


AUTHOR QUERY FORM

	Journal: Theoretical Population Biology Article Number: 2282	Please e-mail or fax your responses and any corrections to: E-mail: corrections.essd@elsevier.river-valley.com Fax: +44 1392 285879
---	--	--

Dear Author,

Please check your proof carefully and mark all corrections at the appropriate place in the proof (e.g., by using on-screen annotation in the PDF file) or compile them in a separate list. Note: if you opt to annotate the file with software other than Adobe Reader then please also highlight the appropriate place in the PDF file. To ensure fast publication of your paper please return your corrections within 48 hours.

For correction or revision of any artwork, please consult <http://www.elsevier.com/artworkinstructions>.

Location in article	Query / Remark click on the Q link to go Please insert your reply or correction at the corresponding line in the proof
Q1	Please confirm that given names and surnames have been identified correctly.
	<div style="border: 1px solid black; padding: 5px; display: inline-block;"> Please check this box if you have no corrections to make to the PDF file <input type="checkbox"/> </div>

Thank you for your assistance.



Contents lists available at SciVerse ScienceDirect

Theoretical Population Biology

journal homepage: www.elsevier.com/locate/tpb

Multiple merger gene genealogies in two species: Monophyly, paraphyly, and polyphyly for two examples of Λ coalescents

 **Marki Eldon**^{a,*}, **James H. Degnan**^b

^a Department of Statistics, University of Oxford, One South Parks Road, Oxford OX1 3TG, UK

^b Department of Mathematics and Statistics, University of Canterbury, Christchurch, Private Bag 4800, New Zealand

ARTICLE INFO

Article history:

Received 22 February 2012

Available online xxxxx

Keywords:

Gene genealogies

Species tree

Taxonomic distinctiveness

Multiple merger coalescent processes

Monophyly

ABSTRACT

Probabilities of monophyly, paraphyly, and polyphyly of two-species gene genealogies are computed for modest sample sizes and compared for two different Λ coalescent processes. Coalescent processes belonging to the Λ coalescent family admit asynchronous multiple mergers of active ancestral lineages. Assigning a timescale to the time of divergence becomes a central issue when different populations have different coalescent processes running on different timescales. Clade probabilities in single populations are also computed, which can be useful for testing for taxonomic distinctiveness of an observed set of monophyletic lineages. The coalescence rates of multiple merger coalescent processes are functions of coalescent parameters. The effect of coalescent parameters on the probabilities studied depends on the coalescent process, and if the population is ancestral or derived. The probability of reciprocal monophyly tends to be somewhat lower, when associated with a Λ coalescent, under the null hypothesis that two groups come from the same population. However, even for fairly recent divergence times, the probability of monophyly tends to be higher as a function of the number of generations for coalescent processes that admit multiple mergers, and is sensitive to the parameter of one of the example processes.

© 2012 Published by Elsevier Inc.

1. Introduction

The coalescent (Kingman, 1982a,c) has proved to be a very useful tool for inference in population genetics (Hudson, 1990; Donnelly and Tavaré, 1995; Möhle, 2000; Nordborg, 2001; Rosenberg and Nordborg, 2002; Wakeley, 2009), and phylogenetics (Satta et al., 2000; Ting et al., 2000; Liu and Pearl, 2007; Degnan and Rosenberg, 2009; Liu et al., 2009). Genetic information drawn from a set of taxa may not yield unequivocal resolution of the corresponding species tree (Hudson, 1983; Nei, 1986; Neigel and Avise, 1986; Doyle, 1992; Ruvolo, 1994; Maddison, 1997; Nichols, 2001; Nordborg, 2001). To understand why, it is helpful to think about the gene genealogies of the sampled DNA sequences embedded within the phylogeny of the species in question (Fig. 1). Reciprocal monophyly (Rosenberg, 2003) is illustrated in Fig. 1a; the lineages from both populations (A and B) reach their respective most recent common ancestor (MRCA) before any coalescence event involving lineages from both A and B occurs. Polyphyly (Rosenberg, 2003) is illustrated in Fig. 1c, d, in which lineages from A and B coalesce more quickly than a MRCA is reached in either population. Although

phenomena such as recombination, sampling error (Cummings et al., 1995; Otto et al., 1996), and gene duplication can result in gene genealogies being discordant with species trees, lack of reciprocal monophyly is often expected to be widespread for closely related populations (Knowles and Carstens, 2007), whose divergence occurred recently. The proportion of genes which are reciprocally monophyletic or paraphyletic is a reflection of the time of divergence between populations.

Probabilities of monophyly, paraphyly, and polyphyly under the Kingman coalescent model (Kingman, 1982a,c,b) have been the focus of previous work (Hudson and Coyne, 2002; Rosenberg, 2003). The Kingman coalescent can be derived from the usual Fisher–Wright (Fisher, 1930; Wright, 1931) and the Moran (1958, 1962) population models, which can be classified as low offspring number models. In low offspring number models, individuals have very many offspring with only negligible probability in large populations. Indeed, convergence to the Kingman coalescent follows from conditions on higher moments of Cannings (1974) reproduction law, of which the Fisher–Wright and the Moran models are special cases. Large offspring number models (Schweinsberg, 2003; Eldon and Wakeley, 2006; Sargsyan and Wakeley, 2008) in which individuals can have very many offspring – up to the order of the population size – with non-negligible probability in large populations, give rise to multiple merger coalescent processes (Donnelly and Kurtz, 1999; Pitman, 1999; Sagitov, 1999; Schweinsberg, 2000a; Möhle and Sagitov, 2001). In multiple merger coalescent

* Corresponding author.

E-mail addresses: eldon@stats.ox.ac.uk, beldon11@gmail.com (B. Eldon).

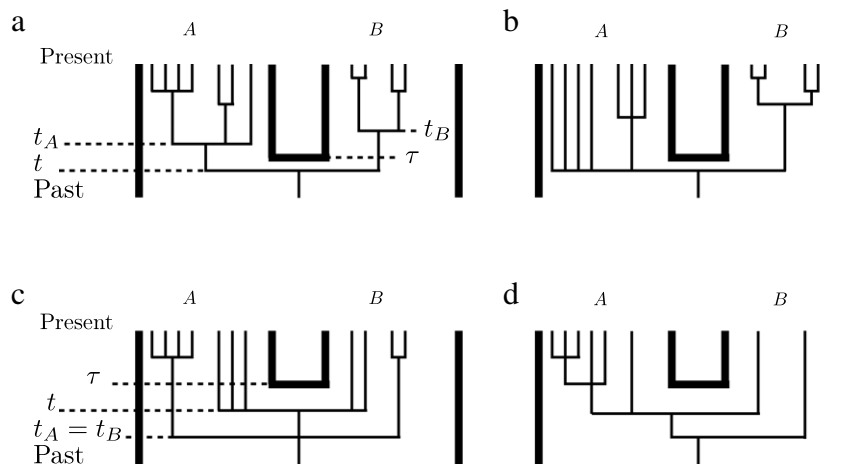


Fig. 1. Examples of gene genealogies relating genetic information drawn from two populations *A* and *B* diverged at time τ . Thick lines demarcate the populations, while thin lines trace ancestral lineages from the present into the past. The gene genealogies display monophyly of *A* and *B* (a); parphyly of *A* relative to *B* (b); polyphyly (c) and (d). Population *A* and the common ancestral population both admit multiple mergers of ancestral lineages. The time of the MRCA of *A* and *B* is denoted by t_A and t_B , respectively; t denotes the first time at which lineages from *A* and *B* coalesce.

processes, any number of active ancestral lineages can coalesce to a different common ancestral sequence at the same time (Λ coalescent). Simultaneous multiple merger coalescent processes (Schweinsberg, 2000a) admit the coalescence of different groups of active ancestral lineages to different ancestors at the same time (Ξ coalescent). Fig. 1 carries examples of gene genealogies with multiple mergers, in population *A* and the common ancestral population. In Fig. 1a, for example, the first merger in population *A* is a merger of four lineages, followed by a merger of two lineages. The three remaining ancestral lineages finally reach the most recent common ancestor of the *A* lineages in a single merger.

Large offspring number models may be better approximations for highly fecund organisms than the usual low offspring number Fisher–Wright and Moran models. Schweinsberg (2003) considers a population model in discrete generations in which the distribution of potential offspring has heavy tails. The population is then regenerated by sampling from the pool of potential offspring. If the tails are ‘heavy enough’, the resulting coalescent process admits multiple mergers of ancestral lineages. Eldon and Wakeley (2006) and Sargsyan and Wakeley (2008) consider large offspring number models and predictions about genetic diversity, and argue that large offspring number models may be appropriate for highly fecund marine organisms such as Pacific oysters. Indeed, sweepstake-style reproduction, in which few parents have very many offspring, was proposed by Beckenbach (1994), Hedgecock et al. (1982), and Hedgecock (1994) when considering data on Pacific oysters. Árnason (2004) raises similar ideas in relation to Atlantic cod. Genetic evidence for large offspring numbers in different marine taxa continues to be subject to investigation (Boudry et al., 2002; Flowers et al., 2002; Petersen et al., 2008). Ingvarsson (2010) proposes that large offspring number models may be appropriate for forest trees.

We compute probabilities of monophyly, parphyly, and polyphyly, of two-species gene genealogies (Fig. 1) when the coalescent process in any of the three populations (*A*, *B*, or the common ancestral population) are special cases of the Λ coalescent; i.e., admitting asynchronous multiple mergers of ancestral lineages. The approach is similar to the one taken by Hudson and Coyne (2002); conditioning on the number of lineages that have coalesced more recently than the species divergence, and using recursions for the ancestral population.

The simple techniques we employ can be applied to any multiple merger coalescent. Special cases of the Λ coalescent derived by Schweinsberg (2003) and Eldon and Wakeley (2006) are

considered in detail. The presence of multiple mergers limits the computation of exact probabilities to modest, but very relevant, sample sizes. By way of example, Waters and Roy (2004) study the phylogeography of a New Zealand sea star by sampling, on average, five specimens of *Patiriella regularis* at different locations around the coast of New Zealand. As different populations can have different coalescent processes running on different timescales, assigning a timescale to the time of divergence becomes a central issue in the computations. In addition, we consider probabilities that a subset of lineages form a clade in the case of a single population. Clade probabilities are useful for determining whether observed levels of monophyly can be considered statistically significant when treating a single population as a null hypothesis (Rosenberg, 2007).

2. Theory and results

2.1. One population

The two special cases of a Λ coalescent we will consider are the ones introduced by Schweinsberg (2003) and Eldon and Wakeley (2006). At each timestep in Schweinsberg (2003)’s model, individual i independently produces a random number X_i of potential offspring; N offspring are then drawn without replacement from the pool of the potential offspring. The assumption $\mathbb{E}[X_i] > 1$ assures that $X_1 + \dots + X_N \geq N$ with sufficiently high probability (Schweinsberg, 2003). The X_i are independent and identically distributed with tail probabilities

$$\mathbb{P}[X_i \geq k] \sim Ck^{-\alpha}, \quad k > 0$$

in which \sim means that the ratio of the two sides tends to 1 as $k \rightarrow \infty$, and C is a constant. The usual Kingman coalescent is obtained when $\alpha \geq 2$. In the Kingman coalescent, each pair of active ancestral lineages coalesces with rate 1. When $1 \leq \alpha < 2$, the coalescent process is a Λ coalescent, and the time during which there are i active ancestral lineages is exponential with rate $\sum_{k=2}^i \lambda_{i,k}$, in which

$$\lambda_{i,k} = \binom{i}{k} \frac{\Gamma(k-\alpha)\Gamma(i-k+\alpha)}{\Gamma(i)\Gamma(2-\alpha)\Gamma(\alpha)}, \quad 1 < \alpha < 2. \quad (1)$$

Schweinsberg (2003). The quantity $\lambda_{i,k}$ is the rate at which k out of i active ancestral lineages coalesce, and is referred to as the coalescence rate. Refer to the coalescent with rate (1) as the Beta-coalescent. Eldon and Wakeley (2006) consider a discrete-time

modified Moran-type model in which a single individual chosen uniformly at random from the population contributes a random number U of offspring at each timestep, and persists. Eldon and Wakeley (2006) give U the distribution

$$\mathbb{P}[U = u] = \begin{cases} 1 - N^{-\gamma} & \text{if } u = 1, \gamma > 0 \\ N^{-\gamma} & \text{if } u = \lfloor \psi N \rfloor, 0 < \psi < 1, \end{cases}$$

in which γ and ψ are constants. A simple case of a Λ coalescent results when $0 < \gamma < 2$, in which the coalescence rate is

$$\lambda_{i,k} = \binom{i}{k} \psi^k (1 - \psi)^{i-k}, \quad 2 \leq k \leq i, 0 < \psi < 1. \quad (2)$$

Refer to the coalescent with rate (2) as the ψ -coalescent. The Kingman coalescent is recovered from the Beta coalescent by taking $\alpha = 2$. The Kingman coalescent is not recovered from the ψ -coalescent. To obtain the ψ -coalescent, one has already assumed that large offspring number events ($U = \lfloor \psi N \rfloor$) are much more frequent than ordinary Moran reproduction ($U = 1$) by taking $\gamma < 2$. Thus, ψ is not a timescale parameter (unlike α in the Beta coalescent), but rather reflects the size of the large offspring number event. Hence, if ψ is quite small, the resulting genealogy will look just like a Kingman coalescent genealogy by consisting only of binary mergers, but running on a much shorter timescale.

It will sometimes be convenient to denote by $q_{i,j} = \lambda_{i,k}$ the rate at which i active lineages change to $j = i - k + 1$ active lineages by the merging of k lineages. Also for ease of presentation, write

$$q_i = q_{i,i} = - \sum_{k=2}^i \lambda_{i,k} = - \sum_{j=1}^{i-1} q_{i,k}.$$

The reason for the negative sign is that the q_i terms are the diagonals in the instantaneous rate matrix for the continuous-time Markov chain $A_r(t)$ that counts the number of active lineages in a single population (see Appendix). The process $A_r(t)$ starts at time $t = 0$ in state $A_r(0) = r$, which means that at time zero we have r active ancestral lineages. The process stops as soon as $A_r(t) = 1$, in which case the r lineages have reached their most recent common ancestor.

A key quantity in our computations is the probability $\mathbb{P}[A_i(t) = j]$ for $1 \leq j \leq i$. Following Tavaré (1984) we write $g_{i,j}(t) = \mathbb{P}[A_i(t) = j]$, where t is measured in units associated with the coalescent process used. The units of time of the two Λ coalescent processes considered are shorter than the ones associated with the Kingman coalescent. Letting t_g represent time measured in generations and N the number of copies of a gene in a population, we use $t = t_g/N^{\alpha-1}$ for the Beta-coalescent. Substituting $\alpha = 2$ yields the natural coalescent units for the Kingman coalescent, $t = t_g/N$. Thus, the parameter α is a timescale parameter. Under the ψ -coalescent, the time t_g is based on time in the Moran model, and we have $t = t_g/N^\gamma$, $0 < \gamma < 2$. The issue of timescales is treated in more depth as a separate subsection below.

An explicit expression for $g_{i,j}(t)$ can be obtained when associated with the Kingman coalescent (Griffiths, 1979; Watterson, 1982; Tavaré, 1984). A generalization of the functions $g_{i,j}(t)$ to multiple merger coalescents is needed. The parameter π indicates the coalescent process used, where $\pi = 2$ indicates a Kingman coalescent, $1 < \pi < 2$ indicates a Beta-coalescent with $\alpha = \pi$, and $0 < \pi < 1$ indicates a ψ -coalescent with $\psi = \pi$. By $g_{i,j,\pi}$ we denote the probability that i lineages coalesce into j lineages by time t under a coalescent process with parameter π . Thus, $g_{i,j,\alpha}$ or $g_{i,j,\psi}$ will refer to the stated probability when associated with a particular process. An explicit expression is difficult to obtain for $g_{i,j,\pi}(t)$ when associated with a multiple merger coalescent. In Appendices B1 and B2 two ways to compute $g_{i,j,k}(t)$ for any Λ coalescent are presented. One method enumerates all the paths of $A_r(t)$, but is

not practical for $i > 20$, approximately, due to the exponential increase in number of paths for $A_i(t)$ to go from i to $j < i$ lineages as i increases. Another method to compute $g_{i,j,\pi}(t)$ involves finding the spectral decomposition of the rate matrix (12). Tavaré (1984) applies this technique to obtain the $g_{i,j}(t)$ for the Kingman coalescent. This method is computationally more feasible than listing all the paths of $A_i(t)$.

The probability that two lineages coalesce before time t is $g_{2,1,\pi}(t) = 1 - g_{2,2,\pi}(t) = 1 - \exp(-tq_2)$. Under the Beta-coalescent, $-q_2 = 1$, and $g_{2,1}(t)$ is therefore the same for both the Beta-coalescent and the Kingman coalescent. One must remember, however, that the units of t depend on the particular process. Two lineages coalesce with probability $1 - \exp(-1)$ within N generations under the Kingman coalescent, or within $N^{\alpha-1}$ generations for the Beta-coalescent. Thus, two lineages will tend to coalesce faster, in number of generations, under the Beta-coalescent than the Kingman coalescent. Under the point-mass coalescent, $g_{2,1,\psi}(t) = 1 - \exp(-t\psi^2)$, in which the timescale is also shorter than the corresponding one (proportional to N^2 timesteps) associated with the Kingman coalescent.

Considering three active lineages, one obtains

$$\begin{aligned} g_{3,1,\alpha}(t) &= 1 - \frac{3}{2}e^{-t} + \frac{1}{2}e^{-(1+\alpha)t} \\ g_{3,2,\alpha}(t) &= \frac{3}{2} \left[e^{-t} - e^{-(1+\alpha)t} \right], \\ g_{3,3,\alpha}(t) &= e^{-(1+\alpha)t}. \end{aligned} \quad (3)$$

In a population with the Beta-coalescent. In a population with the ψ -coalescent,

$$\begin{aligned} g_{3,1,\psi}(t) &= 1 - \frac{3}{2}e^{-t\psi^2} + \frac{1}{2}e^{-t[\psi^3+3(1-\psi)\psi^2]}, \\ g_{3,2,\psi}(t) &= \frac{3}{2} \left\{ e^{-t\psi^2} - e^{-t[\psi^3+3(1-\psi)\psi^2]} \right\}, \\ g_{3,3,\psi}(t) &= e^{-t(\psi^3+3(1-\psi)\psi^2)}. \end{aligned} \quad (4)$$

The quantities $g_{3,j,\pi}(t)$ in (3) and (4) have similar form as those obtained for the Kingman coalescent,

$$\begin{aligned} g_{3,1}(t) &= 1 - \frac{3}{2}e^{-t} + \frac{1}{2}e^{-3t} \\ g_{3,2}(t) &= \frac{3}{2} (e^{-t} - e^{-3t}) \\ g_{3,3}(t) &= e^{-3t}. \end{aligned} \quad (5)$$

In particular, $g_{i,j,\alpha}(t) = g_{i,j}(t)$ when $\alpha = 2$ for any i and j , which follows from the form (1) of the rate of coalescence associated with the Beta-coalescent. In case of the ψ -coalescent, one obtains $\lim_{\psi \rightarrow 0} g_{i,i,\psi}(t) = 1$.

In general, $g_{i,j,\pi}(t)$ is not a monotone function of j , and can have more than one peak under the ψ coalescent. Figs. 2 and 3 show graphs of $g_{i,j,\pi}(t)$ as a function of j for different values of π and time t of divergence. The value of j that maximizes $g_{i,j,\psi}(t)$ depends on ψ for intermediate values of t . Similar conclusions hold for $g_{i,j,\alpha}(t)$ (Fig. 2). When α is close to 1, and t is small enough, the values of $g_{i,j,\alpha}(t)$ become more evenly distributed over j as compared to the values of $g_{i,j}(t)$ associated with the Kingman coalescent. One must keep in mind, though, that α is a timescale parameter of the Beta-coalescent. Hence, different values of α result in different units of time. By way of example, consider the rate $\lambda_b \equiv \lambda_{b,2} + \dots + \lambda_{b,b}$, with $\lambda_{b,k}$ associated with the Beta-coalescent given by Eq. (1). The overall coalescent rate, λ_b , is an increasing function of α when measured in units of $N^{\alpha-1}$ generations (Appendix D), which means that the probability $\mathbb{P}[A_i(t) = i] = e^{-t\lambda_b}$ that none of the i lineages

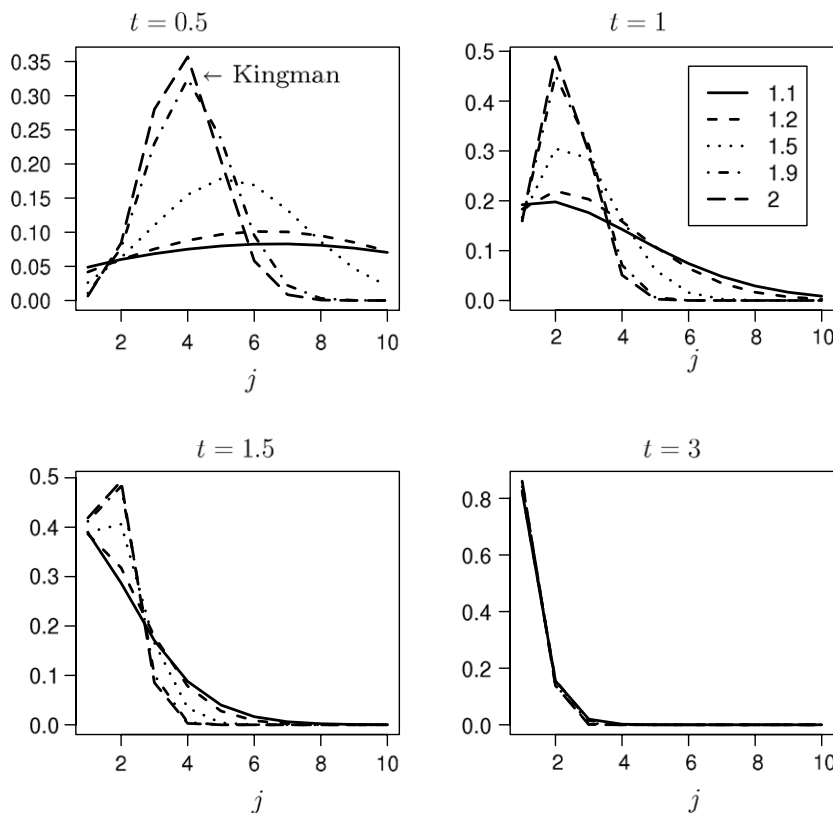


Fig. 2. The probability $g_{30,j,\alpha}(t)$ as a function of j as α varies over the values shown in the legend. Results for the Kingman coalescent ($\alpha = 2$) are shown for reference. Each plot is for different values of t .

have coalesced by time t decreases as α increases when t is fixed. As α increases, however, the unit of time $N^{\alpha-1}$ increases also. As a result, the expected time until the next coalescence is increasing in α when time is measured in generations and the sample size is not a large fraction of the population size N (Appendix D).

Comparing Figs. 2 and 3, one observes that $g_{i,1,\alpha}(t)$ tends to one much more quickly than $g_{i,1,\psi}(t)$ as t increases, in particular for low values of ψ . Comparing Eqs. (3) and (4) is helpful in this context. The expressions for $g_{3,j,\psi}(t)$ (Eq. (4)) show that only when $t \gg 1/\psi^2$ is $g_{3,1,\psi}(t) \approx 1$.

Exact computations involving coalescent processes with multiple mergers are generally only possible for moderate sample sizes. It would be highly desirable to be able to distill the most likely paths of $A_i(t)$, but this is not a simple matter. By way of example, one obtains for the Beta-coalescent

$$q_{i,i-1} > q_{i,i-2} > \dots > q_{i,1}$$

for any value of $\alpha \in (1, 2)$, and hence a merger of two active lineages is always the most likely merger in a Beta-coalescent. However, a path of only 2-mergers may not be the most likely path. Table 1 reports the probability $P_2(\pi)$ of a path consisting of only 2-mergers as a function of sample size and π , and shows that $P_2(\pi)$ decreases quickly as a function of sample size. Table 2 reports the most likely sequence of mergers for the two A coalescents for different values of π , and sample size (i), for paths of $A_i(t)$ going from i to 1. The most likely paths of $A_i(t)$ can be those beginning with a large merger, at least for the small sample sizes considered, suggesting that it is more likely for tree topologies to be unresolved near the tips than towards the root of the tree. Table 2 also reveals that, in some cases, the vast majority of paths will have low probability compared to the highest probability.

Table 1

The probability $P_2(\pi)$ a gene genealogy consists of only binary mergers, as a function of sample size n , and π . The parameter π stands for α when $1 < \pi < 2$, and for ψ when $0 < \pi < 1$.

α	ψ	n	$P_2(\alpha)$	$P_2(\psi)$
1.01	0.01	5	0.319	0.980
		25	0.000	0.391
		50	0.000	0.017
1.05	0.05	5	0.343	0.900
		25	0.000	0.006
		50	0.000	0.000
1.2	0.1	5	0.438	0.802
		25	0.000	0.000
		50	0.000	0.000
1.5	0.2	5	0.643	0.611
		25	0.004	0.000
		50	0.000	0.000
1.9	0.5	5	0.928	0.157
		25	0.388	0.000
		50	0.115	0.000

2.2. Two extant populations

Additional notation is needed to write down the recursion for the probability of reciprocal monophyly. Let n_A and n_B denote the number of sequences sampled for species A and B, respectively. Of the n_A and n_B lineages, m_A and m_B pass into the common ancestral population from A and B, respectively. Let $m = m_A + m_B$. By P we denote the probability of monophyletic concordance (Fig. 1a) between the species tree of species A and B, and the gene genealogy of the sample of n_A lineages from species A, and n_B lineages from species B. Denote by T the time when lineages from both derived populations A and B are first involved in a coalescence event. Let T_A and T_B denote the times of the most recent common ancestors

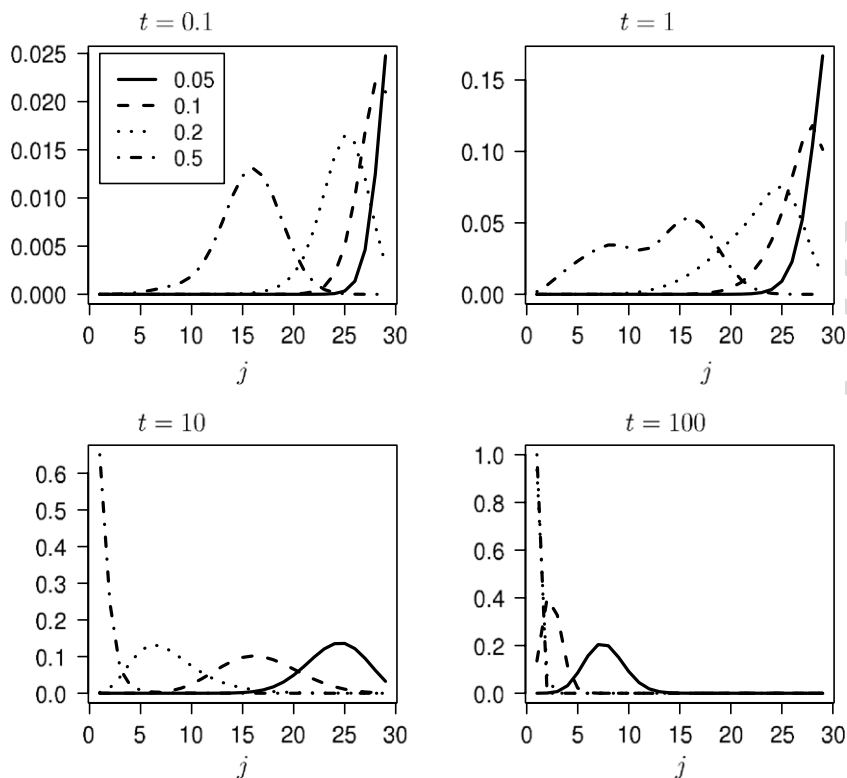


Fig. 3. The probability $g_{30,j,\psi}(t)$ as a function of j as ψ varies over the values shown in the legend. Each plot is for different values of t .

Table 2

The paths of $A_i(t)$ with the highest probability for the α - and ψ -coalescents. The parameter π stands for α when $1 < \pi < 2$, and for ψ when $0 < \pi < 1$. Sample size is denoted by i . By c^* denote the sequence of mergers corresponding to the path with the highest probability. The % column denotes the fraction of sequences whose probability is $< 1/1000$ of the probability of c^* .

π	i	%	c^*
1.01	10	0	(2, ..., 2)
	20	66.9	(19, 2)
1.05	10	0	(2, ..., 2)
	20	58.6	(19, 2)
1.2	10	0	(2, ..., 2)
	20	39.8	(18, 2, 2)
1.5	10	18.8	(2, ..., 2)
	20	96.6	(2, ..., 2)
1.9	10	90.2	(2, ..., 2)
	20	99.9	(2, ..., 2)
0.01	10	96.5	(2, ..., 2)
	20	99.9	(2, ..., 2)
0.05	10	84.4	(2, ..., 2)
	20	99.2	(2, ..., 2)
0.1	10	61.7	(2, ..., 2)
	20	84.1	(3, 3, 2, ..., 2)
0.2	10	14.1	(3, 2, ..., 2)
	20	59.3	(5, 4, 4, 3, 2, ..., 2)
0.5	10	23.8	(6, 3, 2, 2)
	20	98.9	(11, 6, 3, 2, 2)

of the lineages from A and B , respectively. By τ we denote the time of divergence. One can now express the probability $P(n_A, n_B, \tau)$ of reciprocal monophyly as

$$P(n_A, n_B, \tau) = \mathbb{P}[T > T_A, T > T_B].$$

Similarly to the approach of Hudson and Coyne (2002), we condition on the number of lineages entering the common ancestral population from populations A and B . The probability $P(n_A, n_B, \tau)$ is a function of the sample sizes n_A and n_B and the time

τ of divergence, and is obtained recursively by

$$P(n_A, n_B, \tau) = \sum_{m_A=1}^{n_A} \sum_{m_B=1}^{n_B} P(m_A, m_B, 0) \times g_{n_A, m_A, \pi_A}(\tau) g_{n_B, m_B, \pi_B}(\tau) \quad (6)$$

and, with $m = m_A + m_B$,

$$P(m_A, m_B, 0) = \sum_{k=2}^{\max(m_A, m_B)} \left[\frac{\binom{m_A}{k}}{\binom{m}{k}} p(m, k) P(m_A, m_B - k + 1, 0) - k + 1, m_B, 0 \right] + \frac{\binom{m_B}{k}}{\binom{m}{k}} p(m, k) P(m_A, m_B - k + 1, 0) \quad (7)$$

with $P(1, 1, 0) = 1$, and

$$p(m, k) = \frac{q_{m, m-k+1}}{-q_m}, \quad 2 \leq k \leq m,$$

is the probability of a k merger given m active ancestral lineages. The coefficient $\binom{u}{v} = 0$ if $v > u$, and $g_{n_A, m_A, \pi_A}(\tau)$ denotes the probability that m_A of n_A lineages pass into the common ancestral population.

Fig. 4 reports values of the probability of monophyly P (6) as a function of time τ of divergence when all populations have the same Λ coalescent with same parameter value. The values of P associated with a Beta-coalescent are not directly comparable for different values of α since α is a timescale parameter. However, Fig. 4 reveals that about four coalescent time units result in approximately 80% probability of monophyly for any Beta-coalescent. Values of P are directly comparable for different values of ψ , and P varies quite a bit with ψ . Knowing ψ is thus crucial for estimating time of divergence in a population with the ψ -coalescent, since P can be high even for a short time of divergence, if ψ is high.

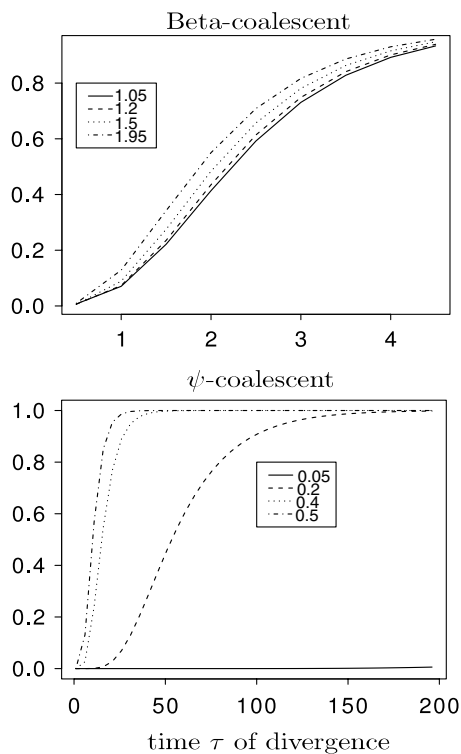


Fig. 4. The probability P of monophyly as a function of time τ of divergence when all populations have the same Λ coalescent. Different lines in each plot represent different values of π associated with each process (see legends). Initial sample size is $n_A = n_B = 40$.

The probability of monophyly depends strongly on values of ψ in the descendent populations when all populations have the ψ -coalescent, and only weakly on values of ψ in the common ancestral population (Fig. 5).

2.3. Timescales

Probabilities of monophyly and $g_{i,j,\pi}(\cdot)$ are functions of time. The appropriate way to scale time is in units of c_N generations (timesteps) in which c_N is the probability that two lineages coalesce in one timestep in the discrete-time model (Möhle, 2000). If v_i denotes the number of offspring of a single individual (i), $c_N = \mathbb{E}[v_i(v_i - 1)] / (N - 1)$, if the v_i are identically distributed. In a Fisher-Wright haploid population of constant size N , $c_N = 1/N$. In the usual Moran model, $c_N \approx 1/N^2$. For the ψ -coalescent, the leading term of c_N is ψ^2/N^γ , with $0 < \gamma < 2$. However, to facilitate interpretation of the results, we scale time in units of N^γ for the ψ -coalescent, with $1 < \gamma < 2$. We interpret γ to be a fixed parameter, and its exact value is not important, as none of the quantities we are concerned with are functions of γ . The timescale N^γ for the ψ -coalescent should be compared to the one associated with the usual Moran model. Thus, the ψ -coalescent runs on a shorter timescale than the corresponding timescale associated with the Kingman coalescent. By way of example, $t = 1$ on the Kingman coalescent timescale corresponds to $t = N^{2-\gamma}$ on the ψ -coalescent timescale. Similarly, c_N is proportional to $N^{1-\alpha}$ with $1 < \alpha < 2$ when associated with the Beta-coalescent (Schweinsberg, 2003). Thus, $t = 1$ on the corresponding Kingman coalescent timescale in units of N generations corresponds to $t = N^{2-\alpha}$ on the Beta-coalescent timescale for a given value of α . Since α is a timescale parameter, Beta-coalescents with different values of α have different units of time, each proportional to $N^{\alpha-1}$.

To compare the Beta-coalescent to the Kingman coalescent, it is helpful to convert units of time into generations, so that the

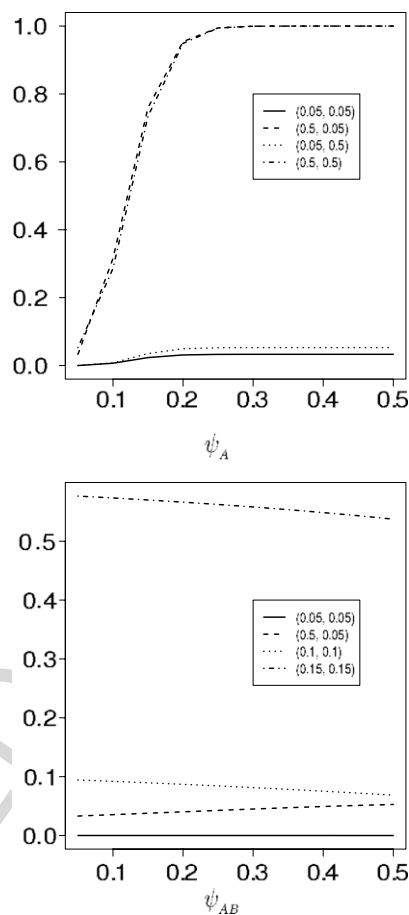


Fig. 5. The probability P as a function of ψ_A when all populations have the ψ -coalescent, with $n_A = n_B = 40$, time $\tau = 100$. Different lines represent different values of the vector (ψ_B, ψ_{AB}) , or for (ψ_A, ψ_B) (lower panel). See legends for values of parameters.

timescale does not depend on the parameter α . To do this we have computed probabilities of monophyly as a function of the number of generations since divergence. Fig. 6 gives the probability of reciprocal monophyly when the divergence time ranges from 0 to 6N generations, using $N = 10^5$ (Fig. 6a) and $N = 10^4$ (Fig. 6b). Here we see that the probability of monophyly as a function of time in generations can depend strongly on α , with smaller values of α reaching high probabilities of monophyly with much more recent species divergences than the Kingman coalescent processes with larger α values. The probability of monophyly is not strictly decreasing in α , however, if the number of generations is sufficiently small. This is shown by zooming in to the probabilities of monophyly for a very small number of generations (Fig. 6, top right panel).

The probability of monophyly measured in generations depends on N more strongly for smaller values of α , which is a result of the probability depending on $t/N^{\alpha-1}$ rather than t/N (where t is time measured in generations), as in the Kingman coalescent. Thus, the shapes of the curves for the Kingman coalescent ($\pi = 2$ in Fig. 6a and b) are exactly the same (ignoring the scaling of the time axis) when changing from $N = 10^5$ to $N = 10^4$, but the shapes of the curves change slightly for smaller α values, with the probability of monophyly being slightly higher for larger N when measured in units of N generations. We also see that the shapes of the curves are similar with $n_A = n_B = 4$ lineages versus $n_A = n_B = 40$ lineages per population, with monophyly being achieved slightly faster for the smaller sample size.

One might be interested in knowing, for each process, the effects of the coalescent parameter π on the time by which all

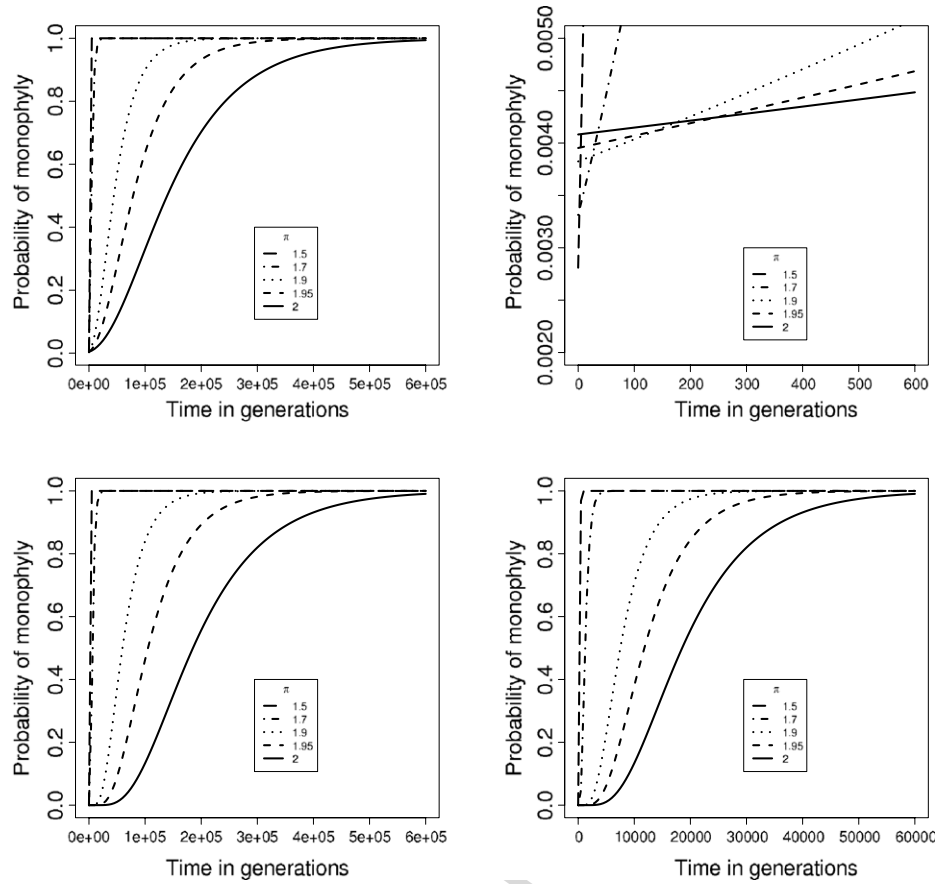


Fig. 6. Probabilities of monophyly as a function of time in generations for the Beta-coalescent and Kingman coalescent (solid line). In the top row, $n_A = n_B = 4$ and $N = 10^5$. In the bottom row, $n_A = n_B = 40$ with $N = 10^5$ (bottom left) and $N = 10^4$ (bottom right).

Table 3

Estimates of $t^*(\pi) = \inf\{t > 0 : g_{i,1,\pi}(t) > 0.999999\}$ for different values of i , α , and ψ .

α	ψ	i	$t^*(\alpha)$	$t^*(\psi)$
1.01	0.01	5	14.55	145,000
		25	15.15	148,000
		50	15.31	149,000
1.05	0.05	5	14.55	5,806
		25	15.13	5,952
		50	15.29	5,977
1.2	0.1	5	14.54	1,452
		25	15.07	1,493
		50	15.21	1,502
1.5	0.2	5	14.53	363
		25	14.97	376
		50	15.06	380
1.9	0.5	5	14.52	59
		25	14.86	62
		50	14.91	63

lineages have coalesced with a ‘high’ probability. Table 3 reports estimates of $t^*(\pi) \equiv \inf\{t > 0 : g_{i,1,\pi}(t) \geq 0.999999\}$, or the smallest values of time t at which i lineages have all coalesced with a high probability in a single population. The results in Table 3 show a much stronger effect of ψ than α on t^* . Sample size appears to matter little for t^* for both processes. The Beta-coalescent ‘comes down from infinity’ (Schweinsberg, 2000b), which means that the number of active ancestral lineages becomes finite in finite time, if the process started with an infinite number of lineages. The point mass process does not come down from infinity, as one can check using Schweinsberg (2000b)’s results. Thus, $t^*(\alpha)$ should approach a limit as sample size tends to infinity.

Different populations may have different population sizes. If a Fisher–Wright population has population size bN for some constant $b > 0$, on a timescale of N generations, then each pair of ancestral lineages coalesces with rate $1/b$. In a Moran population with only one parent each timestep, the rate is $1/b^2$ (Eldon, 2009) on a timescale proportional to N^2 . In the presence of large families, the scaling by b depends on the specifics of the model. In the modified Moran model of Eldon and Wakeley (2006), the constant b cancels from the coalescence rate of multiple mergers (Eldon, 2009). Consider Schweinsberg’s (2003) model, and let c_{bN} denote the timescale in a population of size bN . As c_N is proportional to $N^{1-\alpha}$,

$$\lim_{N \rightarrow \infty} \frac{c_{bN}}{c_N} = b^{1-\alpha}, \quad b > 0; \quad (8)$$

in which $c_{bN} = \mathbb{E}[(v_{1,Nb})_2] / (Nb - 1)$ and the random variable $v_{1,Nb}$ denotes the number of offspring of individual 1 in a population of size Nb . One can follow Schweinsberg (2003) to establish that, in a large population of size bN ,

$$\lambda_{n,k} = b^{1-\alpha} \binom{n}{k} \frac{\Gamma(k - \alpha) \Gamma(n - k + \alpha)}{\Gamma(n) \Gamma(2 - \alpha) \Gamma(\alpha)}, \quad (9)$$

$$1 < \alpha < 2, \quad b > 0,$$

when the timescale is c_N .

Let c_A and c_B denote the relative population size scaling constants (8) for populations A and B , respectively. Fig. 7 shows the probability of monophyly as a function of time of divergence for different values of c_B and α ($n_A = n_B = 40$), with $c_A = 1$ in all cases. In Fig. 7, the size of population B can be taken as being scaled relative to A , which has population size N . The form of the

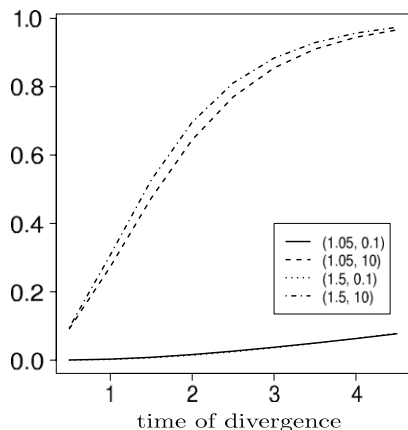


Fig. 7. Graphs of P as a function of time τ of divergence when all populations have the same Beta-coalescent. Values of α and the population size scaling constant c_B vary as shown in the legend (as (α, c_B)).

coalescence rates (9) results in the probability $g_{i,1,\alpha}(t)$ being small for large t when c_B is small, i.e., when the population size bN of B is large relative to the one of A . On the other hand, P increases quickly with time when c_B is large, i.e., when the two populations differ significantly in size.

2.4. Paraphyly and polyphyly

Paraphyly occurs when the lineages from only one of the two descendent populations are monophyletic (Fig. 1b). Let $P_A \equiv \mathbb{P}[T > T_A]$, in which T was the first time at which lineages from both A and B coalesce. The probability P_B^* of paraphyly of B with respect to A is

$$P_B^* = \mathbb{P}[T > T_A, T \leq T_B] \\ = \mathbb{P}[T > T_A] - \mathbb{P}[T > T_A, T > T_B].$$

Polyphyly is the event $\{T \leq T_A\} \cap \{T \leq T_B\}$ (Fig. 1c, d), which occurs with probability

$$P^* = \mathbb{P}[T \leq T_A, T \leq T_B] = 1 - P_A - P_B + P.$$

The probabilities P_A and P_B can be obtained recursively analogously to Eqs. (6) and (7) with P replaced by P_A (or P_B), with boundary conditions

$$P_A(1, m_B, \tau) = 1, \quad P_B(m_A, 1, \tau) = 1, \quad \tau \geq 0.$$

As expected, the probability of polyphyly increases with sample size for the Beta- and ψ -coalescents if time of divergence was recent (results not shown). Polyphyly is most likely to occur when the time of divergence was recent (Fig. 8). The values of monophyly, paraphyly, and polyphyly associated with the Beta-coalescent (Fig. 8a) for different values of α must be interpreted in light of the timescale property of α already mentioned. When the value of ψ is low (Fig. 8b), monophyly becomes most likely for more ancient divergence times (in ψ -coalescence time units), compared to populations with high value of ψ (Fig. 8c).

2.5. Clade probabilities

Monophyly for a set of lineages can be used as evidence that the lineages should be considered a separate taxonomic group (Hudson and Coyne, 2002). Using the null hypothesis that all lineages are from the same population, clade probabilities have been used as a measure of taxonomic distinctiveness for a set of lineages in a population (or subpopulation) that form a clade (Rosenberg, 2007). Here we can compare probabilities of clades

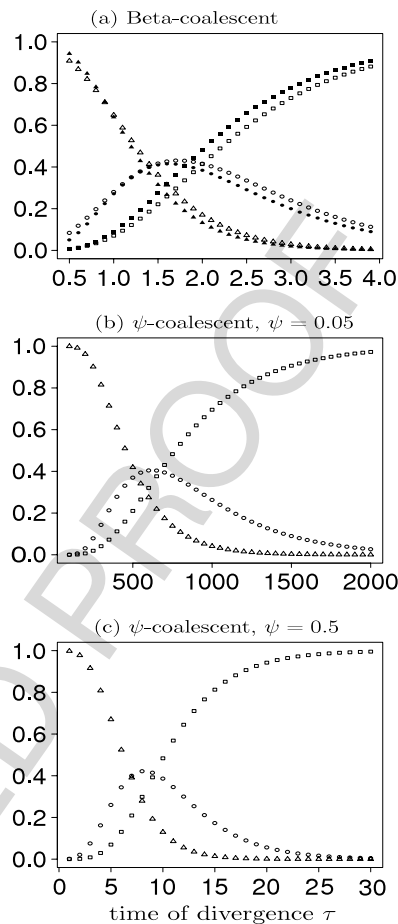


Fig. 8. Probabilities of monophyly (squares), paraphyly (circles), and polyphyly (triangles) for two species as a function of time τ of divergence, and $n_A = n_B = 40$. In (a) $\alpha = 1.05$ (open symbols); $\alpha = 1.5$ (closed symbols).

in single populations for multiple merger coalescents and the Kingman coalescent.

The quantities P_A and P_B are clade probabilities, which are also of interest in single populations—i.e., when the species divergence time τ is zero. In particular, $P_A(i, j)$ with boundary condition $P_A(1, j) = 1$ is the probability that i lineages from A form a clade when $i + j$ lineages have been drawn from a single population.

Probabilities of clades under multiple merger models have considerably different properties from clade probabilities under strictly bifurcating models of trees such as the Kingman coalescent and Yule models. For example, a constraint that exists for models of binary, rooted trees is that the sum of the clade probabilities must be exactly $n - 2$ (Allman et al., 2011), where n is the number of species. Thus if C_i is a clade, we use $\mathbb{P}(C_i)$ to denote the probability that C_i is a clade given a random genealogy. There are $k = 2^n - n - 2$ nontrivial clades for a genealogy with n lineages, and the clade probabilities must satisfy the constraint

$$\sum_{i=1}^k \mathbb{P}(C_i) = n - 2.$$

However, for multiple merger models, we show in Appendix C that the sum of clade probabilities is strictly less than $n - 2$:

$$\sum_{i=1}^k \mathbb{P}(C_i) < n - 2. \quad (10)$$

The result is due to the fact that multiple merger trees have fewer clades, thus reducing the chances for many clades to occur in a

Table 4

Probabilities of monophyly of lineages from group *A* with sample sizes from group *A* (rows) and group *B* (columns) under the null hypothesis that all lineages are from the same unstructured population. $\pi = 2$ is the Kingman coalescent.

Lineages from <i>A</i>	π	Lineages in group <i>B</i>									
		1	2	3	4	5	6	7	8	9	10
2	2	0.3333	0.2222	0.1667	0.1333	0.1111	0.0952	0.0833	0.0741	0.0667	0.0606
2	1.5	0.3000	0.1857	0.1333	0.1035	0.0842	0.0709	0.0611	0.0536	0.0477	0.0430
2	0.1	0.3214	0.2047	0.1476	0.1137	0.0913	0.0754	0.0636	0.0544	0.0472	0.0413
3	2	0.1667	0.0833	0.0500	0.0333	0.0238	0.0179	0.0139	0.0111	0.0091	0.0076
3	1.5	0.1571	0.0714	0.0405	0.0260	0.0180	0.0132	0.0100	0.0079	0.0064	0.0052
3	0.1	0.1666	0.0798	0.0466	0.0304	0.0212	0.0156	0.0118	0.0093	0.0074	0.0060
4	2	0.1000	0.0400	0.0200	0.0114	0.0071	0.0048	0.0033	0.0024	0.0018	0.0014
4	1.5	0.1000	0.0359	0.0169	0.0092	0.0056	0.0036	0.0025	0.0018	0.0013	0.0010
4	0.1	0.1033	0.0394	0.0192	0.0108	0.0066	0.0043	0.0030	0.0021	0.0016	0.0012
5	2	0.0667	0.0222	0.0095	0.0048	0.0026	0.0016	0.0010	0.0007	0.0005	0.0003
5	1.5	0.0706	0.0209	0.0084	0.0040	0.0021	0.0013	0.0008	0.0005	0.0003	0.0002
5	0.1	0.0712	0.0224	0.0094	0.0046	0.0025	0.0015	0.0009	0.0006	0.0004	0.0003
6	2	0.0476	0.0136	0.0051	0.0023	0.0011	0.0006	0.0004	0.0002	0.0001	9.5e-05
6	1.5	0.0531	0.0134	0.0047	0.0020	0.0010	0.0005	0.0003	0.0002	0.0001	7.2e-05
6	0.1	0.0525	0.0140	0.0051	0.0022	0.0011	0.0006	0.0003	0.0002	0.0001	8.8e-05
7	2	0.0357	0.0089	0.0030	0.0012	0.0005	0.0003	0.0001	8.3e-05	5.0e-05	3.1e-05
7	1.5	0.0418	0.0092	0.0028	0.0011	0.0005	0.0002	0.0001	6.7e-05	4.0e-05	2.4e-05
7	0.1	0.0406	0.0094	0.0030	0.0012	0.0005	0.0003	0.0001	8.1e-05	4.8e-05	3.0e-05
8	2	0.0278	0.0062	0.0019	0.0007	0.0003	0.0001	6.5e-05	3.5e-05	1.9e-05	1.1e-05
8	1.5	0.0340	0.0066	0.0018	0.0006	0.0003	0.0001	5.5e-05	2.9e-05	1.6e-05	9.2e-06
8	0.1	0.0326	0.0066	0.0019	0.0006	0.0003	0.0001	6.4e-05	3.4e-05	1.9e-05	1.1e-05
9	2	0.0222	0.0044	0.0012	0.0004	0.0002	6.7e-05	3.1e-05	1.6e-05	8.2e-06	4.6e-06
9	1.5	0.0283	0.0049	0.0012	0.0004	0.0001	6.0e-05	2.7e-05	1.3e-05	6.9e-06	3.8e-06
9	0.1	0.0269	0.0048	0.0013	0.0004	0.0002	6.8e-05	3.1e-05	1.6e-05	8.2e-06	4.5e-06
10	2	0.0182	0.0033	0.0008	0.0003	9.1e-05	3.6e-05	1.6e-05	7.5e-06	3.7e-06	2.0e-06
10	1.5	0.0240	0.0038	0.0009	0.0003	8.7e-05	3.4e-05	1.4e-05	6.6e-06	3.2e-06	1.7e-06
10	0.1	0.0227	0.0037	0.0009	0.0003	9.4e-05	3.7e-05	1.6e-05	7.6e-06	3.8e-06	2.0e-06

random tree from one of these models. For example, a four-taxon binary tree such as $((A_1, A_2), B_1), B_2)$, with two lineages from *A* and two from *B*, has two clades, one with two lineages from *A*, and one clade with three lineages (A_1, A_2 , and B_1). However, if a 3-merger occurs on a 4-taxon tree, such as for the tree $((A_1, A_2, B_1), B_2)$, then there is only one non-trivial clade, which in this case has three lineages. The inequality (10) holds more generally than just for the Δ -coalescents considered in this paper. In particular, it just needs to be assumed that at least one multifurcating tree has positive probability.

As another example, under a strictly bifurcating model, a tree must have at least one *cherry*, a clade with exactly two lineages (McKenzie and Steel, 2000). Consequently, for a process that produces a strictly bifurcating tree with n descendants labeled A_1, \dots, A_n , clade probabilities must satisfy the constraint

$$\sum_{i=1}^{n-1} \sum_{j=i+1}^n \mathbb{P}[(A_i, A_j) \text{ is a clade}] \geq 1. \quad (11)$$

The sum in (11) can be greater than 1.0 because not all clades are mutually exclusive if $n > 3$. Under multiple merger models, however, it is possible for the sum of the probabilities of cherries to be less than 1 if multiple mergers are sufficiently likely.

Probabilities of n_A lineages forming a clade out of $n_A + n_B$ total lineages are given in Table 4 under the point-mass coalescent (with $\psi = 0.1$), Beta-coalescent with $\alpha = 1.5$, and the Kingman coalescent. Values for the Kingman coalescent are also reported in Table 1 of Rosenberg (2007). Tables analogous to those in Rosenberg (2007) can be constructed for different values of ψ and α to determine significance levels for observed levels of monophyly and reciprocal monophyly. The probabilities listed in the table can be interpreted as p -values for having observed monophyly of the a lineages sampled from species *A* under the null hypothesis of all lineages being from the same population. This p -value gives a test of the taxonomic distinctiveness of *A*.

From Table 4, it appears that for small numbers of lineages in *A*, there is stronger evidence against the null hypothesis of a single

population under typical multiple merger coalescents than under the Kingman coalescent. This trend strengthens if the number of lineages in *A* is kept small and the total number of lineages increases. For example, with a total of 12 lineages sampled, two of which are from *A*, the probability that the two from *A* are monophyletic is 0.06 under the Kingman coalescent and about 0.04 under the two multiple merger coalescent models tried.

The intuition behind these results is that multiple merger coalescents will tend to put increased probability on larger clades, so observing small clades can be better evidence (compared to a Kingman coalescent model) against the hypothesis that all lineages are from a single population. The p -value for small clades also will tend to decrease as ψ increases and as α decreases, since larger ψ and smaller α result in increased probability of multiple mergers. For example, with 12 lineages sampled, two of which are sampled from *A*, the probability of monophyly for the *A* lineages decreases from 0.041 to 0.026 as ψ changes from 0.1 to 0.2.

The flip side of this intuition is that when the number of lineages sampled from population *A* is relatively large compared to the number of lineages sampled from *B*, the probability of monophyly of the *A* lineages can be larger under a multiple merger coalescent than under the Kingman coalescent (e.g., Table 4 with 10 lineages from *A* and 3 or less not from *A*).

A similar comparison between different coalescent models can be made for probabilities of reciprocal monophyly for two groups under the null hypothesis of a single unstructured population. In most examples tried, as long as the number of lineages in population *B* is greater than 1, then the probability of reciprocal monophyly is slightly lower under multiple merger models than under a Kingman model (Table 5; see also Table 6 of Rosenberg (2007)), suggesting that an observation of reciprocal monophyly under a multiple merger model is often slightly stronger evidence against the null hypothesis of a single population than is obtained under the Kingman coalescent. Consistent with this observation, as the parameter changes to make multiple mergers more likely (i.e., smaller α or larger ψ), the probability of reciprocal monophyly decreases, can be computed using $P(m_A, m_B, 0)$ (Table 6).

Table 5
Probabilities of reciprocal monophyly with sample sizes from group *A* (rows) and not *A* (columns) under the null hypothesis that all lineages are from the same unstructured population.

Lineages from <i>A</i>	π	Lineages from <i>B</i>									
		1	2	3	4	5	6	7	8	9	10
2	2	0.3333	0.1111	0.0500	0.0267	0.0159	0.0102	0.0069	0.0049	0.0036	0.0028
2	1.5	0.3000	0.0857	0.0381	0.0208	0.013	0.0085	0.0060	0.0044	0.0034	0.0026
2	0.1	0.3214	0.0996	0.0448	0.0242	0.0146	0.0095	0.0066	0.0047	0.0035	0.0027
3	2	0.1667	0.0500	0.0200	0.0095	0.0051	0.0030	0.0019	0.0012	0.0008	0.0006
3	1.5	0.1571	0.0381	0.0144	0.0068	0.0037	0.0022	0.0014	0.0009	0.0007	0.0005
3	0.1	0.1666	0.0448	0.0176	0.0083	0.0045	0.0026	0.0016	0.0011	0.0007	0.0005
4	2	0.1000	0.0267	0.0095	0.0041	0.0020	0.0011	0.0006	0.0004	0.0002	0.0002
4	1.5	0.1000	0.0208	0.0068	0.0028	0.0014	0.0007	0.0004	0.0003	0.0002	0.0001
4	0.1	0.1033	0.024	0.0083	0.0035	0.0017	0.0009	0.0005	0.0003	0.0002	0.0001
5	2	0.0667	0.0159	0.0051	0.0020	0.0009	0.0004	0.0002	0.0001	7.7e-05	4.8e-05
5	1.5	0.0706	0.0128	0.0037	0.0014	0.0006	0.0003	0.0002	8.7e-05	5.2e-05	3.3e-05
5	0.1	0.0712	0.0146	0.0045	0.0017	0.0008	0.0004	0.0002	0.0001	6.5e-05	4.1e-05
6	2	0.0476	0.0102	0.0030	0.0011	0.0004	0.0002	9.7e-05	5.1e-05	2.9e-05	1.7e-05
6	1.5	0.0531	0.0085	0.0022	0.0007	0.0003	0.0001	6.3e-05	3.3e-05	1.9e-05	1.1e-05
6	0.1	0.0525	0.0095	0.0026	0.0009	0.0004	0.0002	8.2e-05	4.3e-05	2.4e-05	1.4e-05
7	2	0.0357	0.0069	0.0019	0.0006	0.0002	9.7e-05	4.5e-05	2.2e-05	1.2e-05	6.4e-06
7	1.5	0.0418	0.0060	0.0014	0.0004	0.0002	6.3e-05	2.9e-05	1.4e-05	7.4e-06	4.1e-06
7	0.1	0.0406	0.0066	0.0016	0.0005	0.0002	8.2e-05	3.7e-05	1.8e-05	9.7e-06	5.3e-06
8	2	0.0278	0.0049	0.0012	0.0004	0.0001	5.1e-05	2.2e-05	1.0e-05	5.1e-06	2.7e-06
8	1.5	0.0340	0.0044	0.0009	0.0002	8.7e-05	3.3e-05	1.4e-05	6.5e-06	3.2e-06	1.7e-06
8	0.1	0.0326	0.0047	0.0011	0.0003	0.0001	4.3e-05	1.8e-05	8.6e-06	4.2e-06	2.2e-06
9	2	0.0222	0.0036	0.0008	0.0002	7.7e-05	2.9e-05	1.2e-05	5.1e-06	2.4e-06	1.2e-06
9	1.5	0.0283	0.0034	0.0007	0.0002	5.2e-05	1.9e-05	7.4e-06	3.2e-06	1.5e-06	7.5e-07
9	0.1	0.0269	0.0035	0.0007	0.0002	6.5e-05	9.6e-06	1.1e-05	4.2e-06	2.0e-06	9.8e-07
10	2	0.0182	0.0028	0.0006	0.0002	4.8e-05	1.7e-05	6.4e-06	2.7e-06	1.2e-06	5.7e-07
10	1.5	0.0240	0.0026	0.0005	0.0001	3.3e-05	1.1e-05	4.1e-06	1.7e-06	7.5e-07	3.5e-07
10	0.1	0.0227	0.0027	0.0005	0.0001	4.1e-05	1.4e-05	5.3e-06	2.2e-06	9.8e-07	4.6e-07

Table 6
Values of $P(m, m)$ (7) for different values of α_{AB} , ψ_{AB} , and $m = m_A = m_B$.

α	ψ	m	$P(\alpha)$	$P(\psi)$
1.01	0.01	4	0.00163	0.00403
		20	$9.756 \cdot 10^{-14}$	$3.632 \cdot 10^{-13}$
1.05	0.05	4	0.00172	0.00380
		20	$1.056 \cdot 10^{-13}$	$3.286 \cdot 10^{-13}$
1.2	0.1	4	0.00207	0.00353
		20	$1.384 \cdot 10^{-13}$	$2.881 \cdot 10^{-13}$
1.5	0.2	4	0.00281	0.00148
		20	$2.151 \cdot 10^{-13}$	$2.175 \cdot 10^{-13}$
1.9	0.5	4	0.00382	0.00148
		20	$3.377 \cdot 10^{-13}$	$8.229 \cdot 10^{-14}$

In addition to considering probabilities of monophyly for a single group and for two groups, formulas are available for computing probabilities that k groups are each monophyletic when there are more than k groups within a single population. In particular, supposing that the groups are A_1, A_2, \dots, A_s with m_1, m_2, \dots, m_s lineages, respectively. The probability that some subsets, $A_{i_1}, A_{i_2}, \dots, A_{i_k}$, $k \leq s$, are each monophyletic can be computed analytically under the Kingman coalescent (Zhu et al., 2011). The framework used in this paper could be extended to compute similar quantities under multiple merger coalescents by extending the recursion (7). In particular, the probability that the k groups are each monophyletic is

$$P(m_{i_1}, \dots, m_{i_k}, 0) = \sum_{k=2}^{\max(m_{i_1}, \dots, m_{i_k})} \sum_{j=1}^k \binom{m_j}{k} \binom{m}{k} \times [p(m, k)P(m_1, \dots, m_j - k + 1, \dots, m_s, 0)]$$

where $m = \sum_{j=1}^s m_j$. The boundary condition for the recursion is $P(\mathbf{v}, 0) = 1$ where \mathbf{v} is a list of s 1s and 0s specifying whether a group is required to be monophyletic:

$$v_j = \begin{cases} 1 & j \in \{i_1, \dots, i_k\} \\ 0 & \text{otherwise.} \end{cases}$$

In principle, probabilities of monophyly for more than two groups when there has been multiple species divergences or a polytomy at the species level could also be developed by similarly generalizing Eq. (6).

3. Discussion

Monophyly, paraphyly, and polyphyly of ancestral lineages are important concepts in molecular ecology and phylogeography (Neigel and Avise, 1986; Nee et al., 1996; Avise, 1989; Palumbi et al., 2001). Wakeley and Hey (1997) infer population histories of the ancestral and the two descendent populations using coalescent methods. Quantifying probabilities of monophyly, paraphyly, and polyphyly under various demographic scenarios is therefore helpful in understanding evolutionary histories of populations, both ancestral and derived. The present focus is on ancestral lineages in an isolation model of two species, in which the coalescent processes admit multiple mergers of ancestral lineages.

Probabilities of monophyly, paraphyly, and polyphyly, are computed, using recursion, in a model of two species. Rosenberg (2003) obtains closed-form expressions for the three probabilities in question, when the coalescent process is the Kingman coalescent. The multiple-merger nature of the coalescent processes presently considered makes obtaining closed-form expressions an arduous task, and limits exact computations to modest sample sizes. However, important insights can still be obtained.

3.1. Multiple merger coalescent processes

Multiple merger coalescent processes (Donnelly and Kurtz, 1999; Pitman, 1999; Schweinsberg, 2000a; Möhle and Sagitov, 2001) arise from population models in which individuals have a non-negligible propensity for having very many offspring (Schweinsberg, 2003; Eldon and Wakeley, 2006; Sargsyan and Wakeley, 2008). The question of which large offspring number model applies to which population remains, however, very much open. Resolving this question will require detailed knowledge of

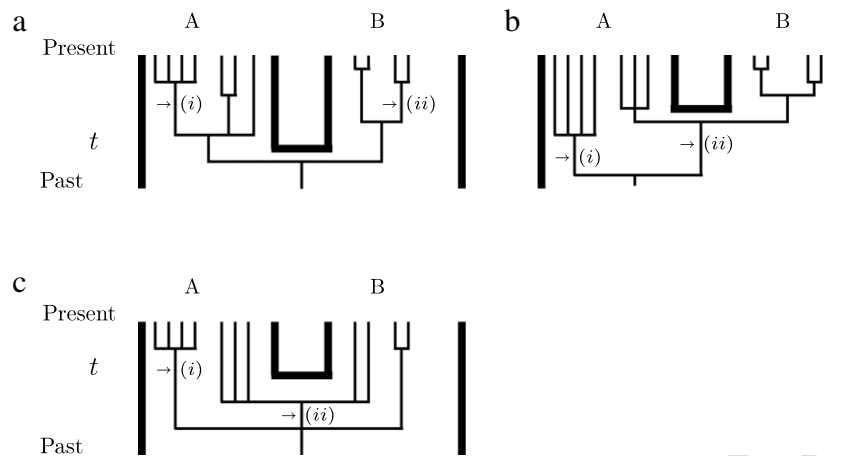


Fig. 9. Effects of (a) monophyly, (b) paraphyly, and (c) polyphyly on observed genetic variation in sequences drawn from populations A and B. Mutations are denoted by arrows (→). In (a), mutations (i) and (ii) are polymorphic in both populations A and B, respectively. In (b), mutation (i) is polymorphic in A, while mutation (ii) is polymorphic in A but fixed in B. In (c), mutation (i) is polymorphic in A while mutation (ii) is polymorphic in both A and B.

the biology of the organism in question, in addition to comparison of multi-loci genetic data to different models. The present focus is on two special cases (Schweinsberg, 2003; Eldon and Wakeley, 2006) of a Λ coalescent (Pitman, 1999) that have been derived from specific population models. Both processes introduce an extra parameter, which can be estimated from data (Birkner and Blath, 2008; Eldon, 2011). As Table 2 shows, the two processes are quite different, with implications for inference.

3.2. Monophyly and units of time

In many cases, one may wish to compare populations with different coalescent processes, or at least with different parameter values of the same coalescent process. Different Beta-coalescents run on different timescales, as time scales proportionally to $N^{\alpha-1}$ (Schweinsberg, 2003). By way of example, let $\alpha_{AB} < \min(\alpha_A, \alpha_B)$. If the time of divergence is taken in units according to α_{AB} , then the coalescent processes in the descendent populations are running on longer timescales, leading to low frequencies of monophyly. On the other hand, if $\alpha_A < \alpha_B$, say, and one scales the divergence time according to α_B , then all the A lineages will have coalesced before time of divergence. In that case, drawing more sequences from the A population will add little. The interpretation of Fig. 4 must be done with the timescale property of α in mind.

In contrast, ψ is not a timescale parameter. The probabilities of monophyly, paraphyly, and polyphyly for different values of ψ can therefore be directly compared. The probability of monophyly increases quickly with values of ψ in the descendent populations when all populations have the ψ -coalescent. The probability of monophyly is almost invariant to changes in ψ in the ancestral population. Thus, the effects of coalescent parameters on the probability of monophyly depends on the coalescent process, and if the population is ancestral or derived. Obtaining estimates of ψ for the derived populations will therefore be important for accurately computing probabilities of monophyly.

Distinguishing between the Kingman and Λ coalescents will be important for determining the unit of time of divergence. The timescales of both Beta- and ψ -coalescents are shorter than the usual Fisher–Wright timescale. As will probably frequently be the case, one may wish to apply computations for two species to data from closely related taxa, which will then most likely have very similar coalescent processes. However, should one wish to apply these computations to data from populations with very different coalescent processes, then the issue of the unit of time of divergence becomes central to the inference.

3.3. Paraphyly and polyphyly

Genetic and phenotypic datasets sometimes give different results in terms of monophyly or paraphyly of groups of taxa. One example is the evolutionary relationship of lampreys and hagfish to other chordates (Stock and Whitt, 1992; Forey and Janvier, 1993; Meyer, 1996). The effect of coalescent parameters on the probability of paraphyly of ancestral lineages depends on the coalescent process (Fig. 8). If monophyly is repeatedly observed among two populations, they will have been separated by at least two coalescence-time units, if all populations have the Beta-coalescent. On the other hand, if polyphyly is consistently observed, then the time of divergence was quite recent. Similar conclusions hold for paraphyly.

In the context of inference, it is helpful to keep in mind the effects of paraphyly, polyphyly, and monophyly on patterns of genetic diversity observed in the ancestral populations (Fig. 9). If the time of divergence was very recent, then polyphyly will be frequent, and many sites will be either fixed or polymorphic in both populations (Fig. 9c). If the time of divergence was more ancestral, then any site experiencing mutation more ancestral in time than the most recent common ancestor of the derived monophyletic population may be polymorphic in the derived population displaying paraphyly (Fig. 9b). In ψ -populations, most sites will be either fixed or polymorphic in both populations if the time of divergence was recent. Even if time of divergence was more ancient (in ψ -coalescence units) polyphyly may still frequently occur if ψ is low.

3.4. Conclusions

Probabilities of monophyly, paraphyly, and polyphyly for two species are computed by recursion for populations with coalescent processes that allow multiple mergers of ancestral lineages. The effects of coalescent parameters on these probabilities depend on the coalescent process and the population. The timescale of the time of divergence becomes a key issue when different populations have different coalescent processes running on different timescales. By estimating coalescent parameters, one should be able to distinguish between recent and ancient divergence times, at least when the timescale is a function of the same coalescent parameters.

Our calculations have shown that monophyly for a subset of lineages sampled from one panmictic population can be less likely under the α and ψ coalescent processes than under the

Kingman coalescent. However, even for quite recent divergences, reciprocal monophyly is likely to be achieved in fewer generations under these multiple merger coalescent processes than under the Kingman coalescent when there has been some divergence without subsequent gene flow. Consequently, the amount of incomplete lineage sorting (ILS) among lineages failing to coalesce more recently than the species divergence (Degnan and Rosenberg, 2009) is likely to be reduced. This suggests that estimates of population divergence times measured in years or generations would typically be shorter under a multiple merger coalescent than under a Kingman coalescent given the same observed levels of monophyly.

Acknowledgments

The present work was partially supported by Marsden fund to J.D., and by a Royal Society travel grant to the authors. J. D. was also supported by a **Sabbatical Fellowship at the National Institute for Mathematical and Biological Synthesis**, an Institute sponsored by the National Science Foundation, the **US Department of Homeland Security**, and the **US Department of Agriculture** through NSF Award #EF-0832858, with additional support from The **University of Tennessee, Knoxville**. B.E. was supported by EPSRC grant EP/G052026/1.

Appendix

A.1. The rate matrix of a Λ coalescent

Let $(A_r(t); t \geq 0)$ denote the gene genealogical process in a single population starting from $A_r(0) = r$, and **stopping** at the time $\inf\{t \geq 0 : A_r(t) = 1\}$. The pure-death process $A_r(t)$ counts the number of active ancestral lineages, and therefore has state space $[r] \equiv \{1, 2, \dots, r\}$. The infinitesimal generator $Q = (q_{i,j})$ of $A_r(t)$ is

$$q_{i,j} = \begin{cases} \binom{i}{i-j+1} \int_0^1 x^{i-j+1} (1-x)^{j-1} x^{-2} \Lambda(dx) & \text{if } 1 \leq j < i \\ -q_{i,1} - \dots - q_{i,i-1} & \text{if } j = i \\ 0 & \text{otherwise} \end{cases} \quad (12)$$

in which Λ is a finite measure on the unit interval (Pitman, 1999). The rate at which k active lineages out of i coalesce, the coalescence rate, will be denoted by $\lambda_{i,k} = q_{i,i-k+1}$, $2 \leq k \leq i$.

A.2. Enumerating the paths of $A_r(t)$

One way to compute $g_{i,j,\pi}(t)$ is by enumerating all the paths of $A_r(t)$. In a Kingman coalescent, $A_r(t)$ visits every state from r to 1. In a Λ coalescent, however, $A_r(t)$ may go directly from r to 1 in a single merger. Indeed, the total number of paths $A_r(t)$ can take in going from i to $j < i$ is 2^{i-j-1} (the number of subsets of the integers from j to i that include i and j). Let $\mathcal{A}(i,j)$ denote the set of all paths $A_r(t)$ takes in going from i to $j \leq i$. Let $a = (a_1, \dots, a_\ell) \in \mathcal{A}(i,j)$ denote an element in $\mathcal{A}(i,j)$, where $i = a_1$ and ℓ denotes the number of coalescence events in path a . Let $p(a)$ be the probability of a , given by

$$p(a) = \frac{q_{i,a_2} q_{a_2,a_3} \dots q_{a_\ell,j}}{(-1)^\ell q_i q_{n_2} \dots q_{a_\ell}} \quad (13)$$

since the probability of a transition of $A_r(t)$ from i to $j < i$ is $q_{i,j}/(-q_i)$. Let $g_{i,j,\pi}(t)$ denote the probability $g_{i,j}(t)$ when associated with a Λ coalescent with parameter π . If associated with the Kingman coalescent, simply write $g_{i,j}(t)$. Let $T(a) \equiv T_{n_1} + \dots + T_{n_\ell}$ denote the sum of independent exponentials T_k

with rate $-q_k$. Each T_k denotes the time during which there are k active ancestral lineages. By $g_{i,j,\pi}(t, a)$ denote the probability $g_{i,j,\pi}(t)$ conditional on a , given by

$$g_{i,j,\pi}(t, a) = \begin{cases} \mathbb{P}[T(a) \leq t, T(a) + T_j > t] & \text{if } 2 \leq j < i \\ \mathbb{P}[T(a) \leq t] & \text{if } j = 1 \\ e^{q_i t} & \text{if } j = i. \end{cases} \quad (14)$$

The probability $g_{i,j,\pi}(t)$ is now given by

$$g_{i,j,\pi}(t) = \sum_{a \in \mathcal{A}(i,j)} g_{i,j,\pi}(t, a) p(a). \quad (15)$$

A.3. The spectral decomposition of the rate matrix

Under a Λ coalescent, the rate matrix (12) is triangular. The left and right eigenvectors are therefore obtained recursively. Let $\ell^{(k)} = (\ell_1^{(k)}, \dots, \ell_n^{(k)})$ and $r^{(k)} = (r_1^{(k)}, \dots, r_n^{(k)})$ denote the left and right eigenvectors, respectively, corresponding to eigenvalue $\lambda_k = q_k$ for $k \geq 1$ with $\lambda_1 = 0$. Then $\ell_j^{(1)} = \delta_{1j}$, $\ell_j^{(k)} = 0$ if $j > k$, $\ell_k^{(k)} = 1$, and

$$\ell_j^{(k)} = \frac{q_{j+1,j} \ell_{j+1}^{(k)} + \dots + q_{k,j} \ell_k^{(k)}}{q_k - q_j}, \quad 1 \leq j < k. \quad (16)$$

For the right eigenvectors we have $r_j^{(1)} = 1$, $r_k^{(k)} = 1$, $r_j^{(k)} = 0$ if $j < k$, and

$$r_j^{(k)} = \frac{q_{j,k} r_k^{(k)} + \dots + q_{j,j-1} r_{j-1}^{(k)}}{q_k - q_j}, \quad 1 < k < j \leq n. \quad (17)$$

One confirms that, for sample size two, $\ell^{(1)} = (1, 0)$, $\ell^{(2)} = (-1, 1)$, $r^{(1)} = (1, 1)$, $r^{(2)} = (0, 1)$, yielding $\mathbb{P}[A_2(t) = 1] = 1 - \mathbb{P}[A_2(t) = 2] = 1 - e^{q_2 t}$.

The probability $g_{i,j}(t)$ can now be computed as

$$g_{i,j}(t) = \sum_{k=j}^i e^{t q_k} r_i^{(k)} \ell_j^{(k)}, \quad 1 \leq j \leq i. \quad (18)$$

If R denotes a matrix whose columns are the right eigenvectors of Q , D is a diagonal matrix whose entries are all zero except the diagonal which contains the eigenvalues of Q , and $L = R^{-1}$ whose rows contain the left eigenvectors of Q , then $Q = RDL$.

A.4. Sum of clade probabilities

Here we show that under a multiple merger coalescent model for rooted gene trees, the sum of the clade probabilities is less than $n - 2$. The approach is similar to that used for binary trees (Allman et al., 2011). We use C_i to denote an arbitrary clade on n taxa (i.e., nontrivial subset of the taxa), and T_j to denote a rooted tree which can be either binary or nonbinary. We note that $\mathbb{P}(C_i|T_j) = 1$ if clade C_i is a clade on tree T_j ; otherwise, $\mathbb{P}(C_i|T_j) = 0$. Thus $\sum_i \mathbb{P}(C_i|T_j)$ counts the number of clades on tree T_j . Thus,

$$\begin{aligned} \sum_i \mathbb{P}(C_i) &= \sum_i \sum_{T_j} \mathbb{P}(C_i|T_j) \mathbb{P}(T_j) \\ &= \sum_{T_j} \mathbb{P}(T_j) \sum_i \mathbb{P}(C_i|T_j) \\ &= \sum_{T_j \text{ binary}} \mathbb{P}(T_j) \sum_i \mathbb{P}(C_i|T_j) \\ &\quad + \sum_{T_j \text{ nonbinary}} \mathbb{P}(T_j) \sum_i \mathbb{P}(C_i|T_j) \\ &< \sum_{T_j \text{ binary}} (n-2) \mathbb{P}(T_j) + \sum_{T_j \text{ nonbinary}} (n-2) \mathbb{P}(T_j) \\ &= n - 2 \end{aligned}$$

where the inequality is due to nonbinary trees having fewer than $n - 2$ clades.

A.5. Sum of coalescence rates

The sum $\lambda_b \equiv \lambda_{b,2} + \dots + \lambda_{b,b}$ of the coalescence rates $\lambda_{b,k}$ associated with the Beta-coalescent and given by Eq. (1) increases with α for any $b \geq 3$, as we now show.

At first, time is measured in coalescent units, i.e. units of $N^{\alpha-1}$ generations. It is straightforward to show that λ_b increases with α for $b = 3, 4$. Assume now that λ_b increases with α for some b . We start with the representation

$$\lambda_b = \int_0^1 (1 - (1-x)^b - bx(1-x)^{b-1}) x^{-2} \Lambda(dx) \quad (19)$$

in which

$$\Lambda(dx) = (B(2 - \alpha, \alpha))^{-1} x^{1-\alpha} (1-x)^{\alpha-1} dx$$

and $B(v, w) = \Gamma(v)\Gamma(w)/\Gamma(v+w)$ denotes the beta function. By rewriting the integrand in Eq. (19) for $b+1$ one obtains

$$\begin{aligned} 1 - (1-x)^{b+1} - (1+b)x(1-x)^b \\ &= 1 - (1-x)^b(1-x) - x(1-x)^b - bx(1-x)^b \\ &= 1 - (1-x)^b - bx(1-x)^b \\ &= 1 - (1-x)^b - bx(1-x)^{b-1} + bx^2(1-x)^{b-1}. \end{aligned}$$

The representation (19) now gives

$$\lambda_{b+1} = \lambda_b + b \int_0^1 (1-x)^{b-1} \Lambda(dx).$$

The properties of the beta and gamma functions finally yield

$$b \int_0^1 (1-x)^{b-1} \Lambda(dx) = (b-2+\alpha)(b-3+\alpha) \dots \alpha$$

and which is clearly increasing in α .

Although the expected waiting time $\mathbb{E}[T_b] = 1/\lambda_b$ until the next coalescence with time measured in coalescent units, $1/\lambda_b$, is decreasing in α , $\mathbb{E}[T_b]$ with time measured in generations is increasing in α . The expected waiting time in generations is $N^{\alpha-1}/\lambda_b$. We therefore wish to show that $\lambda_b N^{1-\alpha}$ is decreasing in α . This is straightforward to verify for $b = 3$, and we use induction for $b > 3$. To check that $\lambda_{b+1} N^{1-\alpha}$ is decreasing in α if $\lambda_b N^{1-\alpha}$ is decreasing in α , we let

$$f_b(\alpha) := (b-2+\alpha)(b-3+\alpha) \dots \alpha = \prod_{j=0}^{b-2} (\alpha+j).$$

It is sufficient to show that $f_b(\alpha) N^{1-\alpha}$ is decreasing in α . One obtains

$$f'_b(\alpha) = \sum_{i=0}^{b-2} \prod_{\substack{j=0 \\ j \neq i}}^{b-2} (\alpha+j), \quad \text{and}$$

$$\frac{f'_b(\alpha)}{f_b(\alpha)} = \sum_{j=0}^{b-2} \frac{1}{\alpha+j} < \sum_{j=1}^{b-1} \frac{1}{j} < 1 + \log(b-1).$$

We can check when

$$\frac{d}{d\alpha} f_b(\alpha) N^{1-\alpha} = f'_b(\alpha) N^{1-\alpha} - f_b(\alpha) N^{1-\alpha} \log(N)$$

is less than 0 to find that $\lambda_b N^{1-\alpha}$ is decreasing in α when

$$N > \exp\left(\sum_{j=0}^{b-2} \frac{1}{\alpha+j}\right),$$

which can be seen to be satisfied when $N > e \cdot (b-1)$. Since b is the number of lineages that can merge, the result therefore holds as long as the sample size is not a large fraction of the population size N .

References

- Allman, E.S., Degnan, J.H., Rhodes, J.A., 2011. Determining species tree topologies from clade probabilities under the coalescent. *J. Theoret. Biol.* 289, 96–106.
- Árnason, E., 2004. Mitochondrial cytochrome *b* variation in the high-fecundity Atlantic cod: trans-Atlantic clines and shallow gene genealogy. *Genetics* 166, 1871–1885.
- Avise, J.C., 1989. Gene trees and organismal histories: a phylogenetic approach to population biology. *Evolution* 43, 1192–1208.
- Beckenbach, A.T., 1994. Mitochondrial haplotype frequencies in oysters: neutral alternatives to selection models. In: Golding, B. (Ed.), *Non-Neutral Evolution*. Chapman & Hall, New York, pp. 188–198.
- Birkner, M., Blath, J., 2008. Computing likelihoods for coalescents with multiple collisions in the infinitely many sites model. *J. Math. Biol.* 57, 435–465.
- Boudry, P., Collet, B., Cornette, F., Hervouet, V., Bonhomme, F., 2002. High variance in reproductive success of the Pacific oyster (*Crassostrea gigas*, Thunberg) revealed by microsatellite-based parentage analysis of multifactorial crosses. *Aquaculture* 204, 283–296.
- Cannings, C., 1974. The latent roots of certain Markov chains arising in genetics: a new approach, I. Haploid models. *Adv. Appl. Probab.* 6, 260–290.
- Cummings, M.P., Otto, S.P., Wakeley, J., 1995. Sampling properties of DNA sequence data in phylogenetic analysis. *Mol. Biol. Evol.* 12, 814–822.
- Degnan, J.H., Rosenberg, N.A., 2009. Gene tree discordance, phylogenetic inference, and the multispecies coalescent. *Trends Ecol. Evol.* 24, 332–340.
- Donnelly, P., Kurtz, T.G., 1999. Particle representations for measure-valued population models. *Ann. Probab.* 27, 166–205.
- Donnelly, P., Tavaré, S., 1995. Coalescents and genealogical structure under neutrality. *Annu. Rev. Genet.* 29, 401–421.
- Doyle, J.J., 1992. Gene trees and species trees: molecular systematics as one-character taxonomy. *Syst. Bot.* 17, 144–163.
- Eldon, B., 2009. Structured coalescent processes from a modified Moran model with large offspring numbers. *Theor. Popul. Biol.* 76, 92–104.
- Eldon, B., 2011. Estimation of parameters in large offspring number models and ratios of coalescence times. *Theor. Popul. Biol.* 80, 16–28.
- Eldon, B., Wakeley, J., 2006. Coalescent processes when the distribution of offspring number among individuals is highly skewed. *Genetics* 172, 2621–2633.
- Fisher, R.A., 1930. *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford.
- Flowers, J.M., Schroeter, S.C., Burton, R.S., 2002. The recruitment sweepstakes has many winners: genetic evidence from the sea urchin *Strongylocentrotus purpuratus*. *Evolution* 56, 1445–1453.
- Forey, P., Janvier, P., 1993. Agnathans and the origin of jawed vertebrates. *Nature* 361, 129–134.
- Griffiths, R.C., 1979. Exact sampling distributions from the infinite neutral alleles model. *Adv. Appl. Probab.* 11, 326–354.
- Hedgecock, D., 1994. Does variance in reproductive success limit effective population sizes of marine organisms? In: Beaumont, A. (Ed.), *Genetics and Evolution of Aquatic Organisms*. Chapman and Hall, London, pp. 1222–1344.
- Hedgecock, D., Tracey, M., Nelson, K., 1982. *Genetics*. In: Abele, L.G. (Ed.), *The Biology of Crustacea*, 2. Academic Press, New York, pp. 297–403.
- Hudson, R.R., 1990. Gene genealogies and the coalescent. In: Futuyma, D.J., Antonovics, J. (Eds.), *Oxford Surveys in Evolutionary Biology*, 7. Oxford University Press, Oxford, pp. 1–44.
- Hudson, R.R., 1983. Testing the constant-rate neutral allele model with protein sequence data. *Evolution* 37, 203–217.
- Hudson, R.R., Coyne, J.A., 2002. Mathematical consequences of the genealogical species concept. *Evolution* 56, 1557–1665.
- Ingvarsson, P.K., 2010. Nucleotide polymorphism, linkage disequilibrium and complex trait dissection in *Populus*. In: Jansson, S., Bhalerao, R., Groover, A. (Eds.), *Genetics and Genomics of Populus*. In: *Plant Genetics and Genomics: Crops and Models*, vol. 8. Springer, New York, pp. 91–111.
- Kingman, J.F.C., 1982a. The coalescent. *Stoch. Proc. Appl.* 13, 235–248.
- Kingman, J.F.C., 1982b. Exchangeability and the evolution of large populations. In: Koch, G., Spizzichino, F. (Eds.), *Exchangeability in Probability and Statistics*. North-Holland, Amsterdam, pp. 97–112.
- Kingman, J.F.C., 1982c. On the genealogy of large populations. *J. Appl. Probab.* 19A, 27–43.
- Knowles, L.L., Carstens, B.C., 2007. Delimiting species without monophyletic gene trees. *Syst. Biol.* 56, 887–895.
- Liu, L., Pearl, D.K., 2007. Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Syst. Biol.* 56, 504–514.
- Liu, L., Yu, L., Pearl, D.K., Edwards, S.V., 2009. Estimating species phylogenies using coalescence times among sequences. *Syst. Biol.* 58, 468–477.
- Maddison, W.P., 1997. Gene trees in species trees. *Syst. Biol.* 46, 523–536.
- McKenzie, A., Steel, M., 2000. Distributions of cherries for two models of trees. *Math. Biosci.* 164, 81–92.
- Meyer, A., 1996. The evolution of body plans: HOM/Hox Cluser evolution, model systems, and the importance of phylogeny. In: Harvey, P.H., Brown, A.J.L., Smith, J.M., Nee, S. (Eds.), *New Uses for New Phylogenies*. Oxford University Press, Oxford, pp. 323–340.
- Möhle, M., 2000. Ancestral processes in population genetics: the coalescent. *J. Theor. Biol.* 204, 629–638.

- Möhle, M., Sagitov, S., 2001. A classification of coalescent processes for haploid exchangeable population models. *Ann. Probab.* 29, 1547–1562.
- Moran, P.A.P., 1958. Random processes in genetics. *Proc. Camb. Philos. Soc.* 54, 60–71.
- Moran, P.A.P., 1962. *Statistical Processes of Evolutionary Theory*. Clarendon Press, Oxford.
- Nee, S., Holmes, E.C., Rambaut, A., Harvey, P.H., 1996. Inferring population histories from molecular phylogenies. In: Harvey, P.H., Brown, A.J.L., Smith, J.M., Nee, S. (Eds.), *New Uses for New Phylogenies*. Oxford University Press, Oxford, pp. 66–80.
- Nei, M., 1986. Stochastic errors in DNA evolution and molecular phylogeny. In: Gershowitz, H., Rucknagel, D.L., Tashian, R.E. (Eds.), *Evolutionary Perspectives and the New Genetics*. A R Liss, New York, pp. 133–147.
- Neigel, J.E., Avise, J.C., 1986. Phylogenetic relationships of mitochondrial DNA under various demographic models of speciation. In: Karlin, S., Nevo, E. (Eds.), *Evolutionary Processes and Theory*. Academic Press, New York, pp. 515–534.
- Nichols, R., 2001. Gene trees and species trees are not the same. *Trends Ecol. Evol.* 16, 358–364.
- Nordborg, M., 2001. Coalescent theory. In: Balding, D.J., Bishop, M.J., Cannings, C. (Eds.), *Handbook of Statistical Genetics*. John Wiley & Sons, Chichester, England, pp. 179–212.
- Otto, S.P., Cummings, M.P., Wakeley, J., 1996. Inferring phylogenies from DNA sequence data: the effects of sampling. In: Harvey, P.H., Brown, A.J.L., Smith, J.M., Nee, S. (Eds.), *New Uses for New Phylogenies*. Oxford University Press, Oxford, pp. 103–115.
- Palumbi, S.R., Cipriano, F., Hare, M.P., 2001. Predicting nuclear gene coalescence from mitochondrial data: the three-times rule. *Evolution* 55, 859–868.
- Petersen, J.L., Ibarra, A.M., Ramirez, J.L., May, B., 2008. An induced mass spawn of the hermaphroditic Lion-Paw scallop *Nodipecten subnodosus*: Genetic assignment of maternal and paternal parentage. *J. Heredity* 99, 337–348.
- Pitman, J., 1999. Coalescents with multiple collisions. *Ann. Probab.* 27, 1870–1902.
- Rosenberg, N.A., 2003. The shapes of neutral gene genealogies in two species: probabilities of monophyly, paraphyly, and polyphyly in a coalescent model. *Evolution* 57, 1465–1477.
- Rosenberg, N.A., 2007. Counting coalescent histories. *J. Comp. Biol.* 14, 360–377.
- Rosenberg, N.A., Nordborg, M., 2002. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nat. Rev. Genet.* 3, 380–390.
- Ruvolo, M., 1994. Molecular evolutionary processes and conflicting gene trees: the Hominoid case. *Am. J. Phys. Anthropol.* 94, 89–113.
- Sagitov, S., 1999. The general coalescent with asynchronous mergers of ancestral lines. *J. Appl. Probab.* 36, 1116–1125.
- Sargsyan, O., Wakeley, J., 2008. A coalescent process with simultaneous multiple mergers for approximating the gene genealogies of many marine organisms. *Theor. Popul. Biol.* 74, 104–114.
- Satta, Y., Klein, J., Takahata, N., 2000. DNA archives and our nearest relative: the trichotomy problem revisited. *Mol. Phylogenet. Evol.* 14, 259–275.
- Schweinsberg, J., 2003. Coalescent processes obtained from supercritical Galton–Watson processes. *Stoch. Proc. Appl.* 106, 107–139.
- Schweinsberg, J., 2000a. Coalescents with simultaneous multiple collisions. *Electron. J. Probab.* 5, 1–50.
- Schweinsberg, J., 2000b. A necessary and sufficient condition for the Λ -coalescent to come down from infinity. *Electron. Comm. Probab.* 5, 1–11.
- Stock, D.W., Whitt, G.S., 1992. Evidence from 18s ribosomal RNA sequences that lampreys and hagfishes form a natural group. *Science* 257, 787–789.
- Tavaré, S., 1984. Lines-of-descent and genealogical processes, and their application in population genetic models. *Theor. Popul. Biol.* 26, 119–164.
- Ting, C., Tsaur, S., Wu, C., 2000. The phylogeny of closely related species as revealed by the genealogy of a speciation gene, *Odysseus*. *Proc. Natl. Acad. Sci. USA* 97, 5313–5316.
- Wakeley, J., 2009. *Coalescent Theory*. Roberts & Company, Greenwood Village, Colorado.
- Wakeley, J., Hey, J., 1997. Estimating ancestral population parameters. *Genetics* 145, 847–855.
- Waters, J.M., Roy, M.S., 2004. Phylogeography of a high-dispersal New Zealand sea star: does upwelling block gene-flow? *Mol. Ecol.* 13, 2797–2806.
- Watterson, G.A., 1982. Mutant substitutions at linked nucleotide sites. *Adv. Appl. Probab.* 14, 206–224.
- Wright, S., 1931. Evolution in Mendelian populations. *Genetics* 16, 97–159.
- Zhu, S., Degnan, J.H., Steel, M., 2011. Clades, clans, and reciprocal monophyly under neutral evolutionary models. *Theor. Popul. Biol.* 79, 220–227.