

Corridors for migration between large subdivided populations, and the structured coalescent

John Wakeley^{a,*}, Sabin Lessard^b

^a*Department of Organismic and Evolutionary Biology, Harvard University, 2102 Biological Laboratories, 16 Divinity Avenue, Cambridge, MA 02138, USA*

^b*Département de Mathématiques et de Statistique, Université de Montréal, CP 6128 succursale Centre-Ville, Montréal QC, Canada H3C 3J7*

Received 19 October 2005

Available online 12 June 2006

Abstract

We study the ancestral genetic process for samples from two large, subdivided populations that are connected by migration to, from, and within a small set of subpopulations, or demes. We consider convergence to an ancestral limit process as the numbers of demes in the two large, subdivided populations tend to infinity. We show that the ancestral limit process for a sample includes a recent instantaneous adjustment to the sample size and structure followed by a more ancient process that is identical to the usual structured coalescent, but with different scaled parameters. This justifies the application of a modified structured coalescent to some hierarchically structured populations.

© 2006 Elsevier Inc. All rights reserved.

Keywords: Population structure; Migration; Coalescence

1. Introduction

Natural populations are typically distributed over the landscape such that it is unusual for an individual organism to traverse the entire species range within its lifetime. For this reason it is common to find some correspondence between genetic diversities and geographic locations (Slatkin, 1987). Observations of this sort are not predicted by the standard models, which assume a well-mixed, or panmictic, population and which form the basis of the ancestral limit process called the *coalescent* (Kingman, 1982; Hudson, 1983; Tajima, 1983). The standard coalescent describes the ancestry of a sample of genetic data at a locus without selection or recombination from a population of constant size over time. In order to develop an understanding of structured populations and to provide tools for making inferences about migration rates from genetic data, a number of alternative models which include population structure have been proposed.

The most commonly used model of a subdivided population is the model behind the *structured coalescent* (Notohara, 1990; Herbots, 1994, 1997; Wilkinson-Herbots, 1998). This model accounts for geographic–genetic structure by supposing that migration rates between populations are of the same order of magnitude as the inverse of the local population size, and that the local population size is very large. Here we study a model of hierarchical population structure in which geographic–genetic structure develops without small migration rates, as a result of a constriction in the species range. We show that the structured coalescent may be applied in this situation if it is modified slightly to account for the hierarchical structure of the population.

The structured coalescent underlies most software packages for inferring migration rates from genetic data, including MIGRATE (Beerli and Felsenstein, 2001), GENE-TREE (Bahlo and Griffiths, 2000), MDIV (Nielsen and Wakeley, 2001) and IM (Hey and Nielsen, 2004). The structured coalescent considers the ancestral process for a sample of genetic data from a population made up of a number of local populations that exchange migrants. We follow the usual population genetic terminology and use

*Corresponding author. Fax: +1 617 496 5854.

E-mail address: wakeley@fas.harvard.edu (J. Wakeley).

deme to mean local population. The number of demes is D , the deme sizes are N_i , and m_{ij} is the fraction of deme i that is replaced by migrants from deme j each generation. Reproduction is assumed to occur according to the Wright–Fisher model (Fisher, 1930; Wright, 1931), although other models give essentially the same result. The structured coalescent holds for a fixed number of demes in the limit as the deme sizes tend to infinity with the products $N_i m_{ij}$ assumed to be finite. Thus, this model is intended to approximate the dynamics of a population in which N_i is large for every deme i , and m_{ij} is small for every pair of demes i and j . The description above assumes the organisms are haploid, but the structured coalescent may be applied to diploid organisms, with some restrictions (Nagylaki, 1998), if N_i is replaced with $2N_i$.

While the structured coalescent can be expected to apply over a broad range of values of N_i and m_{ij} , in situations where the products $N_i m_{ij}$ are either very small or very large other limit models will provide a better approximation to the ancestral process. One such model is the strong-migration limit, which Nagylaki (1980) proved for the forward-time diffusion of allele frequencies and Notohara (1993) put into a gene genealogical context. In this limit, the products $N_i m_{ij}$ tend to infinity as the deme sizes N_i tend to infinity, and the dynamics of the population become identical to those of a single well-mixed population with an effective size that depends on the parameters N_i and m_{ij} .

Nordborg and Krone (2002) introduced a more general model that combines aspects of both the structured coalescent and the strong-migration limit. Specifically, they considered a subdivided population in which $N_i m_{ij}$ remains finite for some pairs of demes but diverges for other pairs of demes. Nordborg and Krone (2002) used a result due to Möhle (1998) to prove convergence of the ancestral process to a structured coalescent. The difference between this model and the usual structured coalescent is that structure collapses within sets of demes that are connected by strong migration, so that migration is restricted only *between* sets of demes for which all the pairwise $N_i m_{ij}$ (where i is in one set and j is in another set) remain finite. This model is intended as an approximation to the coalescent process in a subdivided population containing a finite number of large demes in which some rates of migration are high and others are low.

Slatkin and Voelm (1991) studied inbreeding coefficients in a model of hierarchical population structure (Nei, 1973; Carmelli and Cavalli-Sforza, 1976; Sawyer and Felsenstein, 1983), but made different assumptions than Nordborg and Krone (2002) about the relative sizes of migration fractions within and between sets of demes. Slatkin and Voelm (1991) assumed that demes of a single size N were organized into “neighborhoods” and that migration rates between demes within neighborhoods, m_w , could be different than migration rates between demes in different neighborhoods, m_b . Their analysis of expected pairwise coalescence times showed that population structure can be appreciable both between demes in the same neighborhood

and between demes in different neighborhoods if the products Nm_w and NDm_b are not too large. Note that m_w and m_b are called m_1 and m_2 in Slatkin and Voelm (1991), whereas we reserve the latter for use below.

The ancestral process for a sample from a subdivided population of this sort was considered in Wakeley (2000), where it was shown that a result analogous to that of Nordborg and Krone (2002) is obtained in the limit as the number of demes tends to infinity. Two subdivided populations, each with many demes, were assumed to be connected by migration. Migration among demes within each population was assumed to occur according to the island model (Wright, 1931), with migration fraction m for each deme. Another migration fraction m_{12} accounted for movement between demes in different populations. A heuristic analysis showed that if Nm and NDm_{12} are finite as both D and N tend to infinity, then a modified version of the structured coalescent described the ancestral limit process. The modifications were to recognize that the time scale of coalescence within populations depends on Nm in addition to the total population size ND and to include a “scattering phase” for cases when multiple samples are taken from single demes (Wakeley, 1999).

All of the above results, as well as some more recent findings (Wilkinson-Herbots and Ettridge, 2004), reflect the fact that in order for there to be any appreciable effect of subdivision between large populations, the migration fractions between populations or between demes in different populations must be of the same order of magnitude as the inverse of the total population size. This is true whether one considers very large demes or very large numbers of demes per population or both. When the rates of movement between populations are small in this sense, i.e. scaled inversely with the population size, then the structured coalescent is an appropriate limit model to use to approximate the ancestral process and to make inferences from genetic data. However, this can be achieved without assuming that the migration fractions themselves are small. In this article, we consider a different kind of hierarchical subdivision, in which this scaling of migration rates results from a constriction in the habitat, and is obtained for arbitrary migration fractions. We imagine that migration is hierarchically structured by the relatively low abundance of gateway demes which give individuals access to a “migration corridor” between populations.

The model we introduce in the next section is suggested by recent empirical work on a variety of organisms. For example, Vollmer and Palumbi (2002) obtained sequence data at two nuclear loci and one mitochondrial locus in three sympatric species of Caribbean corals in the genus *Acropora*. All the individuals of the relatively rare species *Acropora prolifera* were shown to be F_1 hybrids between two relatively more common species *Acropora cervicornis* and *Acropora palmata*. The genetic data provided evidence for non-zero, but low levels of gene flow between the two “parent” species. The work of Krings et al. (1999) provides another example, in which the authors used sequences of

hypervariable region 1 of human mitochondrial DNA to investigate long-term patterns of movement between Africa and Eurasia along the Nile river valley. The data showed evidence for gene flow in both directions, as well as a pattern of isolation by distance (Wright, 1943) in which the extent of northern versus southern affiliation changed as one moved along the valley.

2. A model of two overlapping structured populations

We begin with the idealized model depicted in Fig. 1. Two populations are each subdivided into a large number of demes. Between these populations, 1 and 2, sits a third population that is subdivided into a relatively small number of demes. We assume that there is no direct migration between population 1 and population 2, but that individuals can move from one to the other by passing through population 3. Let D_1 and D_2 be the numbers of demes in populations 1 and 2, and let d be the number of demes in population 3. The total number of demes is thus $D_1 + D_2 + d$. We consider the ancestral genetic process for this model in the limit as D_1 and D_2 tend to infinity for a fixed, or constant, value of d .

There are other parameters in the model. We assume that every deme is of the same size N . The backward migration probability for a deme is the fraction of its membership that is replaced by migrants each generation. We assume that generations are non-overlapping: all adults die and are replaced by offspring each generations. The order of events in the model forward in time is Wright–Fisher sampling, or reproduction, followed by migration, but we do not deal explicitly with the forward-time process. Instead, we assume that lineages migrate independently of one another backward in time. We let demes in different populations have different backward migration fractions, which we denote m_1 , m_2 , and m_3 . Using three migration rates allows us to illustrate the way in which the scaled migration parameters in the limit process depend on movement to and from population 3. Thus, migration is conservative (Nagylaki, 1980; Strobeck, 1987; Herbots, 1997) only when $m_1 = m_2 = m_3$.

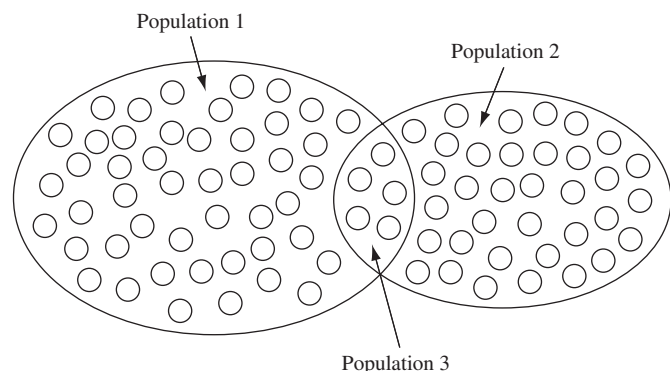


Fig. 1. An illustration of the model.

Migration occurs according to the island model (Wright, 1931), but structured in the following way. A migrant in population 1 is equally likely to have come from any deme in population 1 or population 3 (including the deme it is in currently) and a migrant in population 2 is equally likely to have come from any deme in population 2 or population 3. For the demes in population 3 we assume that migrants are equally likely to have come from any deme in the total population ($1 + 2 + 3$). Finally, we assume that the deme size and the three migration probabilities are greater than zero—the migration probabilities also cannot be larger than one—and that all four of these parameters are constant in the limit as D_1 and D_2 tend to infinity.

The ancestry of a sample of size one reveals some important aspects of the ancestral process. The single ancestral lineage is either in population 1, in population 2, or in population 3. Call these state 1, state 2, and state 3, respectively. Movement of the lineage among these states occurs by migration, but some events are much more likely than others. We can separate the probable from the improbable events by writing the single-generation transition matrix as the sum of two matrices,

$$\mathbf{M} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \frac{m_3 D_1}{D_1 + D_2 + d} & \frac{m_3 D_2}{D_1 + D_2 + d} & 1 - m_3 \end{pmatrix} + \begin{pmatrix} -\frac{m_1 d}{D_1 + d} & 0 & \frac{m_1 d}{D_1 + d} \\ 0 & -\frac{m_2 d}{D_2 + d} & \frac{m_2 d}{D_2 + d} \\ 0 & 0 & \frac{m_3 d}{D_1 + D_2 + d} \end{pmatrix}, \quad (1)$$

which differ greatly in magnitude when D_1 and D_2 are both large. To explain Eq. (1), the entry $(\mathbf{M})_{ij}$, in row i and column j , is the probability of moving from state i to state j in a single generation back in time.

We let $D = D_1 + D_2 + d$ and assume that the ratios D_1/D and D_2/D are constant in the limit $D \rightarrow \infty$. The non-zero entries in the first matrix on the right-hand side of Eq. (1) are all of order 1 and are constant in this limit, while those in the second matrix are all of order $1/D$ and therefore decrease in proportion to $1/D$. For the lineage to move from population 1 to population 2 it must go through population 3. This requires a migration event of order $1/D$ that takes the lineage from population 1 to population 3 followed by a migration event of order 1 that takes the lineage from population 3 to population 2. A mathematical formalism is available (Möhle, 1998) for letting $D \rightarrow \infty$ to obtain an ancestral limit process. We will use this formalism to obtain the ancestral limit process for samples larger than one, under the additional assumption

that the total sample size is much less than the total number of demes D .

Notice that Eq. (1) uses a collapsed state space: due to the island-type structure of the model, within each population every deme is equivalent to every other deme, so it is only necessary to know which population the lineage is in, not which deme. A full accounting of states for a larger number of lineages would label all the demes and record the numbers of lineages in each, but this is unnecessary for two reasons. The first is, again, the assumed island-type structure of the populations. The second reason is the vast difference in probabilities of certain events in the limit we consider. In particular, migration events that bring lineages into demes already occupied by another lineage (or lineages) and migration events to population 3 occur with probabilities of order $1/D$. On the other hand, migration events that move a lineage from a multiply occupied deme to an unoccupied deme, coalescent events within multiply occupied demes, and migration events out of population 3 occur with probabilities of order 1 when they are possible. As the number of demes increases to infinity, the lineages will almost surely be distributed such that some number sit alone in demes in population 1 and the rest sit alone in demes in population 2.

Therefore, we can classify the possible distributions of lineages among demes into five disjoint sets of states, and use a reduced notation for states within each set. S_1 includes all distributions where every lineage is in a separate deme and none are in population 3. We use $(n_1, n_2, 0)$ to denote a state in S_1 . S_2 includes all distributions where every lineage is in a separate deme and one lineage is in population 3; i.e. $(n_1, n_2, 1)$. The set S_2 is needed because migration in the limit process occurs by the movement of a single lineage into population 3. S_3 includes all distributions where a pair of lineages is in the same deme in population 1, and every other lineage is in a separate deme and none are in population 3. We use $(n_1+, n_2, 0)$ to denote a state in S_3 . S_4 includes all distributions where a pair of lineages is in the same deme in population 2, and every other lineage is in a separate deme and none are in population 3; i.e. $(n_1, n_2+, 0)$. The sets S_3 and S_4 are needed because coalescent events in the limit process depend on migration events to demes that already contain a lineage. S_5 includes all other states, and this set is needed to account for the original sample, whose distribution is determined by the experimenter.

3. The ancestral limit process

Here we state our main result, which is proved in the next section. We use the results of Möhle (1998) to prove that the ancestral limit process for a sample in which each lineage is in a separate deme and none are in population 3 has a structure identical to the usual structured coalescent reviewed above. Time in this process is rescaled by the total population size, so that it is measured in units of ND

generations. Transitions in the ancestral process are given by

$$(n_1, n_2, 0) \rightarrow \begin{cases} (n_1 - 1, n_2, 0) & \text{with rate} \\ & \binom{n_1}{2} (1 - F_1)/\alpha_1, \\ (n_1, n_2 - 1, 0) & \text{with rate} \\ & \binom{n_2}{2} (1 - F_2)/\alpha_2, \\ (n_1 - 1, n_2 + 1, 0) & \text{with rate } n_1 M_{12}^{(3)}, \\ (n_1 + 1, n_2 - 1, 0) & \text{with rate } n_2 M_{21}^{(3)}, \end{cases} \quad (2)$$

in which $\alpha_i = D_i/D$ is the fraction of individuals that live in population i ,

$$F_i = \frac{(1 - m_i)^2}{Nm_i(2 - m_i) + (1 - m_i)^2} \quad (3)$$

for $i = 1, 2$, is the usual inbreeding coefficient and

$$M_{ij}^{(3)} = Ndm_i \frac{\alpha_j}{\alpha_i} \quad (4)$$

is the scaled, backward rate of migration from population i to population j , for $(i, j) = (1, 2), (2, 1)$. The superscript in Eq. (4) is in recognition of the fact that migration between populations 1 and 2 has to occur via population 3.

The dynamics of the ancestral process are understood as follows. In the limit as the number of demes D tends to infinity and with time rescaled by the total population size ND , the ancestral processes within population 1 and within population 2 become identical to the many-demes coalescent described previously; e.g. see Wakeley (1998) and, more recently, Lessard and Wakeley (2004). The rate of coalescence in the many-demes limit depends on the size of the population and its inbreeding coefficient, often called F . Here, because we measure time in units of the total population size, the factors $\alpha_1/(1 - F_1)$ and $\alpha_2/(1 - F_2)$ appear in Eq. (2) as the relative sizes of the two populations. Migration between population 1 and population 2 proceeds by the unlikely event that a lineage migrates to one of the finite number of demes in intersection population 3 (see Fig. 1) then migrates from there to the other population. The rescaled migration rate in Eq. (4) is finite because it records events of order $1/D$ on a time scale proportional to D .

As mentioned above, this limit result does not require any of the migration fractions to be small. It might seem curious that m_3 does not appear in Eqs. (2)–(4). However, this is a straightforward consequence of the assumption that m_3 is constant in the limit $D \rightarrow \infty$. When a lineage is in population 3, it will spend an average of $1/m_3$ generations there, and then it will move either to population 1 or to population 2. The total number of generations that lineages spend in population 3 up to the most recent common ancestor of the sample becomes negligible in the limit process with $D \rightarrow \infty$. However, in

thinking about applying our result to populations in nature, the smaller m_3 is, the larger D would have to be for Eq. (2) to be a good approximation to the ancestral process.

4. Convergence of the ancestral process

Here we establish the ($D \rightarrow \infty$) continuous-time ancestral limit process for a sample. We assume that $D_1/D \rightarrow \alpha_1$ and $D_2/D \rightarrow \alpha_2 = 1 - \alpha_1$ with $0 < \alpha_1, \alpha_2 < 1$ as $D = D_1 + D_2 + d$ tends to infinity, and that all other quantities (d , and N, m_i , and the sample size) are constant. Note that the number of lineages ancestral to the sample is always less than or equal to the sample size. We express the single-generation transition probability matrix Π_D for backward migration and coalescence among the lineages as the sum

$$\Pi_D = \mathbf{A} + \mathbf{B}/D + \mathbf{C}_D, \tag{5}$$

in which the matrix \mathbf{A} is defined by $\mathbf{A} = \lim_{D \rightarrow \infty} \Pi_D$, the matrix \mathbf{B} is defined by $\mathbf{B} = \lim_{D \rightarrow \infty} (\Pi_D - \mathbf{A})D$, and the matrix \mathbf{C}_D is defined by Eq. (5). Under the assumptions of our model, all the non-zero entries of \mathbf{A} are less than or equal to one, all the non-zero entries of \mathbf{B} are finite, and all the entries of \mathbf{C}_D are of order $1/D^2$ or smaller.

The matrix \mathbf{A} captures the fast events in the ancestral process: coalescent events when at least one deme contains more than one lineage, and migration events to demes in populations 1 or 2 that do not contain ancestral lineages (unoccupied demes). Further, \mathbf{A} is a stochastic matrix, which means that the sum of the entries in each row of \mathbf{A} is equal to one. The matrix \mathbf{B} captures the slow but crucial events in the ancestral process in which one lineage migrates to an occupied deme or to a deme in population 3, which might be occupied or not. In the case of migration to an occupied deme, it is possible for the incoming lineage to coalesce with one of the resident lineages. In addition, \mathbf{B}/D contains the order $1/D$ parts of Π_D associated with the non-zero entries of \mathbf{A} , so that the sum of the entries in each row of \mathbf{B} is equal to zero. The matrix \mathbf{C}_D captures the events in the ancestral process that become extremely improbable as D grows: for example, two or more migration events to occupied demes in a single generation.

The structure of Eq. (5) means we can apply a convergence theorem for discrete-time Markov processes with two time scales, due to Möhle (1998), to obtain an ancestral limit process in which time becomes continuous and is measured in units of D generations. Under this framework, the limit process depends on the fast events only through the equilibrium matrix $\mathbf{P} = \lim_{r \rightarrow \infty} \mathbf{A}^r$, such that the limit process has transition probability matrix

$$\Pi(t) := \lim_{D \rightarrow \infty} \Pi_D^{[Dt]} = \mathbf{P}e^{\mathbf{G}t}, \tag{6}$$

with infinitesimal generator $\mathbf{G} = \mathbf{PBP}$.

With the classification of states into sets S_1 through S_5 described in Section 2, the matrices \mathbf{A} and \mathbf{B} have a block

structure. For example,

$$\mathbf{B} = \begin{pmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} & \mathbf{B}_{13} & \mathbf{B}_{14} & \mathbf{B}_{15} \\ \mathbf{B}_{21} & \mathbf{B}_{22} & \mathbf{B}_{23} & \mathbf{B}_{24} & \mathbf{B}_{25} \\ \mathbf{B}_{31} & \mathbf{B}_{32} & \mathbf{B}_{33} & \mathbf{B}_{34} & \mathbf{B}_{35} \\ \mathbf{B}_{41} & \mathbf{B}_{42} & \mathbf{B}_{43} & \mathbf{B}_{44} & \mathbf{B}_{45} \\ \mathbf{B}_{51} & \mathbf{B}_{52} & \mathbf{B}_{53} & \mathbf{B}_{54} & \mathbf{B}_{55} \end{pmatrix}, \tag{7}$$

in which the entries of the matrix \mathbf{B}_{ij}/D are the transition probabilities, of order $1/D$, from a state in S_i to a state in S_j . The presentation of the non-zero entries of these matrices occupies the bulk of this section. Note that we will not deal explicitly with the order of states in any of these matrices. The statement that the lineages change state from (n_1, n_2, n_3) to (n'_1, n'_2, n'_3) with probability p should be understood to mean that p is the entry in “row (n_1, n_2, n_3) ” and “column (n'_1, n'_2, n'_3) ” of the transition matrix Π_D .

4.1. Non-zero entries of P

With respect to the transition probabilities of order 1, which are contained in the stochastic matrix \mathbf{A} , the set of states S_1 is absorbing. In other words, to get out of S_1 the lineages must undergo a transition with probability of order $1/D$ or smaller: one or more migration events to an occupied deme or to population 3. Therefore, the matrix $\mathbf{P} = \lim_{r \rightarrow \infty} \mathbf{A}^r$ has non-zero entries only in the left-most blocks:

$$\mathbf{P} = \begin{pmatrix} \mathbf{I} & 0 & 0 & 0 & 0 \\ \mathbf{P}_{21} & 0 & 0 & 0 & 0 \\ \mathbf{P}_{31} & 0 & 0 & 0 & 0 \\ \mathbf{P}_{41} & 0 & 0 & 0 & 0 \\ \mathbf{P}_{51} & 0 & 0 & 0 & 0 \end{pmatrix}, \tag{8}$$

where $\mathbf{I} (= \mathbf{P}_{11})$ is the identity matrix. The set S_1 contains all of the absorbing states (again, with respect to \mathbf{A}) and these can be reached from any state outside of S_1 through a series of coalescence events or migration events to unoccupied demes, whose probabilities are of order 1.

Non-zero entries of \mathbf{P}_{21} : With probability of order 1, lineages in state $(n_1, n_2, 1)$ can move either to state $(n_1 + 1, n_2, 0)$ or to state $(n_1, n_2 + 1, 0)$. Note that the order 1 migration events of lineages within populations 1 and 2, i.e. to unoccupied demes, do not change the state of the lineages under our reduced notation. All other transitions that change the state have probabilities of order $1/D$ or smaller. Since migration from demes in population 3 occurs according to the island model across all three populations, the non-zero entries of \mathbf{P}_{21} are

$$\begin{matrix} & \text{column } (n_1 + 1, n_2, 0) & \text{column } (n_1, n_2 + 1, 0) \\ \text{row } (n_1, n_2, 1) & \alpha_1 & \alpha_2, \end{matrix} \tag{9}$$

where again, in the limit, α_1 is the fraction of demes that are in population 1, and $\alpha_2 = 1 - \alpha_1$ is the fraction of demes

that are in population 2. Again, the transition probabilities in Eq. (9) do not depend on the migration probability m_3 because of the definition $\mathbf{P} = \lim_{r \rightarrow \infty} \mathbf{A}^r$. They are probabilities of ultimate absorption, given that enough generations have passed to guarantee that a migration event to either population 1 or population 2 has occurred.

Non-zero entries of \mathbf{P}_{31} : Order 1 transition probabilities in \mathbf{A} take lineages in state $(n_1+, n_2, 0)$ to either to state $(n_1 + 1, n_2, 0)$ by a migration event to an unoccupied deme in population 1, or to state $(n_1, n_2, 0)$ by a coalescent event. Therefore, the non-zero entries of \mathbf{P}_{31} are

$$\begin{array}{ccc} & \text{column } (n_1 + 1, n_2, 0) & \text{column } (n_1, n_2, 0) \\ \text{row } (n_1+, n_2, 0) & 1 - F_1 & F_1 \end{array} \quad (10)$$

in which

$$\begin{aligned} F_1 &= \frac{(1 - m_1)^2 / N}{1 - (1 - m_1)^2 + (1 - m_1)^2 / N} \\ &= \frac{(1 - m_1)^2}{Nm_1(2 - m_1) + (1 - m_1)^2} \end{aligned} \quad (11)$$

is the probability that a pair of lineages in the same deme in population 1 coalesce before one or the other of them migrates out of the deme.

Non-zero entries of \mathbf{P}_{41} : Similarly to Eq. (10), for the corresponding events in population 2 we have

$$\begin{array}{ccc} & \text{column } (n_1, n_2 + 1, 0) & \text{column } (n_1, n_2, 0) \\ \text{row } (n_1, n_2+, 0) & 1 - F_2 & F_2, \end{array} \quad (12)$$

where F_2 is given by Eq. (11) but with a change of subscripts ($1 \rightarrow 2$).

Non-zero entries of \mathbf{P}_{51} : The states in S_5 are characterized by having more than two lineages in a single deme, or more than one deme containing more than one lineage, or more than one lineage in population 3. Most DNA sequence data sets are of this sort because sampling is not usually limited to a single sequence from each sampled deme. Samples that begin in S_5 will go through a series of coalescent events and migration events to unoccupied demes, with probabilities of order 1, until all remaining lineages are in separate demes and none are in population 3. A process of this sort was called the “scattering phase” in Wakeley (1999). Again, once the lineages are in a state in S_1 , only a transition of order $1/D$ or smaller can change their state. Further, in the limit, the collection of states in S_5 is visited only once, at the time of sampling, because transitions from S_1 to S_5 occur with probabilities of order $1/D^2$.

No simple expressions appear possible for many of the non-zero entries in \mathbf{P}_{51} due to our minimal assumptions about N and m_i . With the additional assumption that $\lim_{N \rightarrow \infty} Nm_i$ is non-zero and finite, then simple expressions are available (Wakeley, 1999), and these are closely related to the Ewens sampling formula (Ewens,

1972). We do not make this additional assumption, and we present only the non-zero entries of \mathbf{P}_{51} that are associated with a particular set of states, (n_1, n_2, n_3) , where $n_3 > 1$ lineages are in population 3 and these are in n_3 distinct demes. These have transition probabilities

$$\begin{array}{ccc} & \text{column } (n_1 + k, n_2 + n_3 - k, 0) & \\ \text{row } (n_1, n_2, n_3) & \binom{n_3}{k} \alpha_1^k \alpha_2^{n_3 - k} & (13) \end{array}$$

for $k = 0, 1, \dots, n_3$. Each lineage stays in its deme for a geometrically distributed number of generations, with mean $1/m_3$, then migrates to a deme chosen uniformly at random from the $D = D_1 + D_2 + d$ demes in the total population. If migration is stronger in the direction of population 1, which here means the fraction α_1 of demes that are in population 1 is high, then lineages will tend to migrate to population 1 as they trace their ancestry back in time.

4.2. Non-zero entries of \mathbf{B}

Due to the structure of \mathbf{P} given in Eq. (8), and the form of the infinitesimal generator, $\mathbf{G} = \mathbf{PBP}$, there is no need to compute the entries of \mathbf{B}_{ij} for $i > 1$. In addition, because transitions from S_1 to S_5 require events with probabilities of order $1/D^2$, which are thus contained in the matrix \mathbf{C}_D , we have $\mathbf{B}_{15} = 0$. Therefore, we need only calculate the entries of \mathbf{B}_{11} , \mathbf{B}_{12} , \mathbf{B}_{13} , and \mathbf{B}_{14} in order to describe the limit process.

Non-zero entries of \mathbf{B}_{11} : Off the main diagonal, these correspond to events in which a lineage migrates to an occupied deme in population 1 or 2 and there is an immediate coalescent event. We have

$$\begin{array}{ccc} & \text{column } (n_1 - 1, n_2, 0) & \text{column } (n_1, n_2 - 1, 0) \\ \text{row } (n_1, n_2, 0) & \binom{n_1}{2} \frac{m_1(2 - m_1)}{\alpha_1 N} & \binom{n_2}{2} \frac{m_2(2 - m_2)}{\alpha_2 N}. \end{array} \quad (14)$$

The diagonal entries of \mathbf{B}_{11} are also non-zero, and represent what was ignored in computing $\mathbf{A}_{11} = \mathbf{P}_{11} = \mathbf{I}$. These entries are actually equal to minus the sums of all other entries of \mathbf{B} on the same rows but it is not necessary to calculate them because they do not enter into the calculation of rates in the limit process.

Non-zero entries of \mathbf{B}_{12} : These correspond to events in which a lineage migrates to a deme in population 3. We have

$$\begin{array}{ccc} & \text{column } (n_1 - 1, n_2, 1) & \text{column } (n_1, n_2 - 1, 1) \\ \text{row } (n_1, n_2, 0) & n_1 m_1 d / \alpha_1 & n_2 m_2 d / \alpha_2. \end{array} \quad (15)$$

Non-zero entries of \mathbf{B}_{13} : These correspond to events in which a lineage in population 1 migrates to an occupied deme in population 1 without an immediate coalescent

event. We have

$$\text{row } (n_1, n_2, 0) \begin{matrix} \text{column } ((n_1 - 1)+, n_2, 0) \\ \binom{n_1}{2} \frac{m_1(2 - m_1)N - 1}{\alpha_1 N} \end{matrix} \quad (16)$$

Non-zero entries of \mathbf{B}_{14} : These correspond to events in which a lineage in population 2 migrates to an occupied deme in population 2 without an immediate coalescent event. We have

$$\text{row } (n_1, n_2, 0) \begin{matrix} \text{column } (n_1, (n_2 - 1)+, 0) \\ \binom{n_2}{2} \frac{m_2(2 - m_2)N - 1}{\alpha_2 N} \end{matrix} \quad (17)$$

4.3. The limit process with generator $\mathbf{G} = \mathbf{PBP}$

From Eq. (5) and the entries of \mathbf{P} given above, it is clear that the ancestral lineages will spend the majority of their history in the set of states S_1 . Specifically, the single-generation probability of leaving S_1 is of order $1/D$ while the single-generation probability of entering S_1 is of order 1. In the limit as D tends to infinity, the dynamics are given by Eq. (6). First, all samples undergo an instantaneous adjustment by the matrix \mathbf{P} , which leaves them in S_1 . After this scattering phase, all further transitions in the limit process are between states in S_1 (from the equation $\mathbf{G} = \mathbf{PBP}$ we have $\mathbf{G}_{ij} = 0$ for $j > 1$). They are either migration events between populations 1 and 2 or coalescent events within populations 1 or 2. Viewed in the context of the original discrete-time process these include compound events in which the lineages move from a state in S_1 to a state in either $S_2, S_3,$ or $S_4,$ then back to another state in S_1 . The rates of these transitions are the entries of

$$\mathbf{G}_{11} = \mathbf{B}_{11} + \mathbf{B}_{12}\mathbf{P}_{21} + \mathbf{B}_{13}\mathbf{P}_{31} + \mathbf{B}_{14}\mathbf{P}_{41}, \quad (18)$$

and we use this equation to compute the overall rates of coalescence and migration for the sample $(n_1, n_2, 0)$.

Coalescence among ancestral lineages in population 1 can occur directly as a single step, with a rate given in \mathbf{B}_{11} (see Eq. (14)). Otherwise, it can occur as the two step event $(n_1, n_2, 0) \rightarrow ((n_1 - 1)+, n_2, 0) \rightarrow (n_1 - 1, n_2, 0)$ whose rate is contained in $\mathbf{B}_{13}\mathbf{P}_{31}$ and is the product of terms from Eqs. (16) and (10). The rate of coalescence among ancestral lineages in population 2 is obtained in the same way, as the sum of terms in \mathbf{B}_{11} and in $\mathbf{B}_{14}\mathbf{P}_{41}$. Under the assumptions we have made, there is no single-step chance for migration between population 1 and 2 in the discrete-time process. Migration of an ancestral lineage from population 1 to population 2, backwards in time, occurs via the two-step event $(n_1, n_2, 0) \rightarrow (n_1 - 1, n_2, 1) \rightarrow (n_1 - 1, n_2 + 1, 0)$ in which a lineage resides temporarily in population 3. The rate of this event in the limit process is the product of terms from Eqs. (15) and (9). The rate for migration of an ancestral lineage from population 2 to population 1 is also given by an entry in $\mathbf{B}_{12}\mathbf{P}_{21}$. In all, after some

simplification we have

$$(n_1, n_2, 0) \rightarrow \begin{cases} (n_1 - 1, n_2, 0) & \text{with rate } \binom{n_1}{2} \frac{1 - F_1}{\alpha_1 N}, \\ (n_1, n_2 - 1, 0) & \text{with rate } \binom{n_2}{2} \frac{1 - F_2}{\alpha_2 N}, \\ (n_1 - 1, n_2 + 1, 0) & \text{with rate } n_1 dm_1 \frac{\alpha_2}{\alpha_1}, \\ (n_1 + 1, n_2 - 1, 0) & \text{with rate } n_2 dm_2 \frac{\alpha_1}{\alpha_2} \end{cases} \quad (19)$$

for the entries of \mathbf{G}_{11} . Again, these are the only entries of \mathbf{G} that contribute to the ancestral limit process.

Recall that the limit process in this section has time measured in units of D generations, whereas in Section 3 time was measured in units of ND generations. Since N is a constant, we simply rescale time in Eq. (19) by N to obtain Eq. (2).

5. Discussion

We have shown that the structured coalescent (Notohara, 1990; Herbots, 1994, 1997; Wilkinson-Herbots, 1998) forms part of the genetic ancestry of a sample taken from a pair of large subdivided populations that are connected by migration to and from a small number of subpopulations, or demes. The ancestral limit process is related to the many-demes coalescent described in Wakeley (1998, 1999), and more recently in Lessard and Wakeley (2004). In particular, the ancestral limit process for a sample involves an instantaneous adjustment, or scattering phase, in which a limited number of coalescent events and migration events can occur, and which ends when the remaining ancestral lineages are each in a separate deme and none are in population 3. The rest of the ancestry of the sample is given by a two-population structured coalescent whose parameters depend on the details of the model.

Importantly, this structured coalescent is obtained even when the single-generation migration probabilities are large between all pairs of demes that can exchange migrants. This stands in contrast to the usual derivation of the structured coalescent, which requires migration rates to be low (on the order of the inverse of the population size). The “low” rate of migration in our model arises because genetic lineages must pass through a relatively small set of demes accessible from both large subdivided populations. Thus, it is a constriction in the habitat that produces structure between the two large assemblages of demes. This justifies the use of the structured coalescent in situations where at first it would seem inappropriate. However, we emphasize that: (1) the parameters estimated from data will not have the usual interpretations, and (2)

the structured coalescent will also include a scattering phase for most samples.

Our result, which depends on a “separation of time scales” between events of order 1 and events of order $1/D$, can be extended to more general situations (e.g. more than two large subdivided populations). Here we consider one case of particular interest in some detail. It should be clear from the previous section that only some of the d demes in population 3 need to be directly accessible via migration from populations 1 and 2, as long as d is finite and the number of accessible demes is non-zero in the limit. In addition, migration among demes in population 3 may assume any pattern, as long as a lineage in population 3 moves either to population 1 or to population 2 in a finite number of generations.

Consider two island-model populations, such as our populations 1 and 2, connected by a stepping-stone (Kimura and Weiss, 1964) migration corridor. Specifically, let us assume that the d demes in population 3 are arranged in a line and labeled 1 through d . Further, assume that deme 1 is the only deme accessible from population 1, and that deme d is the only deme accessible from population 2. If migration occurs between neighboring demes along the corridor in the direction of population 1 with probability μ_1 and in the direction of population 2 with probability μ_2 , then the dynamics of a lineage passing through population 3 are those of a random walk in one dimension. Populations 1 and 2 are the absorbing boundaries, and the process is identical to the classical gambler’s ruin problem; e.g. see Feller (1968, p. 344). The probability that a lineage now in deme i of the corridor exits to population 2 is given by

$$p(i) = \begin{cases} \frac{1 - (\mu_2/\mu_1)^i}{1 - (\mu_2/\mu_1)^{d+2}} & \text{if } \mu_2 \neq \mu_1, \\ \frac{i}{d+2} & \text{if } \mu_2 = \mu_1 \end{cases} \quad (20)$$

for $i = 1, 2, \dots, d$.

Eq. (20) describes the scattering phase for a single sample from population 3, and would also enter into the calculation of the migration rate $M_{ij}^{(3)}$ for this model. A scenario like this is implicit in the discussion of patterns of variation in human mitochondrial DNA (mtDNA) at different sampling points along the Nile river valley by Krings et al. (1999). In particular, Krings et al. (1999) defined mtDNA haplotypes to be of either “northern” or “southern” origin, and showed that the proportions of different haplotypes depended on the distance from the northern and southern source populations.

Acknowledgments

J.W. was supported by a Presidential Early Career Award for Scientists and Engineers (DEB-0133760) from the National Science Foundation. S.L. was supported by grants from the Natural Sciences and Engineering

Research Council of Canada. We thank three anonymous reviewers for their comments.

References

- Bahlo, M., Griffiths, R.C., 2000. Inference from gene trees in a subdivided population. *Theor. Popul. Biol.* 57, 79–95.
- Beerli, P., Felsenstein, J., 2001. Maximum-likelihood estimation of a migration matrix and effective population sizes in n populations by using a coalescent approach. *Proc. Natl. Acad. Sci. USA* 98, 4563–4568.
- Carmelli, D., Cavalli-Sforza, L.L., 1976. Some models of population structure and evolution. *Theor. Popul. Biol.* 9, 329–359.
- Ewens, W.J., 1972. The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.* 3, 87–112.
- Feller, W., 1968. *An Introduction to Probability Theory and Its Applications*, third ed., vol. I. Wiley, New York.
- Fisher, R.A., 1930. *The Genetical Theory of Natural Selection*. Clarendon, Oxford.
- Herbots, H.M., 1994. Stochastic models in population genetics: genealogical and genetic differentiation in structured populations. Ph.D. Thesis, University of London.
- Herbots, H.M., 1997. The structured coalescent. In: Donnelly, P., Tavaré, S. (Eds.), *Progress in Population Genetics and Human Evolution*, IMA Volumes in Mathematics and its Applications, vol. 87. Springer, New York, pp. 231–255.
- Hey, J., Nielsen, R., 2004. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics* 167, 747–760.
- Hudson, R.R., 1983. Testing the constant-rate neutral allele model with protein sequence data. *Evolution* 37, 203–217.
- Kimura, M., Weiss, G.H., 1964. The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics* 49, 561–576.
- Kingman, J.F.C., 1982. The coalescent. *Stochastic Process. Appl.* 13, 235–248.
- Krings, M., Salem, A.H., Bauer, K., Geisert, H., Malek, A.K., Chaix, L., Simon, C., Welsby, D., Di Rienzo, A., Uterman, G., Sajantila, A., Pääbo, S., Stoneking, M., 1999. MtDNA analysis of Nile river valley populations: a genetic corridor or a barrier to migration? *Am. J. Hum. Genet.* 64, 1166–1176.
- Lessard, S., Wakeley, J., 2004. The two-locus ancestral graph in a subdivided population: convergence as the number of demes grows in the island model. *J. Math. Biol.* 48, 275–292.
- Möhle, M., 1998. A convergence theorem for Markov chains arising in population genetics and the coalescent with partial selfing. *Adv. Appl. Probab.* 30, 493–512.
- Nagylaki, T., 1980. The strong-migration limit in geographically structured populations. *J. Math. Biol.* 9, 101–114.
- Nagylaki, T., 1998. The expected number of heterozygous sites in a subdivided population. *Genetics* 149, 1599–1604.
- Nei, M., 1973. Analysis of gene diversity in subdivided populations. *Proc. Natl. Acad. Sci. USA* 70, 3321–3323.
- Nielsen, R., Wakeley, J., 2001. Distinguishing migration from isolation: a Markov Chain Monte Carlo approach. *Genetics* 158, 885–896.
- Nordborg, M., Krone, S.M., 2002. Separation of time scales and convergence to the coalescent in structured populations. In: Slatkin, M., Veuille, M. (Eds.), *Modern Developments in Theoretical Population Genetics*. Oxford University Press, Oxford, UK.
- Notohara, M., 1990. The coalescent and the genealogical process in geographically structured population. *J. Math. Biol.* 29, 59–75.
- Notohara, M., 1993. The strong migration limit for the genealogical process in geographically structured populations. *J. Math. Biol.* 31, 115–122.
- Sawyer, S.A., Felsenstein, J., 1983. Isolation by distance in a hierarchically clustered population. *J. Appl. Probab.* 20, 1–10.

- Slatkin, M., 1987. Gene flow and the geographic structure of natural populations. *Science* 236, 787–792.
- Slatkin, M., Voelm, L., 1991. F_{ST} in a hierarchical island model. *Genetics* 127, 627–629.
- Strobeck, C., 1987. Average number of nucleotide differences in a sample from a single subpopulation: a test for population subdivision. *Genetics* 117, 149–153.
- Tajima, F., 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105, 437–460.
- Vollmer, S., Palumbi, S.R., 2002. Hybridization and the evolution of reef coral diversity. *Science* 296, 2023–2025.
- Wakeley, J., 1998. Segregating sites in Wright's island model. *Theor. Popul. Biol.* 53, 166–175.
- Wakeley, J., 1999. Non-equilibrium migration in human history. *Genetics* 153, 1863–1871.
- Wakeley, J., 2000. The effects of population 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.
- Wilkinson-Herbots, H.M., Ettridge, R., 2004. The effect of unequal migration rates on F_{ST} . *Theor. Popul. Biol.* 66, 185–197.
- Wright, S., 1931. Evolution in Mendelian populations. *Genetics* 16, 97–159.
- Wright, S., 1943. Isolation by distance. *Genetics* 28, 114–138.