1)$ and represents the most recent common ancestor of our three samples.

As the first transition occurs at rate 3, and the second at rate 1, the sub-transition matrix for the transient states is given by

$$\mathbf{T} = \begin{pmatrix} -3 & 3 \\ 0 & -1 \end{pmatrix}.$$

We next derive the joint distribution of the mutation count vector (M_1, M_2, M_3, M_4) that belong to the four branches as shown in Fig. 6.

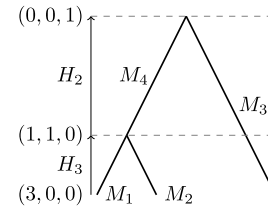


Fig. 6. Mutation patterns and Markov chain structure for a coalescent with $n = 3$ samples. The probability generating function of the four mutation counts M_1, M_2, M_3, M_4 can be obtained by specifying a multivariate reward matrix consisting of four suitable reward vectors.

For this purpose, we use the four linear reward functions specified by the columns of

$$\mathbf{R} = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix}.$$

The first two columns represent the two lineages of equal length H_1 that underlie M_1 and M_2 , the third column is for M_3 (branch length $H_3 + H_2$), and the last one for M_4 . Furthermore,

$$\mathbf{t} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

provides the rates of directly entering the absorbing state from the two transient states $(3, 0, 0)$ and $(1, 1, 0)$. Finally, $\alpha = (1, 0)$ since we start in the first state $(3, 0, 0)$. Together this leads to

$$(\lambda \Delta (\mathbf{R}[\mathbf{e} - \mathbf{z}]) - \mathbf{T})^{-1} = \begin{pmatrix} \frac{1}{3\lambda - \lambda(z_1 + z_2 + z_3) + 3} & \frac{3}{(2\lambda - \lambda(z_3 + z_4) + 1)(3\lambda - \lambda(z_1 + z_2 + z_3) + 3)} \\ 0 & \frac{1}{2\lambda - \lambda(z_3 + z_4) + 1} \end{pmatrix}$$

and the generating function

$$\phi(z_1, z_2, z_3, z_4) = \frac{3}{(2\lambda - (\lambda(z_3 + z_4) + 1) + 1)(3\lambda - (\lambda(z_1 + z_2 + z_3) + 3)) + 3}. \quad (43)$$

Thus the probability of no mutations on all branches is

$$\mathbb{P}(M_1 = 0, M_2 = 0, M_3 = 0, M_4 = 0) = \phi(0, 0, 0, 0) = \frac{3}{(1 + 2\lambda)(3 + 3\lambda)}.$$

As with two branches, a general formula for the joint mutational patterns can be obtained by taking suitable partial derivatives. Indeed, $\mathbb{P}(M_1 = m_1, M_2 = m_2, M_3 = m_3, M_4 = m_4)$ is equal to

$$p(m_1, m_2, m_3, m_4) = \frac{1}{m_1! m_2! m_3! m_4!} \frac{\partial^{m_1 + m_2 + m_3 + m_4}}{(\partial z_1)^{m_1} (\partial z_2)^{m_2} (\partial z_3)^{m_3} (\partial z_4)^{m_4}} \phi(z_1, z_2, z_3, z_4)|_{(0,0,0,0)}.$$

The resulting derivatives may be rewritten in a simpler form as

$$p(m_1, m_2, m_3, m_4) = \frac{3\lambda^{m_1 + m_2 + m_3 + m_4}}{m_1! m_2! m_3! m_4!} \times \sum_{j=0}^{m_3} \binom{m_3}{j} \frac{(m_4 + j)!(m_1 + m_2 + m_3 - j)!}{(1 + 2\lambda)^{m_4 + j + 1} (3 + 3\lambda)^{m_1 + m_2 + m_3 - j + 1}}, \quad (44)$$

with the sum being over the possible allocations of M_3 to the two branch parts with lengths H_3 and H_2 . As a numerical example, we obtain $p(3, 1, 4, 2) \approx 0.000254$ for $\lambda = 2$.

We next consider the labeled coalescent and assign samples a , b , and c to the leaves of the tree in Fig. 6. There are $3! = 6$ possible such assignments that occur with equal prior probability. We want to compute probabilities for the mutational configurations $(M_a, M_b, M_c, M_{ab}, M_{ac}, M_{bc})$. First, notice that under the infinite site model, only one of the doubleton mutation counts can be nonzero. Furthermore, since the configuration of doubleton mutations identifies

Table 2

State number and the corresponding configuration of the 11 states of a three-sample model with migration. The first 10 states are transient, and state 11 is absorbing and corresponds to the most recent common ancestor (MRCA). The first 4 states have three branches and states 5–10 have two branches.

Three branches				Two branches				MRCA		
1	2	3	4	5	6	7	8	9	10	11
$(a, b c)$	(a, b, c)	$(a b, c)$	$(b a, c)$	$(ab c)$	(ab, c)	$(a bc)$	$(b ac)$	(ac, b)	(a, bc)	(abc)

the sequences belonging to the short branches under the infinite sites model, the probability of observing corresponding mutation counts is nonzero for only two out of six possible permutations of a, b, c . These two permutations only differ with respect to the assignments within the two shorter branches and therefore have the same probability. Assume now that there is a nonzero number of doubleton mutations, and let for instance $M_{ab} > 0$. In this case

$$\mathbb{P}(M_a = m_a, M_b = m_b, M_c = m_c, M_{ab} = m_{ab}) = \frac{2}{6} p(m_a, m_b, m_c, m_{ab}).$$

For other nonzero doubleton mutation counts, the formula is analogous.

If there are no doubleton mutations, i.e. $M_{ab} = M_{ac} = M_{bc} = 0$, then all six permutations lead to nonzero probabilities for corresponding mutational configurations. The six assignments of the label permutations to the tree in Fig. 6 can then be partitioned into three pairs of equal probability (again due to permutations within the shorter branches), and therefore

$$\begin{aligned} &\mathbb{P}(M_a = m_a, M_b = m_b, M_c = m_c, M_{ab} = 0, M_{bc} = 0, M_{ac} = 0) \\ &= \frac{1}{3} \left(p(m_a, m_b, m_c, 0) + p(m_a, m_c, m_b, 0) + p(m_c, m_b, m_a, 0) \right). \end{aligned}$$

5.2. Two sub-populations with three samples

We now consider a model with two subdivided populations (or demes) and migration at rate μ between the populations. Likelihood inference for such models has been considered for instance by Costa and Wilkinson-Herbots (2017, 2021). In this work, (composite) likelihoods for two samples are obtained via general Markov chain theory. Here we obtain likelihoods for three samples using phase-type theory. The mutation rate is assumed to be λ in both populations, and the two populations are assumed to be of equal size. We assume that a and b are sampled in the first population and c is taken from the other. We denote this configuration by $(a, b|c)$. As mutation and coalescence rates are equal for both populations and the migration rate between the two populations is also equal, we have full symmetry. We, therefore, do not distinguish between configurations that are swapped across populations. For instance, $(a, b|c)$ and $(c|a, b)$ are combined into a single state. Following this strategy, we have a total of 11 states, with $(a, b|c)$ being the initial state, and $(abc|)$ being the final absorbing state. Samples that are not separated by a vertical bar or a comma are assumed to have coalesced. For instance, abc represents the state where all three lines have coalesced. Table 2 summarizes the 11 states of our three-lineage model with two subdivided populations.

The sub-transition matrix T between the transient states is given in Box 1. In Fig. 7, we show the adjacency graph for the three sample model with migration.

There are six types of lineages characterized by the labeled leaves that they subtend: singleton a , b and c , and doubleton ab , ac and bc (recall Table 2). The rewards are constructed to measure the contribution of the states to these lineages. They are represented by the following 10×6 reward matrix R , with the rows representing the transient states

ordered in the same way as in T .

$$R = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \quad (46)$$

The column entries are either zero or one, depending on whether the lineage represented by a column is present in a given state. The columns of R are allocated to the lineages in the order singleton mutation in a , singleton mutation in b , singleton mutation in c , doubleton mutation in a and c , doubleton mutation in b and c , and doubleton mutation in a and b . As before, the generating function can be obtained as

$$\phi(z_1, z_2, z_3, z_4, z_5, z_6) = \left(\lambda \Delta (R(e - z)) - T \right)^{-1},$$

with $z = (z_1, z_2, \dots, z_6)$. Probabilities for a given mutational pattern can again be computed by taking suitable partial derivatives of $\phi(\cdot)$ and evaluating at $z = (0, \dots, 0)$. The partial derivatives of $\phi(\cdot)$ can either be computed directly or by using the matrix identity

$$\frac{\partial}{\partial z_i} A^{-1}(z) = A^{-1}(z) \left[\frac{\partial}{\partial z_i} A(z) \right] A^{-1}(z),$$

and the chain rule. By plugging in values for the mutation rate λ and the migration rate μ , likelihoods can be obtained for arbitrary parameter values. Analytic computation of the likelihood is possible but leads to long formulas. We are currently exploring the possibility of a matrix analytic computation.

5.3. Likelihood based statistical inference

The multivariate probability distributions obtained in the above subsections can be used for statistical inference. As our first illustrative example, we consider a standard coalescent model for a sample of size three. Suppose for instance that $(m_a, m_b, m_c, m_{ab}) = (0, 1, 1, 3)$. Then Eq. (44) and its extension to the labeled coalescent may be used to obtain the likelihood function. Fig. 8 shows the maximum likelihood estimator (MLE) using our formula (44), as well as the Ewens–Watterson estimator proposed both by Ewens (1974) and Watterson (1975). They are clearly different.

We next look at the situation where there are $l = 20$ independent loci each with a sample of size three, and θ is estimated from their combined information. Fig. 9(a) displays the log-likelihood (summed across all loci) and shows that the MLE is more accurate than the Ewens–Watterson estimator.

In Fig. 9(b), we consider one locus, and a sample of size $n = 20$. Although there are methods such as *genetree* (see Bahlo and Griffiths (2000)) available that approximate the full likelihood for a coalescent with twenty lineages, we rely on our formula for three lineages and use a composite likelihood approach. For this purpose, we took 1000 subsamples from the original sample, each of size three without replacement. The likelihood is then computed pretending that these subsamples are independent. With 20 starting lineages, it would be easily possible to consider all $\binom{20}{3} = 1140$ possible configurations. We relied on subsampling, however, to illustrate that the approach can be easily generalized to any number of lineages.

To show that our illustrations indeed represent typical cases, we look at the average behavior over 100 simulation runs for the three scenarios considered above. Table 3 provides the root mean squared errors, i.e. the square root of the average squared differences between the estimates and the true parameter value. As expected from coalescent theory, data from 20 independent loci lead to the most accurate

$$T = \begin{pmatrix} -1-3\mu & \mu & \mu & \mu & 1 & 0 & 0 & 0 & 0 & 0 \\ \mu & -3-3\mu & \mu & \mu & 0 & 1 & 0 & 0 & 1 & 1 \\ \mu & \mu & -1-3\mu & \mu & 0 & 0 & 1 & 0 & 0 & 0 \\ \mu & \mu & \mu & -1-3\mu & 0 & 0 & 0 & 1 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & -2\mu & 2\mu & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2\mu & -1-2\mu & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -2\mu & 0 & 0 & 2\mu \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -2\mu & 2\mu & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2\mu & -1-2\mu & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2\mu & 0 & 0 & -1-2\mu \end{pmatrix} \quad (45)$$

Box I.

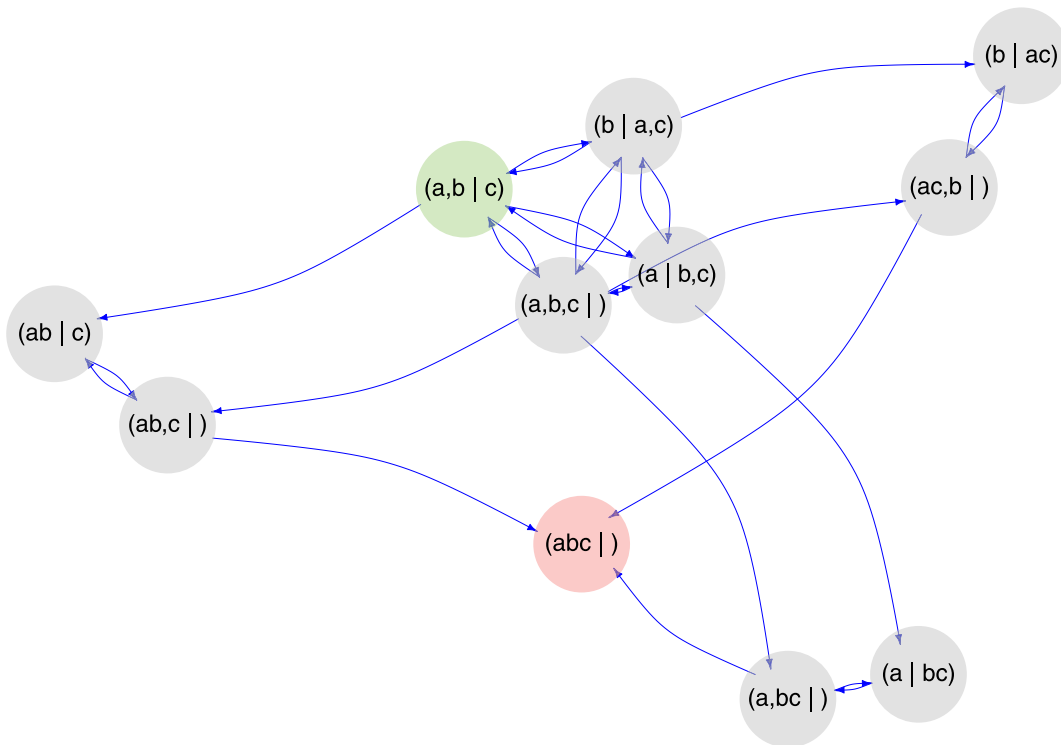


Fig. 7. Adjacency graph for the three sample model with migration. The states correspond to those specified in Table 2. The initial and the absorbing state are colored in green and red, respectively. Possible transitions, either due to migration or coalescence events, are represented by arrows. The transition rates can be found in the sub-transition matrix T . PhaseTypeR was used to create the graph, and the R code can be found in the accompanying script.

Table 3

Root mean squared error of the M(C)LE (maximum (composite) likelihood estimate) versus Ewens–Watterson estimate from 100 simulation runs. Scenarios: (1) MLE for sample of size $n = 3$ from $l = 1$ locus; (2) sample of size $n = 20$ from $l = 1$ locus, composite likelihood estimates based on sub-samples of size 3 drawn from 20 genes are used; (3) MLE from samples of size $n = 3$ from $l = 20$ independent loci.

	M(C)LE	Ewens–Watterson
(1) $n = 3, l = 1$	1.972	2.733
(2) $n = 20, l = 1$	1.569	1.898
(3) $n = 3, l = 20$	0.536	0.541

estimates. With one locus, 20 lineages lead to roughly 20% more accurate estimates compared to three lineages.

While Lohse et al. (2011) and Lohse et al. (2016) also consider a composite likelihood approach, their focus is on multiple linked loci and not multiple subsamples at one single locus. It might therefore be interesting to explore our proposed composite likelihood approach further, for instance in terms of uncertainty quantification.

Table 4

Sample from model with migration $M = 4$ and mutation rate $\theta = 3$. Mutation counts per lineage for three genes and two independent loci.

	m_a	m_b	m_c	m_{ab}	m_{ac}	m_{bc}
l_1	0	1	6	10	0	0
l_2	0	2	2	0	2	0

We next provide an example for the model described in Section 5.2 involving both mutations and migration. For this purpose, we used *ms* (Hudson (2002)) to simulate two independent loci for a coalescent that starts with three sequences. We took $\theta = 3$ as our scaled mutation rate, and the migration rate $M = 4$. Notice that *ms* uses the scaled migration rate $M = 4N\bar{m}$, with N being the population size, and \bar{m} the per generation migration rate. It is related via $M = 2\mu$ to our migration parameter. The resulting mutational pattern can be found in Table 4.

Using the generating function derived in Section 5.2, we computed the partial derivatives that correspond to the mutational patterns

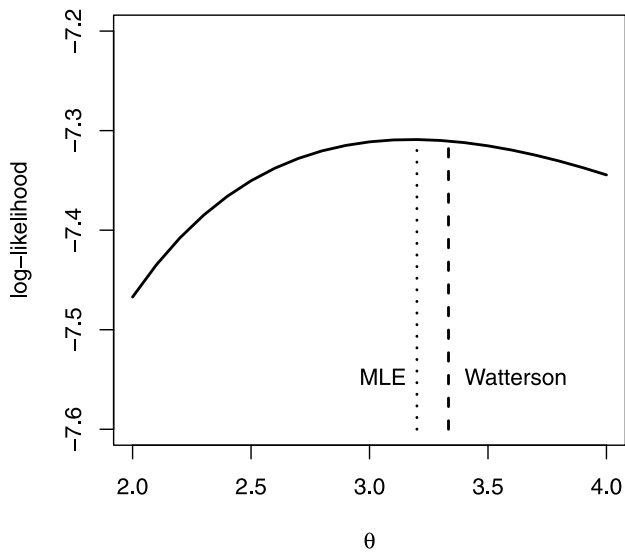


Fig. 8. Likelihood function for the mutational pattern $(m_a, m_b, m_c, m_{ab}) = (0, 1, 1, 3)$ and a sample of size $n = 3$. The maximum likelihood estimate (MLE) as well as the Ewens–Watterson estimate are displayed.

displayed in Table 4. More specifically, we took the derivatives of order 1, 6, and 10 with relation to z_2 , z_3 , and z_4 for locus l_1 , and similarly, the required derivatives for l_2 . This was done symbolically with *Mathematica*. The likelihood is then obtained by evaluating both derivatives at $z_1 = z_2 = \dots = z_6 = 0$ and taking the product. For the likelihood computations, the normalizing constants $(m_1!m_2! \dots m_6!)^{-1}$ may be omitted, as they do not affect the MLE.

Taking the logarithm gives us the log-likelihood. The expression for the likelihood function becomes rather long and can be found in our GitHub repository. Despite its length, however, we still obtained an explicit formula in terms of a function of the unknown parameters. Fig. 10 provides a plot of this composite likelihood function. We observe that the likelihood is much flatter in direction M than in direction θ . Thus, the data provide more information for estimating the mutation rate θ than the migration rate M . We also computed $\hat{M} = 4.50$, and $\hat{\theta} = 3.84$ as the maximum likelihood estimates for our example.

6. Discussion: Extensions and further perspectives

We have demonstrated that phase-type theory provides a general framework to derive basic properties of coalescent trees. Whenever the quantities of interest can be interpreted as (possibly reward weighted) absorption times of a discrete or continuous time Markov chain, the theory provides matrix analytic formulas for computing properties of their distribution. In Section 2 we provided several examples.

In Section 3, we illustrated that joint distributions and their corresponding mixed moments can be obtained by specifying multiple reward patterns that represent quantities of interest such as tree height and length, as well as the external and internal branch lengths. Going beyond the standard coalescent model, phase-type results can be obtained in an analogous way for multiple merger models such as the Λ -coalescent.

Section 4 explained how discrete phase-type distributions can be used to obtain properties of the number of segregating sites, the site frequency spectrum, and coalescence times under consanguineous mating. In a further classical population genetic application, namely fixation or loss in a discrete-time Wright–Fisher model with selection, we demonstrated how to treat a situation with more than one absorbing state.

Distributions and moments of summary statistics, such as the site frequency spectrum, can also be used for statistical inference. In Section 5, we explained how to use phase-type distributions to obtain full likelihoods for small size coalescent trees. Using a composite likelihood approach, we show how these results may be applied to carry out statistical inference under a simple demographic model.

All our applications and examples illustrate that phase-type theory provides a general set of tools that can be applied to a large variety of coalescent models in population genetics.

We end this paper with a short discussion on important extensions of the homogeneous coalescent models, challenges with large sample sizes, and further population genetic models where the phase-type framework seems particularly promising.

6.1. Extensions of the homogeneous coalescent model

Phase-type distributions in population genetics are still in their infancy. In this review, we have focused on time homogeneous coalescent models, but several important extensions are natural to consider. A crucial extension is to consider inhomogeneous coalescent models and the corresponding inhomogeneous phase-type distribution (Albrecher and Bladt, 2019). This model corresponds to a coalescent model with variability in population size (e.g. Wooding and Rogers (2002) and Polanski and Kimmel (2002)). Zeng et al. (2021) discretized time into homogeneous blocks of time with constant population size to analyze genetic diversity for balancing selection. Reward-transformation in inhomogeneous phase-type distributions still needs to be developed.

6.2. Large sample sizes and computational running time

Large sample sizes pose a challenge for the phase-type methodology because the state space increases very fast with the sample size (recall Section 2.4). In Fig. 11 we show the running time of PhaseTypeR for calculating the mean tree height (left) and variance–covariance matrix of the site frequency spectrum (right) for increasing sample size. In our implementation, the calculation of both quantities is cubic in the size of the state space, and the calculation of the variance–covariance matrix can take up to half a minute when the size of the state space is larger than 250.

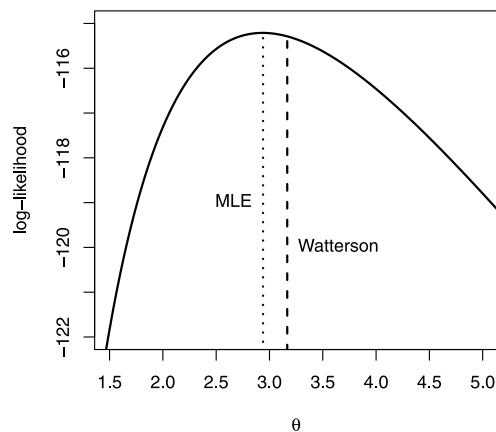
A possible strategy to handle a large number of states is to exploit that the rate matrices are very sparse and have a block structure. Røijker et al. (2022) and Bisschop (2022) represent the rate matrices as graphs and carry out the desired matrix operations on the graph. Røijker et al. (2022) is mainly focused on calculating the moments, and Bisschop (2022) is mainly focused on calculating the likelihood. A natural next step would be to combine the two graph-based approaches in a single software program.

It could be important to have limit theorems for the multivariate phase-type distribution. Limit theorems for the site frequency spectrum are available for the standard coalescent (Dahmer and Kersting, 2015), but a general framework for reward-transformed statistics of homogeneous coalescent models is missing.

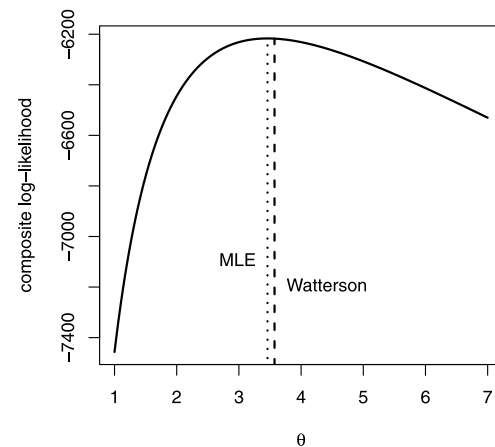
6.3. Discriminating between coalescent models

An application where phase-type theory could perhaps be rather easily applied is in the context of summary statistics for discriminating between coalescent models. Koskela (2018) suggest using summary statistics consisting of the normalized number of singletons and the cumulative number of i -tons with i larger than or equal to 15. Koskela (2018) in his Figure 1 clearly shows that these joint statistics are able to discriminate between the Beta(1 – α , α) – Ξ -coalescent and exponential or algebraic growth for certain values of sample size, number of loci, size of loci, mutation rates, the value of α and growth parameter.

The choice of singletons and cumulative tail probabilities of the site frequency spectrum as summary statistics discriminating between the



(a) Combined log-likelihood computed from 20 independent loci.



(b) Composite log-likelihood for one locus and a sample size $n = 20$.

Fig. 9. Likelihood functions computed from simulated data. Both the maximum composite likelihood estimate (MCLE) and Ewens–Watterson estimate are displayed. We show typical cases where the MLE is closer to the true parameter value $\theta = 3$.

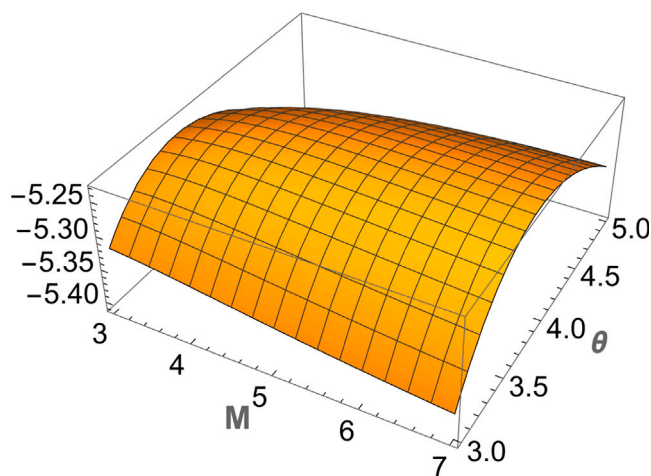


Fig. 10. Log-Likelihood function for two independent loci observed under a two-island model with symmetrical scaled migration rate $M = 4$ and scaled mutation parameter $\theta = 3$. The mutational pattern underlying the likelihood is provided in Table 4. It has been obtained from simulations using *ms*.

models could potentially be improved by using the full site frequency spectrum together with an analytical procedure for determining the means and variances for the models. A further alternative in this setting is the model selection procedure in Freund and Siri-Jégousse (2021) where various growth- and multiple merger coalescent models are simulated in order to train a random forest to distinguish between the genealogical models.

6.4. Phase-type theory for the infinite alleles model

In this review we have focused on the infinite sites model where the number of mutations of different types (e.g. singletons, doubletons, etc.) only depend on the total branch length that can give rise to the mutation type (e.g. total singleton branch length, total doubleton branch length, etc.). The infinite alleles model is another key model in coalescent theory (see e.g. Section 4.2 (Wakeley, 2008)), and it could be interesting to formulate the infinite alleles model in the phase-type framework. In the infinite alleles model, a sample of homologous DNA sequences is summarized by the haplotype frequency vector. Innan

et al. (2005) add the number of segregating sites to the haplotype frequency vector and derive a recursion for the joint probability of the two statistics. It could be intriguing to also understand this joint summary statistics in a phase-type context. In Griffiths and Tavaré (2018), the authors consider the joint distribution of both the number of segregating sites and the number and frequencies of haplotypes for the purpose of inference. These results may be viewed as an extension of Ewens's sampling formula. It would be interesting to explore whether their Markov chain representations (19) and (20) can be used to obtain discrete phase-type distributions for the above-mentioned frequencies.

CRediT authorship contribution statement

Asger Hobolth: Conceptualization, Formal analysis, Methodology, Project administration, Writing – original draft, Writing – review & editing. **Iker Rivas-González:** Conceptualization, Formal analysis, Methodology, Software, Writing – original draft, Writing – review & editing. **Mogens Bladt:** Methodology, Writing – original draft, Writing – review & editing. **Andreas Futschik:** Conceptualization, Formal analysis, Methodology, Project administration, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

We thank Janek Sendrowski for useful comments and suggestions on a previous version of this manuscript. We are grateful to the three reviewers and the editor Noah Rosenberg for the careful reading of our manuscript and many constructive and helpful comments, questions and suggestions. We thank Aarhus University Research Foundation for supporting two visits of Andreas Futschik to Aarhus University.

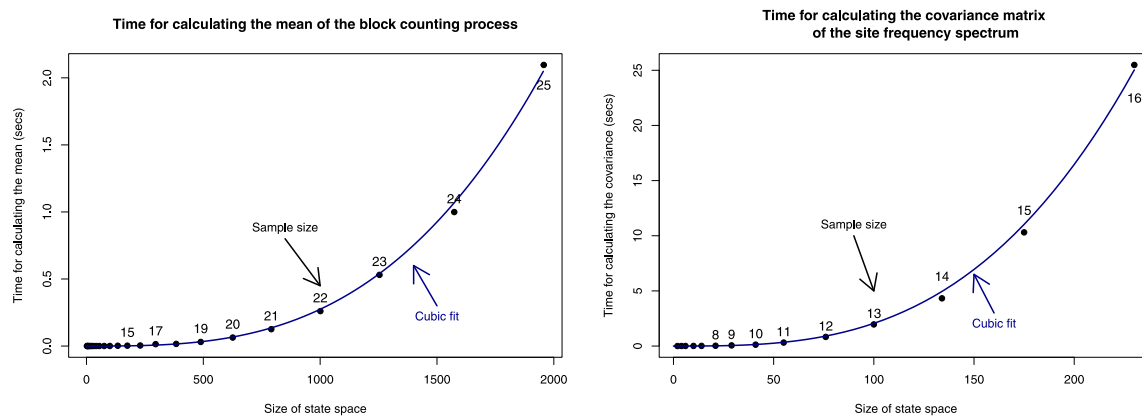


Fig. 11. Left: Running time of PhaseTypeR for calculating the mean tree height of the standard coalescent from the block counting process. Right: Running time of PhaseTypeR for calculating the variance-covariance matrix of the site frequency spectrum from the block counting process.

References

- Albrecher, Hansjörg, Bladt, Mogens, 2019. Inhomogeneous phase-type distributions and heavy tails. *J. Appl. Probabil.* 56 (4), 1044–1064.
- Albrecher, Hansjörg, Bladt, Mogens, Yslas, Jorge, 2022. Fitting inhomogeneous phase-type distributions to data: the univariate and the multivariate case. *Scandinavian J. Stat.* 49 (1), 44–77.
- Alimpiev, Egor, Rosenberg, Noah A., 2022. A compendium of covariances and correlation coefficients of coalescent tree properties. *Theor. Popul. Biol.* 143, 1–13.
- Andersen, Allan T., Neuts, Marcel F., Nielsen, Bo Friis, 2000. PH-distributions arising through conditioning. *Stoch. Models* 16 (1), 179–188.
- Arnold, Taylor, Kane, Michael, Lewis, Bryan W., 2019. *A Computational Approach to Statistical Learning*. Chapman and Hall.
- Asmussen, Søren, Albrecher, Hansjörg, 2010. *Ruin probabilities*, vol. 14, World Scientific.
- Asmussen, Søren, Nerman, Olle, Olsson, Marita, 1996. Fitting phase-type distributions via the EM algorithm. *Scand. J. Stat.* 23, 419–441.
- Bahlo, Melanie, Griffiths, Robert C., 2000. Inference from gene trees in a subdivided population. *Theor. Popul. Biol.* 57 (2), 79–95.
- Baumdicker, Franz, Bisschop, Gertjan, Goldstein, Daniel, Gower, Graham, Ragsdale, Aaron P., Tsambos, Georgia, Zhu, Sha, Eldon, Bjarki, Ellerman, E. Castedo, Galloway, Jared G., et al., 2022. Efficient ancestry and mutation simulation with msprime 1.0. *Genetics* 220 (3), iyab229.
- Bijma, Fetsje, Jonker, Marianne, Van Der Vaart, Add, 2017. *An Introduction to Mathematical Statistics*. Amsterdam University Press.
- Birkner, Mathias, Blath, Jochen, 2021. Genealogies and inference for populations with highly skewed offspring distributions. *Probabil. Struct. Evol.* Chapter 8.
- Bisschop, Gertjan, 2022. Graph-based algorithms for Laplace transformed coalescence time distributions. *PLoS Comput. Biol.* 18 (9), e1010532.
- Bladt, Mogens, Gonzalez, Antonio, Lauritzen, Steffen L., 2003. The estimation of phase-type related functionals using Markov chain Monte Carlo methods. *Scand. Actuar. J.* 2003 (4), 280–300.
- Bladt, Mogens, Nielsen, Bo Friis, 2017. Matrix-exponential distributions in applied probability. In: *Matrix-exponential Distributions in Applied Probability*, Springer.
- Bladt, Martin, Yslas, Jorge, 2021. *Matrixdist: an r package for inhomogeneous phase-type distributions*. arXiv preprint arXiv:2101.07987.
- Blath, Jochen, Buzzoni, Eugenio, Koskela, Jere, Berenguer, Maite Wilke, 2020. Statistical tools for seed bank detection. *Theor. Popul. Biol.* 132, 1–15.
- Blath, Jochen, Casanova, Adrián González, Kurt, Noemi, Wilke-Berenguer, Maite, 2016. A new coalescent for seed-bank models. *Ann. Appl. Probab.* 26 (2), 857–891.
- Blum, Michael G.B., Rosenberg, Noah A., 2007. Estimating the number of ancestral lineages using a maximum-likelihood method based on rejection sampling. *Genetics* 176 (3), 1741–1757.
- Brockmeyer, E., Halstrøm, H.L., Jensen, A., 1948. *The life and works of AK Erlang*. The Academy of Danish Sciences.
- Campbell, R.B., 2015. The effect of inbreeding constraints and offspring distribution on time to the most recent common ancestor. *J. Theoret. Biol.* 382, 74–80.
- Campillo Navarro, Azucena, 2018. *Order statistics and multivariate discrete phase-type distributions*. (Ph.D. thesis). Technical University of Denmark (Copenhagen, Denmark). Department of Applied Mathematics and Computer Science.
- Costa, Rui J., Wilkinson-Herbots, Hilde, 2017. Inference of gene flow in the process of speciation: An efficient maximum-likelihood method for the isolation-with-initial-migration model. *Genetics* 205 (4), 1597–1618.
- Costa, Rui J., Wilkinson-Herbots, Hilde M., 2021. Inference of gene flow in the process of speciation: efficient maximum-likelihood implementation of a generalised isolation-with-migration model. *Theor. Popul. Biol.* 140, 1–15.
- Dahmer, Iulia, Kersting, Götz, 2015. The internal branch lengths of the kingman coalescent. *Ann. Appl. Probab.* 25 (3), 1325–1348.
- Eldon, Bjarki, Wakeley, John, 2006. Coalescent processes when the distribution of offspring number among individuals is highly skewed. *Genetics* 172 (4), 2621–2633.
- Etheridge, Alison, 2011. *Some Mathematical Models from Population Genetics: École D'été de Probabilités de Saint-Flour XXXIX-2009*, vol. 2012, Springer Science & Business Media.
- Ewens, W.J., 1974. A note on the sampling theory for infinite alleles and infinite sites models. *Theor. Popul. Biol.* 6 (2), 143–148.
- Fay, Justin C., Wu, Chung-I, 2000. Hitchhiking under positive darwinian selection. *Genetics* 155 (3), 1405–1413.
- Freund, Fabian, Siri-Jégousse, Arno, 2021. The impact of genetic diversity statistics on model selection between coalescents. *Comput. Stat. Data Anal.* 156, 107055.
- Gardner, Clara Brimnes, Nielsen, Sara Dorthea, Eltvéd, Morten, Rasmussen, Thomas Kjær, Nielsen, Otto Anker, Nielsen, Bo Friis, 2021. Calculating conditional passenger travel time distributions in mixed schedule-and frequency-based public transport networks using Markov chains. *Transp. Res. B* 152, 1–17.
- Goulet, Vincent, Dutang, Christophe, Maechler, Martin, Firth, David, Shapira, Marina, Stadelmann, Michael, 2021. *Expm: Matrix exponential, log, 'etc'*. URL <https://CRAN.R-project.org/package=expm>, R package version 0.999-6.
- Griffiths, Robert C., 1989. Genealogical-tree probabilities in the infinitely-many-site model. *J. Math. Biol.* 27, 667–680.
- Griffiths, Robert C., Tavaré, Simon, 1994. Ancestral Inference in Population Genetics. *Statist. Sci.* 9 (3), 307–319.
- Griffiths, Robert C., Tavaré, Simon, 2018. Ancestral inference from haplotypes and mutations. *Theor. Popul. Biol.* 122, 12–21.
- Hobolth, Asger, Bladt, Mogens, Andersen, L.N., 2021. Multivariate phase-type theory for the site frequency spectrum. *J. Math. Biol.* 83, 1–28.
- Hobolth, Asger, Siri-Jégousse, Arno, Bladt, Mogens, 2019. Phase-type distributions in population genetics. *Theor. Popul. Biol.* 127, 16–32.
- Hudson, Richard R., 2002. Generating samples under a wright-Fisher neutral model of genetic variation. *Bioinformatics* 18 (2), 337–338.
- Hurtado, Paul J., Richards, Cameron, 2021. Building mean field ODE models using the generalized linear chain trick & Markov chain theory. *J. Biol. Dyn.* 15 (sup1), S248–S272.
- Ibe, Oliver C., 2013. *Markov Processes for Stochastic Modeling*, Second ed. Elsevier.
- Innan, H., Zhang, K., Marjoram, P., Tavaré, S., Rosenberg, N.A., 2005. Statistical tests of the coalescent model based on the haplotype frequency distribution and the number of segregating sites. *Genetics* 169, 1763–1777.
- Jensen, Arne, 1953. Markoff chains as an aid in the study of markoff processes. *Skandinavisk Aktuarietidskrift* 36, 87–91.
- Kersting, Götz, Siri-Jégousse, Arno, Wences, Alejandro H., 2021. Site frequency spectrum of the bolthausen-sznitman coalescent. *ALEA, Lat. Am. J. Probab. Math. Stat* 18, 1483–1505.
- Kingman, John F.C., 2009. The first erlang century — and the next. *Queueing Syst.* 63, 3–12.
- Koskela, Jere, 2018. Multi-locus data distinguishes between population growth and multiple merger coalescents. *Stat. Appl. Genet. Molec. Biol.* 17, (3).
- Krukov, Ivan, de Sanctis, Bianca, de Koning, A. P. Jason, 2016. Wright-Fisher exact solver (WFES): scalable analysis of population genetic models without simulation or diffusion theory. *Bioinformatics* 33 (9), 1416–1417.
- Kulkarni, Vidyadhar G., 1989. A new class of multivariate phase type distributions. *Oper. Res.* 37 (1), 151–158.
- Lambert, Amaury, Ma, Chunhua, 2015. The coalescent in peripatric metapopulations. *J. Appl. Probab.* 52 (2), 538–557.

- Lohse, Konrad, Chmelik, Martin, Martin, Simon H., Barton, Nicholas H., 2016. Efficient strategies for calculating blockwise likelihoods under the coalescent. *Genetics* 202 (2), 775–786.
- Lohse, Konrad, Harrison, R.J., Barton, Nick H., 2011. A general method for calculating likelihoods under the coalescent process. *Genetics* 188, 977–987.
- Moler, Cleve, Van Loan, Charles, 2003. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM Rev.* 45 (1), 3–49.
- Neuts, Marcel F., 1975. Probability distributions of phase type. pp. 173–206, *Liber Amicorum Prof. Emeritus H. Florin*.
- Pitman, Jim, 1999. Coalescents with multiple collisions. *Ann. Probab.* 1870–1902.
- Polanski, A., Kimmel, M., 2002. New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. *Genetics* 165, 427–436.
- Rivas-González, Iker, Andersen, Lars Nørvang, Hobolth, Asger, 2023. PhaseTypeR: an R package for phase-type distributions in population genetics. *J. Open Source Softw.* 8 (82), 5054.
- Røikjer, Tobias, Hobolth, Asger, Munch, Kasper, 2022. Graph-based algorithms for phase-type distributions. *Stat. Comput.* 32 (6), 103.
- Rosenberg, Noah A., 2020. Fifty years of theoretical population biology. *Theor. Population Biol.* 133, 1–12.
- Sagitov, Serik, 1999. The general coalescent with asynchronous mergers of ancestral lines. *J. Appl. Probab.* 36 (4), 1116–1125.
- Schrider, Daniel R., Kern, Andrew D., 2018. Supervised machine learning for population genetics: A new paradigm. *Trends Genet.* 34 (4), 301–312.
- Schweinsberg, Jason, 2003. Coalescent processes obtained from supercritical galton–watson processes. *Stoch. Processes Appl.* 106 (1), 107–139.
- Severson, Alissa L., Carmi, Shai, Rosenberg, Noah A., 2019. The effect of consanguinity on between-individual identity-by-descent sharing. *Genetics* 212 (1), 305–316.
- Severson, Alissa L., Carmi, Shai, Rosenberg, Noah A., 2021. Variance and limiting distribution of coalescence times in a diploid model of a consanguineous population. *Theor. Popul. Biol.* 139, 50–65.
- Tavaré, Simon, 2004. Ancestral inference in population genetics. In: *Lectures on Probability Theory and Statistics*. Springer, pp. 1–188.
- Uyenoyama, Marcy K., Takebayashi, Naoki, Kumagai, Seiji, 2019. Inductive determination of allele frequency spectrum probabilities in structured populations. *Theor. Popul. Biol.* 129, 148–159, Special issue in honor of Marcus Feldman's 75th birthday.
- Uyenoyama, Marcy K., Takebayashi, Naoki, Kumagai, Seiji, 2020. Allele frequency spectra in structured populations: Novel-allele probabilities under the labelled coalescent. *Theor. Popul. Biol.* 133, 130–140.
- Wakeley, John, 2008. *Coalescent Theory: An Introduction*. Roberts & Company Publishers.
- Watterson, G.A., 1975. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* 7 (2), 256–276.
- Wooding, Stephen, Rogers, Alan, 2002. The matrix coalescent and an application to human single-nucleotide polymorphisms. *Genetics* 161, 1641–1650.
- Zeng, K., Charlesworth, B., Hobolth, A., 2021. Studying models of balancing selection using phase-type theory. *Genetics* 218.