J. Appl. Prob. **43,** 351–362 (2006)
Printed in Israel

STRUCTURED COALESCENT WITH NONCONSERVATIVE MIGRATION

KOFFI Y. SAMPSON,* *Florida State University*

Abstract

We study the ancestral process of a sample from a subdivided population with stochastically varying subpopulation sizes. The sizes of the subpopulations change very rapidly (almost every generation) with respect to the coalescent time scale. For haploid populations of size N, one coalescence time unit corresponds to N generations. Coalescence and migration events occur on the same time scale. We show that, when the total population size tends to infinity, the structured coalescent is obtained, thus confirming the robustness of the coalescent. Many population structure models have been shown to converge to the structured coalescent (see Herbots (1997), Hudson (1998), Nordborg (2001), Nordborg and Krone (2002), and Notohara (1990)).

Keywords: Structured coalescent; stochastic population size; nonconservative migration; weak convergence

2000 Mathematics Subject Classification: Primary 92D10; 60F05; 60J70
Secondary 92D25

1. Introduction

The n-coalescent or simply the coalescent, also called Kingman's coalescent, is a continuous-time Markov process (see Kingman (1982a), (1982b), (1982c), Tajima (1983), Hudson (1990), and Tavaré (1984)) that describes the ancestry of a sample of n individuals or genes in a large population when time is counted backward from the present into the past. It has been extended to include such biological phenomena as mutation (see Donnelly and Tavaré (1995)), recombination (see Griffiths and Marjoram (1996), Hudson and Kaplan (1988), and Hey and Wakeley (1997)), selection (see Kaplan *et al.* (1988) and Neuhauser and Krone (1997)), populations with a mixture of self-fertilization and random mating (see Fu (1997), Nordborg and Donnelly (1997), and Möhle (1998a)), models with variable population size (see Donnelly and Tavaré (1995), Tajima (1989), Griffiths and Tavaré (1994), Möhle (2002), and Sano *et al.* (2004)), and diploid and two-sex population models, relaxing Kingman's assumption that the population must be haploid (see Möhle (1998b) and Möhle and Sagitov (2003)).

The n-coalescent has also been expanded to subdivided population models and geographically structured models which require an approximation by the structured coalescent, a generalization of Kingman's coalescent (see Notohara (1990), Takahata (1991), Herbots (1994), Wilkinson-Herbots (1998), Hudson (1998), Bahlo and Griffiths (2000), Beerli and Felsenstein (2001), Wakeley (2000), and Nordborg (2001)). The structured coalescent can be applied to other biological situations, for example populations with several forms of selection (see Nordborg (1997)) and populations with partial selfing and balancing selection (see Nordborg (1999)).

Received 19 September 2005; revision received 3 February 2006.
* Postal address: School of Computational Science, Florida State University, 150B Dirac Science Library, Tallahassee, FL 32306-4120, USA. Email address: sampson@csit.fsu.edu

The seed bank model is another example of the structured coalescent (see Kaj *et al.* (2001)). Here the population is formed in each generation from the ancestors belonging to the previous m generations (m is fixed).

A generalized population structure model (see Nordborg and Krone (2002)) has been developed where some migrations occur on a time scale that is much shorter than the coalescent time scale, leading to a separation of time scales.

We are proposing in this paper an island model with stochastically varying population size and nonconservative migration. Nonconservative migration means that the number of lineages migrating out of a subpopulation is not equal all the time to the number of lineages migrating into that subpopulation. Our model is different from the existing models. Although it can be viewed as a generalized population structure model (see Nordborg and Krone (2002)), where some migrations occur much faster than the coalescence events, the subpopulation sizes are not constant over time as proposed in Nordborg and Krone's model. Any model with fixed subpopulation sizes that is equivalent to our model would have the following characteristics. Firstly, the individual lineages would not always 'migrate' backward independently of one another; rather, some groups of lineages would 'migrate' together at the same time to the same subpopulations. Also, subpopulation sizes would be alternating between zero and their actual sizes (a contradiction); in fact, all subpopulation sizes but two would be zero in any given generation. In addition, the pairwise coalescence rates for our model are different from the ones obtained by Nordborg and Krone (2002).

The coalescent is a continuous-time Markov process that plays an important role in population genetics. It is used to explore the genealogy of a sample of n individuals from a sufficiently large population, starting from a particular time (called time 0) and going backward in time until the first common ancestor (most recent common ancestor) of the sample is reached. It serves as a continuous-time approximation for the ancestral structure of a variety of discrete-time population models. A discrete-time ancestral process is a population model that counts the number of ancestral lineages of the original sample in each generation in the past. The basic idea is that the population size is fixed at N. Then, when time is measured in units of N generations, the discrete-time process converges to the coalescent as N tends to infinity. This idea is generalized to structured populations.

Our model is a discrete-time Markov chain. It is based on the Wright–Fisher type of reproduction with some added population structure.

The Wright–Fisher model considers a haploid population (meaning that each individual of the population in any given generation has exactly one parent in the previous generation) of fixed size N for all generations. The population evolves in discrete nonoverlapping generations and the reproduction is neutral (i.e. there is no selection). An equivalent description of the Wright–Fisher model is that, when we look at the process backward in time, individuals are assigned parents in the previous generation randomly and independently of each other and these choices are independent from generation to generation.

The rest of this paper is organized as follows: we introduce the structured coalescent and then describe our model; this is followed by the main result, Theorem 3.1, where we consider a population subdivided in $M \geq 1$ subpopulations; finally, a formal proof of Theorem 3.1 is given.

2. Structured coalescent

Consider a haploid population subdivided into M subpopulations of fixed sizes $N_k = a_k N$, $k = 1, \ldots, M$. Let $b_{k\ell}$ be the backward migration probability that a lineage currently in

https://doi.org/10.1239/jap/1152413727 Published online by Cambridge University Press

subpopulation k was produced in subpopulation ℓ in the previous generation, and suppose that $\beta_{k\ell} := \lim_{N\to\infty} 2Nb_{k\ell}$, $k, \ell = 1, \ldots, M$. Define $\alpha_k := 1/a_k$, $k = 1, \ldots, M$. Consider a sample of size n and let $\boldsymbol{Y}_N(\tau)$ denote the ancestral process that counts the number of ancestral lineges of the sample that are in each subpopulation in generation τ in the past, with $\tau = 0$ corresponding to the current generation (the subscript N in $\boldsymbol{Y}_N(\tau)$ refers to the magnitude of the subpopulations). The process $(\boldsymbol{Y}_N(\tau))_{\tau\in\mathbb{N}}$ has state space

$$E = \left\{ \boldsymbol{r} := (r_1, \ldots, r_M) \in \mathbb{N}^M : 1 \le \sum_{i=1}^{M} r_i \le n \right\}, \tag{2.1}$$

where $\mathbb{N} = \{0, 1, 2, \ldots\}$ and r_k is the number of lineages in subpopulation k. Note that

$$|E| = \binom{n+M}{M} - 1.$$

Let $\boldsymbol{e}_k$, $k = 1, 2, \ldots, M$, be the unit vector in $\mathbb{N}^M$ whose kth component is equal to 1, let $[\cdot]$ denote the integer-part function, and let $D_E[0,\infty)$ denote the space of right-continuous functions f from $[0,\infty)$ to E with left limits (that is, $\lim_{s\to t^+} f(s) = f(t)$ for all $t \ge 0$ and $\lim_{s\to t^-} f(s)$ exists for all $t > 0$). Under reasonable assumptions (see Herbots (1997) and Nordborg (2001)), as $N \to \infty$ the time-scaled ancestral process $\boldsymbol{Y}_N := (\boldsymbol{Y}_N([Nt]))_{t\ge0}$ converges weakly in $D_E[0,\infty)$ to a structured coalescent, $\boldsymbol{Y} := (\boldsymbol{Y}(t))_{t\ge0}$, in which each pair of lineages in subpopulation k coalesces independently at rate α_k and each lineage in k migrates (backward in time) independently to ℓ at rate $\beta_{k\ell}/2$. More precisely, $\boldsymbol{Y}$ is a continuous-time Markov chain with state space E and infinitesimal generator $\boldsymbol{Q} = (q_{\boldsymbol{r},\boldsymbol{r}'})_{\boldsymbol{r},\boldsymbol{r}'\in E}$, where

$$q_{\boldsymbol{r},\boldsymbol{r}'} = \begin{cases} -\displaystyle\sum_{k=1}^{M} \left(\alpha_k \binom{r_k}{2} + r_k \sum_{\substack{\ell=1 \\ \ell\neq k}}^{M} \frac{\beta_{k\ell}}{2} \right) & \text{if } \boldsymbol{r}' = \boldsymbol{r}, \\ \dfrac{r_k \beta_{k\ell}}{2} & \text{if } \boldsymbol{r}' = \boldsymbol{r} - \boldsymbol{e}_k + \boldsymbol{e}_\ell,\ k \neq \ell, \\ \alpha_k \dbinom{r_k}{2} & \text{if } \boldsymbol{r}' = \boldsymbol{r} - \boldsymbol{e}_k, \\ 0 & \text{otherwise.} \end{cases} \tag{2.2}$$

Remark 2.1. Equation (2.2) is valid for some diploid populations (for example monoecious diploid populations), but we need to set $N_k = 2a_kN$ and $\beta_{k\ell} := \lim_{N\to\infty} 4Nb_{k\ell}$ and replace $(\boldsymbol{Y}_N([Nt]))_{t\ge0}$ by $(\boldsymbol{Y}_N([2Nt]))_{t\ge0}$ in the paragraph above.

In the next section, we present our model, which is an extension of the model described by Nordborg (2001).

3. Structured coalescent with nonconservative migration

In many population genetics processes, population sizes actually vary in time. But in most models, the population sizes are assumed to be fixed or evolving deterministically (see Möhle (2002), Griffiths and Tavaré (1994), and Donnelly (1986)). The case where the population size varies stochastically has been studied only for unstructured populations (see Sano *et al.* (2004), Kaj and Krone (2003), and Jagers and Sagitov (2004)). Here we present a structured coalescent model that takes into account this stochastic variation of population sizes.

https://doi.org/10.1239/jap/1152413727 Published online by Cambridge University Press

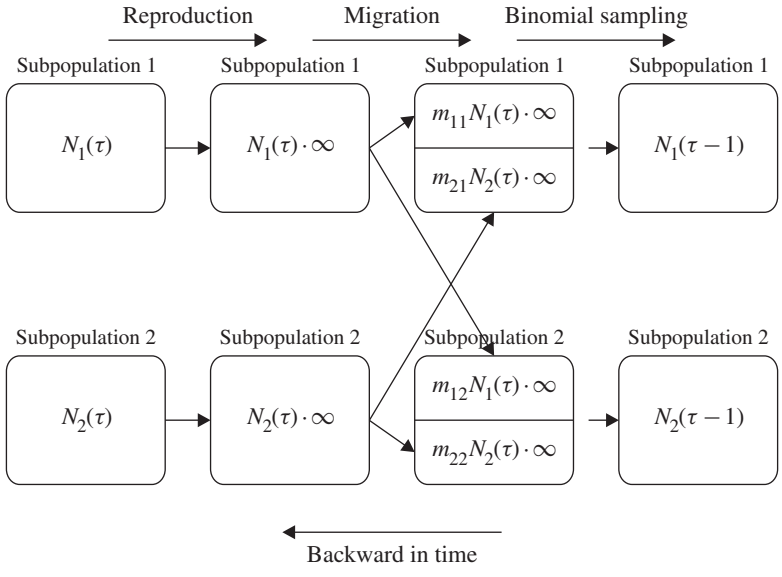

FIGURE 1: Pictorial description of our model for $M = 2$.

3.1. Description of the model

3.1.1. *Forward dynamics.* Consider a haploid population of variable size divided into $M \geq 1$ subpopulations. The subpopulation sizes change stochastically according to an irreducible, aperiodic Markov chain with state space

$$S_N := \{\boldsymbol{a}_i N := (a_{i1}N, \ldots, a_{iM}N) : i \in \{1, \ldots, s\}\}, \tag{3.1}$$

where s and a_{ik} are given positive integer constants, $a_{ik}N$ is the (variable) size of subpopulation k, and $\boldsymbol{a}_i = (a_{i1}, \ldots, a_{iM})$. The integer N controls the subpopulation sizes, and so all subpopulation sizes become large if and only if N becomes large. The population evolves in discrete, nonoverlapping generations. In each generation, every member of the population produces an effectively infinite number of propagules. These propagules then migrate between the M subpopulations independently of one another. After the migration step, the next generation of adults in subpopulation k ($k = 1, \ldots, M$) is formed by selecting randomly and uniformly the appropriate number of propagules from the post-migration propagules in subpopulation k.

Next, for $k, \ell \in \{1, 2, \ldots, M\}$, let $m_{k\ell}$ be the probability that a propagule produced in subpopulation k moves to subpopulation ℓ in the next generation, and assume that the limits

$$\mu_{k\ell} := \lim_{N\to\infty} Nm_{k\ell}, \qquad k \neq \ell, \tag{3.2}$$

exist, i.e. $m_{k\ell} = \mu_{k\ell}/N + o(N^{-1})$ for $k \neq \ell$. So migration of an individual propagule is rare. Of course, $\sum_{\ell=1}^{M} m_{k\ell} = 1$, $k = 1, \ldots, M$.

Figure 1 gives a pictorial illustration of the model for $M = 2$, where $N_k(\tau)$, $k = 1, 2$, represents the number of adults in subpopulation k in generation τ in the past, the current generation corresponding to $\tau = 0$.

In the following section, we consider the backward dynamics of the size process.

https://doi.org/10.1239/jap/1152413727 Published online by Cambridge University Press

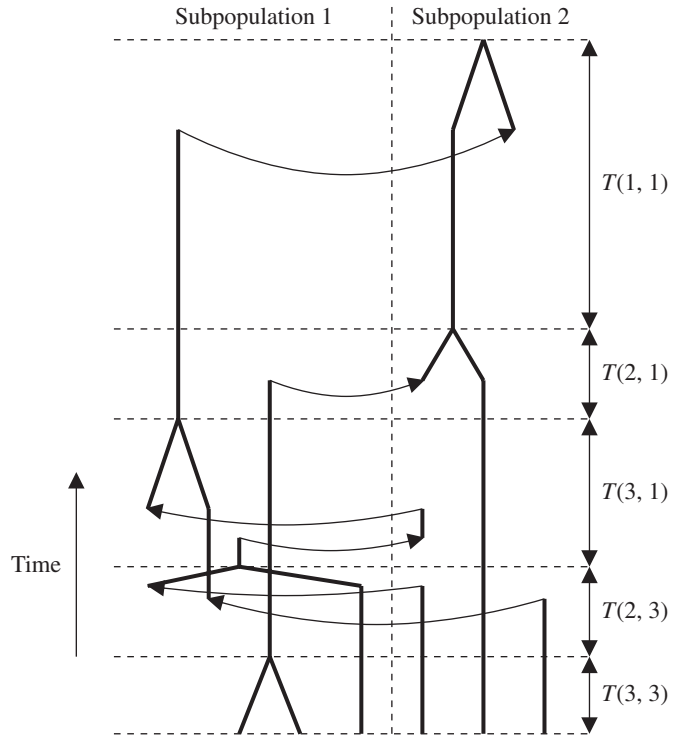

FIGURE 2: Coalescent with migration.

3.1.2. *Backward dynamics.* Let $\boldsymbol{N}(\tau) := (N_1(\tau), \ldots, N_M(\tau))$, where $N_k(\tau)$ is the size of subpopulation k, τ generations into the past. Assume that the backward size process $(\boldsymbol{N}(\tau))_{\tau\in\mathbb{N}}$ is a Markov chain with state space S_N defined in (3.1), one-step transition probability matrix $\boldsymbol{T} = (t_{ij})_{i,j=1,\ldots,s}$, and (unique) stationary distribution $\boldsymbol{\gamma} := (\gamma_1, \ldots, \gamma_s)$, that is,

$$t_{ij} = \mathrm{P}(\boldsymbol{N}(\tau+1) = \boldsymbol{a}_j N \mid \boldsymbol{N}(\tau) = \boldsymbol{a}_i N),$$
$$\boldsymbol{\gamma}\boldsymbol{T} = \boldsymbol{\gamma}.$$

This will be the case if, for example, the forward size process $(\tilde{\boldsymbol{N}}(z))_{z\in\mathbb{N}}$ is a stationary process with state space S_N, meaning that, for all nonnegative integers z, $\mathrm{P}(\tilde{\boldsymbol{N}}(z) = \boldsymbol{a}_j N) = \gamma_j$.

For the remainder of this paper, the Greek letter 'τ', whenever used, refers to the number of generations in the past, the present generation being 0.

We take a random sample of fixed size n from the current generation, $\tau = 0$, and we are interested in tracing the ancestry of the sample back to its most recent common ancestor.

Figure 2 is a simulation of our model for $M = 2$ and a sample of size 6 having three lineages in each subpopulation. Time is measured backward in units of N generations. A migration event is indicated by a thin solid line from one subpopulation to another. Whenever two ancestors merge into one, we have a coalescence event, depicted by two thick solid line-segments converging to a point. The quantities $T(3, 3)$, $T(2, 3)$, $T(3, 1)$, $T(2, 1)$, and $T(1, 1)$ are the coalescence times. For example, $T(2, 3)$ is the time required for a coalescence event to occur when there are two ancestors remaining in subpopulation 1 and three ancestors remaining in subpopulation 2.

https://doi.org/10.1239/jap/1152413727 Published online by Cambridge University Press

Let $Y_{N,k}(\tau)$ denote the number of ancestors of the sample in the kth subpopulation τ generations backward in time. Note that the ancestral process $(\boldsymbol{Y}_N(\tau))_{\tau\in\mathbb{N}}$, $\boldsymbol{Y}_N(\tau) := (Y_{N,1}(\tau), \ldots, Y_{N,M}(\tau))$, is a process with state space E defined by (2.1). We are now able to state our main convergence result. By '$\xrightarrow{\text{W}}$' we denote weak convergence.

Theorem 3.1. *Assume that $\boldsymbol{Y}_N(0) \xrightarrow{\text{W}} \boldsymbol{\omega}$. Then, as $N \to \infty$, the time-scaled ancestral process $(\boldsymbol{Y}_N([Nt]))_{t\geq 0}$ converges weakly in $D_E[0,\infty)$ to a structured coalescent $\boldsymbol{Y} = (\boldsymbol{Y}(t))_{t\geq 0}$ with initial distribution $\boldsymbol{\omega}$ and infinitesimal generator $\boldsymbol{Q} = (q_{\boldsymbol{r},\boldsymbol{r}'})_{\boldsymbol{r},\boldsymbol{r}'\in E}$ with entries (2.2), where now $\alpha_k := \sum_{i=1}^{s} \gamma_i/a_{ik}$ and $\beta_{k\ell} := 2\mu_{\ell k} \sum_{i=1}^{s} \gamma_i a_{i\ell}/a_{ik}$, $k, \ell \in \{1, \ldots, M\}, k \neq \ell$.*

Remark 3.1. Note that $\boldsymbol{Y}(\cdot)$ describes a structured coalescent for which α_k is the average pairwise coalescence rate in subpopulation k. The quantity $\beta_{k\ell}$ is the average rate at which a lineage in subpopulation k migrates (backward) to subpopulation ℓ. The averages mentioned above are with respect to the subpopulation sizes.

Remark 3.2. For $M = 1$ we have $-q_{r,r} = q_{r,r-1} = \alpha_1 r(r-1)/2$. The rescaled limiting process $(Y(t/\alpha_1))_{t\geq 0}$ coincides with the standard Kingman coalescent. Assume now that $M = 2$. If the total population size is constant, i.e. if there exists a positive integer ν such that $a_{11} + a_{12} = \nu = a_{21} + a_{22}$, then the two parameters β_{12} and β_{21} reduce to

$$\beta_{k\ell} = 2\mu_{\ell k} \sum_{i=1}^{s} \gamma_i \frac{\nu - a_{ik}}{a_{ik}} = 2\mu_{\ell k}(\nu\alpha_k - 1).$$

Define $\boldsymbol{X}_N(\tau) := (\boldsymbol{N}(\tau), \boldsymbol{Y}_N(\tau))$ and denote the transition matrix of the process $(\boldsymbol{X}_N(\tau))_{\tau\geq 0}$ by $\boldsymbol{\Pi}_N$, with entries

$$\pi_{(i,\boldsymbol{r}),(j,\boldsymbol{r}')} := \mathrm{P}(\boldsymbol{X}_N(\tau+1) = (\boldsymbol{a}_j N, \boldsymbol{r}') \mid \boldsymbol{X}_N(\tau) = (\boldsymbol{a}_i N, \boldsymbol{r})).$$

The process $(\boldsymbol{X}_N(\tau))_{\tau\in\mathbb{N}}$ has the state space $E_N := S_N \times E$ (see (3.1) and (2.1)). Order the elements of E by level, from level 1 to level n. Within a level the ordering is arbitrary. The level of an element $\boldsymbol{r} = (r_1, \ldots, r_M)$ of E is defined as $|\boldsymbol{r}| = r_1 + \cdots + r_M$, as is customary. We obtain the ordering in E_N by extending the ordering in E, by replacing each element $\boldsymbol{r}$ in E by $(\boldsymbol{a}_1 N, \boldsymbol{r}), \ldots, (\boldsymbol{a}_s N, \boldsymbol{r})$ in this order. We have the following lemma.

Lemma 3.1. *We obtain $\boldsymbol{\Pi}_N = \boldsymbol{A} + \boldsymbol{B}/N + \boldsymbol{o}(N^{-1})$, where the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ have entries*

$$a_{(i,\boldsymbol{r}),(j,\boldsymbol{r}')} = t_{ij}\delta_{\boldsymbol{r}\boldsymbol{r}'}, \quad \textit{where } \delta \textit{ is the Kronecker delta function},$$

$$b_{(i,\boldsymbol{r}),(j,\boldsymbol{r}')} = \begin{cases} -t_{ij} \displaystyle\sum_{k=1}^{M} \left(\binom{r_k}{2} \frac{1}{a_{jk}} + r_k \sum_{\substack{\ell=1 \\ \ell\neq k}}^{M} \mu_{\ell k} \frac{a_{j\ell}}{a_{jk}} \right) & \textit{if } \boldsymbol{r}' = \boldsymbol{r}, \\ \dfrac{t_{ij} r_k \mu_{\ell k} a_{j\ell}}{a_{jk}} & \textit{if } \boldsymbol{r}' = \boldsymbol{r} - \boldsymbol{e}_k + \boldsymbol{e}_\ell,\ k \neq \ell, \\ t_{ij} \dbinom{r_k}{2} \dfrac{1}{a_{jk}} & \textit{if } \boldsymbol{r}' = \boldsymbol{r} - \boldsymbol{e}_k, \\ 0 & \textit{otherwise}, \end{cases}$$

with $(i, \boldsymbol{r}), (j, \boldsymbol{r}') \in \{1, \ldots, s\} \times E$. By $\boldsymbol{o}(N^{-1})$ we denote the matrix of appropriate dimension whose entries are $o(N^{-1})$.

https://doi.org/10.1239/jap/1152413727 Published online by Cambridge University Press

Remark 3.3. The matrix $\boldsymbol{\Pi}_N$ satisfies the conditions of Möhle's lemma (Möhle (1998a)) with $c_N = 1/N$ and $\boldsymbol{B}_N = \boldsymbol{B} + \boldsymbol{o}(1)$, where $\boldsymbol{o}(1)$ is the matrix of appropriate dimension whose entries tend to zero as $N \to \infty$.

For the proof of Lemma 3.1, let $f_{k\ell\,|\,j}$ be the probability that a randomly chosen lineage currently in subpopulation k migrates to subpopulation ℓ in the previous generation, given that the population size in the previous generation is $\boldsymbol{a}_j N$. We have

$$f_{k\ell\,|\,j} = \frac{m_{\ell k} a_{j\ell} N}{\sum_{z=1}^{M} m_{zk} a_{jz} N} = \frac{a_{j\ell} m_{\ell k}}{\sum_{z=1}^{M} a_{jz} m_{zk}}.$$

For $k \neq \ell$, we have, using (3.2),

$$\begin{aligned} f_{k\ell\,|\,j} &= \frac{a_{j\ell}\mu_{\ell k} N^{-1} + o(N^{-1})}{\sum_{z\neq k} a_{jz}\mu_{zk} N^{-1} + a_{jk}(1 - \sum_{z\neq k} \mu_{kz} N^{-1}) + o(N^{-1})} \\ &= \mu_{\ell k}\left(\frac{a_{j\ell}}{a_{jk}}\right) N^{-1} + o(N^{-1}). \end{aligned}$$

So

$$f_{kk\,|\,j} = 1 - N^{-1} \sum_{\ell\neq k} \mu_{\ell k}\left(\frac{a_{j\ell}}{a_{jk}}\right) + o(N^{-1}).$$

Now we proceed to the proof of Lemma 3.1. We start with two cases: $\boldsymbol{r}' = \boldsymbol{r} - \boldsymbol{e}_k$ (binary coalescence) and $\boldsymbol{r}' = \boldsymbol{r} - \boldsymbol{e}_k + \boldsymbol{e}_\ell$, $k \neq \ell$ (one migration). The other cases will follow.

Proof of Lemma 3.1. We have

$$\begin{aligned} \pi_{(i,\boldsymbol{r}),(j,\boldsymbol{r}-\boldsymbol{e}_k)} &= t_{ij}\left(\prod_{\ell=1}^{M} f_{\ell\ell\,|\,j}^{r_\ell}\right)\binom{r_k}{2}\frac{1}{a_{jk}N}\left(\prod_{\ell=1}^{r_k-1}\left(1 - \frac{\ell-1}{a_{jk}N}\right)\right), \\ \prod_{\ell=1}^{M} f_{\ell\ell\,|\,j}^{r_\ell} &= \prod_{\ell=1}^{M}\left(1 - N^{-1}\sum_{z\neq\ell}\mu_{z\ell}\left(\frac{a_{jz}}{a_{j\ell}}\right) + o(N^{-1})\right)^{r_\ell} \\ &= 1 - N^{-1}\sum_{\ell=1}^{M}\sum_{z\neq\ell} r_\ell \mu_{z\ell}\left(\frac{a_{jz}}{a_{j\ell}}\right) + o(N^{-1}), \\ \prod_{\ell=1}^{r_k-1}\left(1 - \frac{\ell-1}{a_{jk}N}\right) &= 1 - \binom{r_k - 1}{2}\frac{1}{a_{jk}N} + o(N^{-1}). \end{aligned}$$

It follows that

$$\pi_{(i,\boldsymbol{r}),(j,\boldsymbol{r}-\boldsymbol{e}_k)} = t_{ij}\binom{r_k}{2}\frac{1}{a_{jk}N} + o(N^{-1}). \tag{3.3}$$

Next, for $k \neq \ell$, we have

$$\begin{aligned} \pi_{(i,\boldsymbol{r}),(j,\boldsymbol{r}-\boldsymbol{e}_k+\boldsymbol{e}_\ell)} &= t_{ij}\left(\prod_{z\neq k} f_{zz\,|\,j}^{r_z}\right) r_k f_{kk\,|\,j}^{r_k-1} f_{k\ell\,|\,j} \\ &= t_{ij} r_k \mu_{\ell k}\left(\frac{a_{j\ell}}{a_{jk}}\right) N^{-1} + o(N^{-1}). \end{aligned} \tag{3.4}$$

https://doi.org/10.1239/jap/1152413727 Published online by Cambridge University Press

Each coalescence event occurs with probability $O(N^{-1})$ and each migration event happens with probability $O(N^{-1})$ (see (3.3) and (3.4)). This implies that, for $|\boldsymbol{r}'| \leq |\boldsymbol{r}|$, $\boldsymbol{r}' \neq \boldsymbol{r}$, $\boldsymbol{r}' \neq \boldsymbol{r} - \boldsymbol{e}_k$, and $\boldsymbol{r}' \neq \boldsymbol{r} - \boldsymbol{e}_k + \boldsymbol{e}_\ell$, $k \neq \ell$,

$$\pi_{(i,\boldsymbol{r}),(j,\boldsymbol{r}')} = o(N^{-1}). \tag{3.5}$$

Obviously, for $|\boldsymbol{r}'| > |\boldsymbol{r}|$,

$$\pi_{(i,\boldsymbol{r}),(j,\boldsymbol{r}')} = 0. \tag{3.6}$$

Equations (3.3)–(3.6) together with $\sum_{\boldsymbol{r}' \in E} \pi_{(i,\boldsymbol{r}),(j,\boldsymbol{r}')} = t_{ij}$ imply that

$$\pi_{(i,\boldsymbol{r}),(j,\boldsymbol{r})} = t_{ij}\left[1 - N^{-1}\sum_{k=1}^{M}\left(\binom{r_k}{2}\frac{1}{a_{jk}} + r_k \sum_{\ell \neq k} \frac{\mu_{\ell k} a_{j\ell}}{a_{jk}}\right)\right] + o(N^{-1}).$$

This concludes the proof of Lemma 3.1.

4. Proof of Theorem 3.1

First, we compute $\lim_{N\to\infty} \boldsymbol{\Pi}_N^{[Nt]}$, using Möhle's lemma (Möhle (1998a)). Then we proceed to the proof of Theorem 3.1 to show the weak convergence of the process $(\boldsymbol{Y}_N(t))_{t\geq 0}$ to the limiting process $(\boldsymbol{Y}(t))_{t\geq 0}$.

We have, by Möhle's lemma (Möhle (1998a)),

$$\lim_{N\to\infty} \boldsymbol{\Pi}_N^{[Nt]} = \boldsymbol{P} - \boldsymbol{I} + \mathrm{e}^{t\boldsymbol{G}},$$

where $\boldsymbol{P} = \lim_{m\to\infty} \boldsymbol{A}^m$ and $\boldsymbol{G} = \boldsymbol{P}\boldsymbol{B}\boldsymbol{P}$, with $\boldsymbol{A}$ and $\boldsymbol{B}$ as in Lemma 3.1. From

$$a_{(i,\boldsymbol{r}),(j,\boldsymbol{r}')} = t_{ij}\delta_{\boldsymbol{r}\boldsymbol{r}'},$$

it follows that $p_{(i,\boldsymbol{r}),(j,\boldsymbol{r}')} = \gamma_j \delta_{\boldsymbol{r}\boldsymbol{r}'}$. The entries of $\boldsymbol{G}$ are easily computed to be

$$g_{(i,\boldsymbol{r}),(j,\boldsymbol{r}')} = \sum_{i',j'=1}^{s} p_{(i,\boldsymbol{r}),(i',\boldsymbol{r})} b_{(i',\boldsymbol{r}),(j',\boldsymbol{r}')} p_{(j',\boldsymbol{r}'),(j,\boldsymbol{r}')} = \gamma_j \sum_{i'=1}^{s} \gamma_{i'} \sum_{j'=1}^{s} b_{(i',\boldsymbol{r}),(j',\boldsymbol{r}')}.$$

Now, substituting the entries $b_{(i',\boldsymbol{r}),(j',\boldsymbol{r}')}$ of $\boldsymbol{B}$ and using $\boldsymbol{\gamma}\boldsymbol{T} = \boldsymbol{\gamma}$ yields

$$g_{(i,\boldsymbol{r}),(j,\boldsymbol{r}')} = \gamma_j q_{\boldsymbol{r},\boldsymbol{r}'}.$$

We use Ethier and Kurtz's theorem and some of their notation (see Ethier and Kurtz (1986)) to prove Theorem 3.1. Define $\eta_N \colon E_N \to E$ by $\eta_N(\boldsymbol{a}_i N, \boldsymbol{r}) := \boldsymbol{r}$. Then η_N is measurable. For $f \in B(E_N)$ and $\boldsymbol{x} \in E_N$, define

$$\mathcal{T}_N f(\boldsymbol{x}) := \sum_{\boldsymbol{y}\in E_N} f(\boldsymbol{y})\pi_{\boldsymbol{x},\boldsymbol{y}}.$$

Here, $B(E_N)$ is the Banach space of bounded functions on E_N with norm

$$\|f\| := \sup_{\boldsymbol{x}\in E_N} |f(\boldsymbol{x})|;$$

https://doi.org/10.1239/jap/1152413727 Published online by Cambridge University Press

$B(E)$ is defined similarly with E_N replaced with E. Thus, for each $f \in B(E)$ and each $(\boldsymbol{a}_i N, \boldsymbol{r}) \in E_N$, we obtain

$$\begin{aligned}\mathcal{T}_N(f \circ \eta_N)(\boldsymbol{a}_i N, \boldsymbol{r}) &= \sum_{(\boldsymbol{a}_j N, \boldsymbol{r}') \in E_N} f \circ \eta_N(\boldsymbol{a}_j N, \boldsymbol{r}')(\boldsymbol{\Pi}^{[Nt]})_{(i,\boldsymbol{r}),(j,\boldsymbol{r}')} \\ &= \sum_{\boldsymbol{r}' \in E} \sum_{j=1}^{s} f(\boldsymbol{r}')(\boldsymbol{\Pi}^{[Nt]})_{(i,\boldsymbol{r}),(j,\boldsymbol{r}')}\end{aligned}$$

and, for each $\boldsymbol{r} \in E$ and $f \in B(E)$, define

$$\mathcal{T}(t) f(\boldsymbol{r}) := \sum_{\boldsymbol{r}' \in E} f(\boldsymbol{r}')(\mathrm{e}^{t\boldsymbol{Q}})_{\boldsymbol{r},\boldsymbol{r}'}.$$

The quantity $(\boldsymbol{\Pi}^{[Nt]})_{(i,\boldsymbol{r}),(j,\boldsymbol{r}')}$ is the probability that the process $(\boldsymbol{X}_N(\tau))_{\tau \in \mathbb{N}}$ jumps from state $(\boldsymbol{a}_i N, \boldsymbol{r})$ to state $(\boldsymbol{a}_j N, \boldsymbol{r}')$ in $[N\tau]$ generations, and $(\mathrm{e}^{t\boldsymbol{Q}})_{t \geq 0}$ is the Feller semigroup for the limiting process $(\boldsymbol{Y}(t))_{t \geq 0}$. Now we proceed to the proof of Theorem 3.1.

Proof of Theorem 3.1. We use a technique borrowed from Kaj *et al.* (2001). To show that $\boldsymbol{Y}_N(\cdot)$ converges weakly to $\boldsymbol{Y}(\cdot)$, it is sufficient to show that, for each function $f\colon E \to \mathbb{R}$ and each state $(\boldsymbol{a}_i N, \boldsymbol{r}) \in E_N$,

$$|\mathcal{T}_N^{[Nt]}(f \circ \eta_N)(\boldsymbol{a}_i N, \boldsymbol{r}) - \mathcal{T}(t) f(\boldsymbol{r})| \to 0 \quad \text{as } N \to \infty.$$

We have

$$|\mathcal{T}_N^{[Nt]}(f \circ \eta_N)(\boldsymbol{a}_i N, \boldsymbol{r}) - \mathcal{T}(t) f(\boldsymbol{r})| = \left| \sum_{\boldsymbol{r}' \in E} f(\boldsymbol{r}') \left\{ \sum_{j=1}^{s} (\boldsymbol{\Pi}_N^{[Nt]})_{(i,\boldsymbol{r}),(j,\boldsymbol{r}')} - (\mathrm{e}^{t\boldsymbol{Q}})_{\boldsymbol{r},\boldsymbol{r}'} \right\} \right|.$$

Since E is finite and $f\colon E \to \mathbb{R}$ is a function, it is enough to show that, for each $\boldsymbol{r}$ and $\boldsymbol{r}'$ in E and $i \in \{1, \ldots, s\}$,

$$\left| \sum_{j=1}^{s} (\boldsymbol{\Pi}_N^{[Nt]})_{(i,\boldsymbol{r}),(j,\boldsymbol{r}')} - (\mathrm{e}^{t\boldsymbol{Q}})_{\boldsymbol{r},\boldsymbol{r}'} \right| \to 0 \quad \text{as } N \to \infty.$$

Recall that $\lim_{N\to\infty} \boldsymbol{\Pi}_N^{[Nt]} = \boldsymbol{P} - \boldsymbol{I} + \mathrm{e}^{t\boldsymbol{G}}$. This implies that

$$\lim_{N\to\infty} \sum_{j=1}^{s} (\boldsymbol{\Pi}_N^{[Nt]})_{(i,\boldsymbol{r}),(j,\boldsymbol{r}')} = \sum_{j=1}^{s} (\mathrm{e}^{t\boldsymbol{G}})_{(i,\boldsymbol{r}),(j,\boldsymbol{r}')}.$$

So it is enough to show that, for each $i \in \{1, \ldots, s\}$ and $\boldsymbol{r}$ and $\boldsymbol{r}'$ in E,

$$\sum_{j=1}^{s} (\mathrm{e}^{t\boldsymbol{G}})_{(i,\boldsymbol{r}),(j,\boldsymbol{r}')} = (\mathrm{e}^{t\boldsymbol{Q}})_{\boldsymbol{r},\boldsymbol{r}'}.$$

We only need to verify by induction that

$$\sum_{j=1}^{s} (\boldsymbol{G}^{\tau})_{(i,\boldsymbol{r}),(j,\boldsymbol{r}')} = (\boldsymbol{Q}^{\tau})_{\boldsymbol{r},\boldsymbol{r}'}. \tag{4.1}$$

https://doi.org/10.1239/jap/1152413727 Published online by Cambridge University Press

Equality (4.1) obviously holds for $\tau \in \{0, 1\}$. For $\tau = 0$, both sides are equal to $\delta_{\boldsymbol{r},\boldsymbol{r}'}$. For $\tau = 1$, (4.1) is equivalent to $\sum_j g_{(i,\boldsymbol{r}),(j,\boldsymbol{r}')} = q_{\boldsymbol{r},\boldsymbol{r}'}$, which is obviously satisfied. Suppose, as the induction hypothesis, that

$$\sum_{j=1}^{s} (\boldsymbol{G}^{\tau-1})_{(i,\boldsymbol{r}),(j,\boldsymbol{r}')} = (\boldsymbol{Q}^{\tau-1})_{\boldsymbol{r},\boldsymbol{r}'}.$$

Then we obtain

$$\begin{aligned}
\sum_{j=1}^{s} (\boldsymbol{G}^{\tau})_{(i,\boldsymbol{r}),(j,\boldsymbol{r}')} &= \sum_{j=1}^{s} \sum_{(i',\boldsymbol{r}'')\in E_N} g_{(i,\boldsymbol{r}),(i',\boldsymbol{r}'')} (\boldsymbol{G}^{\tau-1})_{(i',\boldsymbol{r}''),(j,\boldsymbol{r}')} \\
&= \sum_{(i',\boldsymbol{r}'')\in E_N} g_{(i,\boldsymbol{r}),(i',\boldsymbol{r}'')} \sum_{j=1}^{s} (\boldsymbol{G}^{\tau-1})_{(i',\boldsymbol{r}''),(j,\boldsymbol{r}')} \\
&= \sum_{(i',\boldsymbol{r}'')\in E_N} \gamma_{i'} q_{\boldsymbol{r},\boldsymbol{r}''} (\boldsymbol{Q}^{\tau-1})_{\boldsymbol{r}'',\boldsymbol{r}'} \\
&= \sum_{i'=1}^{s} \gamma_{i'} \sum_{\boldsymbol{r}''\in E} q_{\boldsymbol{r},\boldsymbol{r}''} (\boldsymbol{Q}^{\tau-1})_{\boldsymbol{r}'',\boldsymbol{r}'} \\
&= \sum_{i'=1}^{s} \gamma_{i'} (\boldsymbol{Q}^{\tau})_{\boldsymbol{r},\boldsymbol{r}'} \\
&= (\boldsymbol{Q}^{\tau})_{\boldsymbol{r},\boldsymbol{r}'}.
\end{aligned}$$

Thus, for $\tau = 0, 1, 2, \ldots$, (4.1) holds. We conclude that $\boldsymbol{Y}_N(\cdot)$ converges weakly to $\boldsymbol{Y}(\cdot)$. The proof is complete.

5. Summary

The coalescent is an important tool for modeling the ancestral dynamics of many biological populations. It allows us to trace back in time the ancestry of a sample of genes or individuals chosen from a large population (that may have a complex structure) from the present time until the sample reaches its most recent common ancestor.

For many population models the suitably scaled ancestral process converges to the standard or the structured coalescent, or their time-changed versions. This is referred to as the robustness of the coalescent.

In our model, we consider a subdivided population with stochastically varying subpopulation sizes and slow migration between subpopulations. We point out that our model, although similar to some of the existing models, is different from them, and that, on the coalescent time scale, it converges to a time-changed version of the structured coalescent, thus confirming the robustness of the coalescent.

Acknowledgements

The content of this paper comes mainly from the first part (which has been completely reorganized) of my doctoral thesis (Sampson (2004)). I want to thank Stephen M. Krone for accepting to be my major professor, for his guidance and supervision of my doctoral thesis. My thanks go also to my current postdoctoral advisor, Peter Beerli, for his comments and

suggestions. Finally, I would like to thank two anonymous reviewers for their thoughtful comments and suggestions on the manuscript.

References

Bahlo, M. and Griffiths, R. C. (2000). Inference from gene trees in a subdivided population. *Theoret. Pop. Biol.* **57,** 79–95.

Beerli, P. and Felsenstein, J. (2001). Maximum likelihood estimation of a migration matrix and effective population sizes in *n* subpopulations by using a coalescent approach. *Proc. Nat. Acad. Sci. USA* **98,** 4563–4568.

Donnelly, P. (1986). A genealogical approach to variable population size models in population genetics. *J. Appl. Prob.* **23,** 283–296.

Donnelly, P. and Tavaré, S. (1995). Coalescents and genealogical structure under neutrality. *Ann. Rev. Genet.* **29,** 401–421.

Ethier, S. N. and Kurtz, T. G. (1986). *Markov Processes: Characterization and Convergence.* John Wiley, New York.

Fu, Y. X. (1997). Coalescent theory for a partially selfing population. *Genetics* **146,** 1489–1499.

Griffiths, R. C. and Marjoram, P. (1996). Ancestral inference from samples of DNA sequences with recombination. *J. Comput. Biol.* **3,** 479–502.

Griffiths, R. C. and Tavaré, S. (1994). Sampling theory for neutral alleles in a varying environment. *Phil. Trans. R. Soc. London B* **344,** 403–410.

Herbots, H. M. (1994). Stochastic models in population genetics: genealogy and genetic differentiation in structured populations. Doctoral Thesis, University of London.

Herbots, H. M. (1997). The structured coalescent. In *Progress in Population Genetics and Human Evolution*, eds P. Donnelly and S. Tavaré, Springer, New York, pp. 231–255.

Hey, J. and Wakeley, J. (1997). A coalescent estimator of the population recombination rate. *Genetics* **145,** 833–846.

Hudson, R. R. (1990). Gene genealogies and the coalescent process. In *Oxford Surveys in Evolutionary Biology*, Vol. 7, eds D. Futuyma and J. Antonovics, Oxford University Press, pp. 1–43.

Hudson, R. R. (1998). Island models and the coalescent process. *Mol. Ecol.* **7,** 413–418.

Hudson, R. R. and Kaplan, N. L. (1988). The coalescent process in models with selection and recombination. *Genetics* **120,** 831–840.

Jagers, P. and Sagitov, S. (2004). Convergence to the coalescent in populations of substantially varying size. *J. Appl. Prob.* **41,** 33–48.

Kaj, I. and Krone, S. M. (2003). The coalescent process in a population with stochastically varying size. *J. Appl. Prob.* **40,** 368–378.

Kaj, I., Krone, S. M. and Lascoux, M. (2001). Coalescent theory for seed bank models. *J. Appl. Prob.* **38,** 285–301.

Kaplan, N. L., Darden, T. and Hudson, R. R. (1988). The coalescent process in models with selection. *Genetics* **120,** 819–829.

Kingman, J. F. C. (1982a). Exchangeability and the evolution of large populations. In *Exchangeability in Probability and Statistics*, eds G. Koch and F. Spizzichino, North-Holland, Amsterdam, pp. 97–112.

Kingman, J. F. C. (1982b). On the genealogy of large populations. In *Essays in Statistical Science* (J. Appl. Prob. Spec. Vol. **19A**), eds J. Gani and E. J. Hannan, Applied Probability Trust, Sheffield, pp. 27–43.

Kingman, J. F. C. (1982c). The coalescent. *Stoch. Process. Appl.* **13,** 235–248.

Möhle, M. (1998a). A convergence theorem for Markov chains arising in population genetics and the coalescent with selfing. *Adv. Appl. Prob.* **30,** 493–512.

Möhle, M. (1998b). Coalescent results for two-sex population models. *Adv. Appl. Prob.* **30,** 513–520.

Möhle, M. (2002). The coalescent in population models with time-inhomogeneous environment. *Stoch. Process. Appl.* **97,** 199–227.

Möhle, M. and Sagitov, S. (2003). Coalescent patterns in diploid exchangeable population models. *J. Math. Biol.* **47,** 337–352.

Neuhauser, C. and Krone, S. M. (1997). The genealogy of samples in models with selection. *Genetics* **145,** 519–534.

Nordborg, M. (1997). Structured coalescent processes on different time scales. *Genetics* **146,** 1501–1514.

Nordborg, M. (1999). The coalescent with partial selfing and balancing selection: an application of structured coalescent processes. In *Statistics in Molecular Biology and Genetics* (IMS Lecture Notes Monogr. Ser. **33**), Institute of Mathematical Statistics, Hayward, CA, pp. 56–76.

Nordborg, M. (2001). Coalescent theory. In *Handbook of Statistical Genetics*, eds D. J. Balding, M. J. Bishop and C. Cannings, John Wiley, Chichester, pp. 179–212.

Nordborg, M. and Donnelly, P. (1997). The coalescent process with selfing. *Genetics* **146,** 1185–1195.

Nordborg, M. and Krone, S. M. (2002). Separation of time scales and convergence to the coalescent in structured populations. In *Modern Developments in Theoretical Population Genetics*, eds M. Slatkin and M. Veuille, Oxford University Press, pp. 194–232.

NOTOHARA, M. (1990). The coalescent and the genealogical process in geographically structured populations. *J. Math. Biol.* **29,** 59–75.
SAMPSON, K. Y. (2004). Structured coalescent with nonconservative migration. Doctoral Thesis, Department of Mathematics, University of Idaho.
SANO, A., SHIMIZU, A. AND IIZUKA, M. (2004). Coalescent process with fluctuating population size and its effective size. *Theoret. Pop. Biol.* **65,** 39–48.
TAJIMA, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105,** 437–460.
TAJIMA, F. (1989). The effect of change in population size on DNA polymorphism. *Genetics* **123,** 597–601.
TAKAHATA, N. (1991). Genealogy of neutral genes and spreading of selected mutations in a geographically structured population. *Genetics* **129,** 585–595.
TAVARÉ, S. (1984). Line-of-descent and genealogical processes, and their applications in population genetics models. *Theoret. Pop. Biol.* **26,** 119–164.
WAKELEY, J. (2000). The effects of subdivision on the genetic divergence of populations and species. *Evolution* **54,** 1092–1101.
WILKINSON-HERBOTS, H. M. (1998). Genealogy and subpopulation differentiation under various models of population structure. *J. Math. Biol.* **37,** 535–585.