J. Math. Biol. (1998) 37: 535–585

Journal of
Mathematical
Biology

Genealogy and subpopulation differentiation under various models of population structure

Hilde M. Wilkinson-Herbots

Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, UK. e-mail: hh@stats.ucl.ac.uk

Received: 1 October 1997 / Revised version: 15 March 1998

Abstract. The structured coalescent is used to calculate some quantities relating to the genealogy of a pair of homologous genes and to the degree of subpopulation differentiation, under a range of models of subdivided populations and assuming the infinite alleles model of neutral mutation. The classical island and stepping-stone models of population structure are considered, as well as two less symmetric models. For each model, we calculate the Laplace transform of the distribution of the coalescence time of a pair of genes from specified locations and the corresponding mean and variance. These results are then used to calculate the values of Wright's coefficient F_{ST}, its limit as the mutation rate tends to zero and the limit of its derivative with respect to the mutation rate as the mutation rate tends to zero. From this derivative it is seen that F_{ST} can depend strongly on the mutation rate, for example in the case of an essentially one-dimensional habitat with many subpopulations where gene flow is restricted to neighbouring subpopulations.

Key words: Population structure – Genealogy – Coalescent – Subpopulation differentiation – F_{ST}

1 Introduction

The effect of population structure on the genetic composition of a population has traditionally been analyzed in terms of Wright's hierarchical F-statistics, which are essentially inbreeding coefficients. Studying a population subdivided into distinct colonies, Wright (1951) separated the respective contributions towards inbreeding of non-random mating within colonies and of the population subdivision

itself. The inbreeding-like effect of the population subdivision is represented by Wright's coefficient F_{ST}. At present there appears to be some confusion about the precise meaning of F_{ST} and several definitions are currently in use (see for example Chakraborty and Danker-Hopfe 1991 for a review; see also Nagylaki 1998a). For alleles at a single locus, F_{ST} has been expressed in terms of probabilities of identity as:

$$F_{ST} = \frac{f_0 - \bar{f}}{1 - \bar{f}} \tag{1}$$

(Wright 1951; Nei 1973; Slatkin 1985, 1991), where f_0 is the probability that two genes sampled at random from a single subpopulation are identical (carry the same allele), and $\bar{f}$ is the probability that two genes randomly chosen from the collection of subpopulations considered are identical. In this paper we shall not be concerned with the meaning and interpretation of F_{ST} – we shall simply adopt equation (1) as its definition.

The coefficient F_{ST} serves as a measure of the genetic differentiation among subpopulations. It is used as a correction factor for population structure in forensic calculations of DNA profile match probabilities (Nichols and Balding 1991; Balding et al. 1994; Balding and Nichols 1994, 1995; Foreman et al. 1997 and references therein; also Morton 1992 and Weir 1994). Another major application of F_{ST} in population genetics is to estimating the level of gene flow between subpopulations (Slatkin 1985 and references therein; Slatkin and Barton 1989; Slatkin 1991, 1993; Cockerham and Weir 1993). Denoting by N the number of diploid individuals per subpopulation and by m the fraction of each subpopulation that is replaced by immigrants every generation, the "effective" level of gene flow is estimated using the formula

$$Nm \approx \frac{1}{4}\left(\frac{1}{F_{ST}} - 1\right), \tag{2}$$

which is based on the island model of population structure (Wright 1931, 1951; see also Sect. 4.1) and the neutral Wright–Fisher model of reproduction. Estimates of Nm obtained for pairs of subpopulations can be used to detect isolation by distance in a natural population and to test specific hypotheses about the structure and history of the population (Slatkin 1993). As equation (2) assumes the island model with a large number of subpopulations and a small mutation rate, it is important to understand how F_{ST} depends on the real structure of the population and on the mutation rate. In this paper the theoretical value of F_{ST} is calculated for a range of models of population structure with the infinite alleles model of neutral mutation, showing quite

a different relationship between F_{ST} and Nm according to the model assumed. A detailed discussion (based on results given in this paper), including many figures, of how F_{ST} depends on various parameters of population structure and mutation is found in Herbots (1994).

Slatkin (1991) brought the relationship between F_{ST} and genealogy to the foreground and introduced an approximation for F_{ST} in terms of the mean time since the most recent common ancestor (the mean "coalescence time") of a pair of genes from a single subpopulation and that of a pair of genes from the collection of subpopulations considered. Calculating these mean coalescence times he obtained approximate values of F_{ST} for several models of population subdivision. Underlying this work is the so-called "coalescent approach" (for reviews see Donnelly and Tavaré 1995 and references therein) which is an important recent development in population genetics modeling. The basic idea is to explicitly consider the genealogy of a sample of genes from the current population, tracing back their ancestry and focussing on the events in the past when two or more genes in the sample are descended from the same ancestral gene.

The "coalescent" (Kingman 1982a,b,c) is a stochastic process which is a good description of the genealogy of a panmictic population evolving according to one of a broad class of reproductive models. This process has been extended to incorporate certain forms of population structure (Takahata 1988; Notohara 1990; Herbots 1994, 1997; Barton and Wilson 1995a,b). This paper is based on the "structured coalescent", a stochastic process which describes the genealogical relationships in a subdivided population, under reasonable assumptions about reproduction and migration.

Section 2 contains a brief description of the structured coalescent. In the remaining sections we assume the infinite alleles model of neutral mutation. Slatkin's approximation for F_{ST} is the limit of F_{ST} as the mutation rate tends to zero, which can be expressed in terms of the mean coalescence times of pairs of genes sampled within or among subpopulations (Slatkin 1991). Similarly, the "exact"[1] value of F_{ST} can be expressed in terms of the Laplace transform of the distribution of the coalescence time of two genes sampled at random from a single subpopulation, and that of two genes randomly sampled from the collection of subpopulations considered. This is done in Sect. 3. Being equivalent to the distributions of these coalescence times, these Laplace transforms are also interesting in their own right, as are the

[1] Subject only to the diffusion time-scale approximation (which essentially consists in taking the limit as the subpopulation sizes tend to infinity) inherent in the coalescent approach

corresponding means and variances. Because it is commonly believed that for small mutation rates, F_{ST} is nearly independent of the mutation rate (for example, Slatkin 1985; Slatkin and Barton 1989), we will also consider the derivative of F_{ST} with respect to the mutation rate, in the limit as the mutation rate tends to zero. This limit can be expressed in terms of the means and variances of the coalescence times of pairs of genes (see Sect. 3). In subsequent sections we consider the finite island model and the finite and infinite stepping-stone models in one and two dimensions (Sect. 4) and two less symmetric models of population structure (Sect. 5). For each model we calculate the Laplace transform of the distribution of the coalescence time of a pair of genes, the mean and the variance of this coalescence time, Slatkin's approximation for F_{ST}, the "exact" value of F_{ST} (as a function of the mutation rate) and its derivative with respect to the mutation rate in the limit as the mutation rate tends to zero. For completeness and ease of reference we list results that have been found before as well as new results. In Sect. 6 we discuss some of the results.

The values presented in this paper apply to selectively neutral genes at a single locus, subject to mutation and migration but without intragenic recombination. They are exact only for haploid species, and for diploid species with exclusively gametic migration. Nagylaki (1983, 1998b) and Sawyer (1976; stepping-stone model) have set out conditions under which models of truly diploid migration are well approximated by the model of gamete migration.

2 The structured coalescent

Consider a haploid population divided into a finite or infinite number of subpopulations which are all large and panmictic and which are partially isolated from each other. We denote by $\mathscr{S}$ the set of the subpopulation labels, where we assume that $\mathscr{S}$ is countable. We consider genes at a single locus. (For our purposes, a gene is simply a non-recombining piece of DNA.) The number of homologous genes in subpopulation i is $N_i = 2a_iN$, where a_i is a positive integer constant and N is large. Assume that the population evolves in discrete non-overlapping generations. At a particular generation which we call time zero, we draw a sample of n_0 genes from the total population (where n_0 is finite and fixed) and we trace the ancestry of the genes in the sample. At each time in the past, we count how many distinct ancestors the n_0 sampled genes have in each subpopulation. We denote by $\alpha_N^i(\tau)$ the number of distinct ancestors the sample has in subpopulation i, τ generations ago, and by $\boldsymbol{\alpha}_N(\tau)$ the ordered set $(\alpha_N^i(\tau))_{i\in\mathscr{S}}$, with a

component for each subpopulation. If there are n subpopulations and n is finite, then $\boldsymbol{\alpha}_N(\tau)$ is an n-tuple. If the number of subpopulations is infinite, $\boldsymbol{\alpha}_N(\tau)$ is a sequence with index set $\mathscr{S}$. In standard mathematical notation, we write that $\boldsymbol{\alpha}_N(\tau) \in \mathbb{N}^{\mathscr{S}}$, which is the set of all functions from $\mathscr{S}$ to $\mathbb{N}$, where $\mathbb{N}$ is the set of the natural numbers, including zero. We consider the "ancestral process" $\boldsymbol{\alpha}_N = \{\boldsymbol{\alpha}_N(\tau)\colon \tau = 0, 1, 2, \ldots\}$.

Tracing the ancestral lineages of the genes in the sample, two types of events can occur. Two particular lineages can coalesce at the most recent common ancestor of the corresponding genes in the sample (this can only occur when these lineages reside in the same subpopulation), in which case the number of distinct ancestors in that subpopulation (i.e. the value of $\alpha_N^i(\tau)$) decreases by one. The rate at which such a coalescence event occurs is, for many exchangeable models of reproduction (Cannings 1974), inversely proportional to the size of the subpopulation. When an ancestor in subpopulation i is an immigrant from subpopulation j (which we also describe as a "migration" of the ancestor from subpopulation i to subpopulation j backward in time), the number of distinct ancestors in subpopulation i decreases by one, while that in subpopulation j increases by one. We denote by $\boldsymbol{\varepsilon}^i$ the element of $\mathbb{N}^{\mathscr{S}}$ with components

$$(\boldsymbol{\varepsilon}^i)_j = \delta_{ij} = \begin{cases} 1 & \text{if } j = i \\ 0 & \text{otherwise} \end{cases}$$

and we define addition and subtraction of elements of $\mathbb{N}^{\mathscr{S}}$ to be component-wise, i.e. the sum or difference of two sequences (or n-tuples) in $\mathbb{N}^{\mathscr{S}}$ is obtained by adding or subtracting their corresponding components. If $\boldsymbol{\alpha}_N(\tau) = \boldsymbol{\alpha}$ and two lineages in subpopulation i coalesce, the value of $\boldsymbol{\alpha}_N(\tau)$ is changed to $\boldsymbol{\alpha} - \boldsymbol{\varepsilon}^i$; the migration of an ancestral lineage from subpopulation i to subpopulation j (backward in time) changes the value of $\boldsymbol{\alpha}_N(\tau)$ from $\boldsymbol{\alpha}$ to $\boldsymbol{\alpha} - \boldsymbol{\varepsilon}^i + \boldsymbol{\varepsilon}^j$.

Under reasonable assumptions about reproduction and migration, the ancestral process $\boldsymbol{\alpha}_N$ is, with the appropriate re-scaling of time, well approximated by the "**structured coalescent**", which is the continuous-time Markov chain $\{\boldsymbol{\alpha}(t)\colon t \geqq 0\}$ with Q-matrix $\mathbf{Q}$ whose entries are

$$Q_{\boldsymbol{\alpha},\boldsymbol{\beta}} = \begin{cases} -\sum_{i \in \mathscr{S}} \left\{ \alpha_i \dfrac{M_i}{2} + \dfrac{1}{c_i} \binom{\alpha_i}{2} \right\} & \text{if } \boldsymbol{\beta} = \boldsymbol{\alpha} \\ \alpha_i \dfrac{M_{ij}}{2} & \text{if } \boldsymbol{\beta} = \boldsymbol{\alpha} - \boldsymbol{\varepsilon}^i + \boldsymbol{\varepsilon}^j \ (j \neq i) \\ \dfrac{1}{c_i} \binom{\alpha_i}{2} & \text{if } \boldsymbol{\beta} = \boldsymbol{\alpha} - \boldsymbol{\varepsilon}^i \\ 0 & \text{otherwise,} \end{cases} \quad . \qquad (3)$$

where $M_{ij}/2$ is the scaled migration rate of a lineage from subpopulation i to subpopulation j backward in time (the factor $1/2$ is standard and is convenient in applications involving pairs of genes), $M_i = \sum_{j \neq i} M_{ij}$ and $1/c_i$ is the coalescence rate of a pair of lineages in subpopulation i (in many cases, the c_i correspond to the relative subpopulation sizes) – see Notohara (1990) and Herbots (1994, 1997). For example, consider the case where reproduction in each subpopulation follows the neutral Wright–Fisher model and a constant proportion q_{ij} of the individuals born in subpopulation i migrate to subpopulation j every generation ($i, j \in \mathscr{S}$) where $\forall i \in \mathscr{S}$: $a_i \sum_{j \neq i} q_{ij} = \sum_{j \neq i} a_j q_{ji}$ (i.e. the size of each subpopulation is maintained under migration – recall that a_i is the relative size of subpopulation i). Measuring time in units of $2N$ generations, the ancestral process is well approximated by the structured coalescent: it is proved in Herbots (1994, 1997) that under appropriate technical assumptions, the (re-scaled) ancestral process $\{\boldsymbol{\alpha}_N([2Nt]) : t \geqq 0\}$ converges weakly, as $N \to \infty$, to the structured coalescent $\{\boldsymbol{\alpha}(t) : t \geqq 0\}$ with Q-matrix $\mathbf{Q}$ given by (3) where $c_i = a_i$ and $M_{ij} = \lim_{N \to \infty} (4N \frac{c_j}{c_i} q_{ji})$ for all $i, j \in \mathscr{S}$.

The remainder of this paper does not assume any specific model for reproduction and migration in discrete time – it does assume that (with adequate time-scaling) the structured coalescent is an appropriate description of the genealogy of a sample from the population.

3 F_{ST} in terms of coalescence times

In the remaining sections we assume the infinite alleles model of neutral mutation, in which every mutant gene is assumed to be of a novel type. The coefficient F_{ST} can be related to the respective coalescence times of pairs of genes sampled within and among subpopulations. Slatkin (1991) did this in an approximate way and it can also be done exactly.

3.1 The exact F_{ST} value

In the definition of F_{ST}, equation (1), f_0 and $\bar{f}$ are, respectively, the probability of identity of a pair of genes sampled at random from a single subpopulation and that of a pair of genes randomly sampled from the collection of subpopulations considered. The "collection of subpopulations considered" is in general either the total population (as is usually the case for the finite island model) or a pair of

subpopulations a specified distance apart (as is common for stepping-stone models). Under the infinite alleles model, the probability of identity of two genes is the probability that since their descent from a common ancestor, neither gene has undergone a mutation. Sampling two genes at random from the same subpopulation, the distribution of the time T_0 since their most recent common ancestor follows from the structured coalescent. Given T_0, the probability of identity of the two genes is the probability that no mutation has occurred on either gene's lineage during time T_0. In the coalescent approximation, the probability that a particular gene has not mutated during time T_0 is $e^{-\theta/2\,T_0}$, where θ is the scaled mutation rate (it is assumed that each gene mutates at scaled rate $\theta/2$). For example, if we assume that each subpopulation contains $2N$ genes and evolves according to the neutral Wright–Fisher model, time-scaling is in units of $2N$ generations and $\theta = \lim_{N\to\infty}(4N\mu)$, where μ is the probability of mutation per gene per generation. Assuming that different genes mutate independently, it follows that

$$f_0 = E[e^{-\theta T_0}] \tag{4}$$

(Hudson 1990). Similarly

$$\bar{f} = E[e^{-\theta T}]\,, \tag{5}$$

where T is the coalescence time of two genes randomly sampled from the collection of subpopulations considered. Note that (4) and (5) are the Laplace transforms of the distributions of T_0 and T, respectively, evaluated in θ. The coefficient F_{ST} can then be written as

$$F_{ST} = \frac{E[e^{-\theta T_0}] - E[e^{-\theta T}]}{1 - E[e^{-\theta T}]}\,. \tag{6}$$

As the structured coalescent is a continuous-time Markov chain, conditioning on its first jump easily gives the following system of linear equations: denoting by T_{ij} the coalescence time of a gene from subpopulation i and a gene from subpopulation j,

$$\begin{cases} \left(\dfrac{1}{c_i} + M_i + s\right) E[e^{-sT_{ii}}] - \displaystyle\sum_{k\neq i} M_{ik}\, E[e^{-sT_{ik}}] = \dfrac{1}{c_i} \\ \left(\dfrac{M_i}{2} + \dfrac{M_j}{2} + s\right) E[e^{-sT_{ij}}] - \displaystyle\sum_{k\neq i} \frac{M_{ik}}{2}\, E[e^{-sT_{jk}}] - \sum_{k\neq j} \frac{M_{jk}}{2}\, E[e^{-sT_{ik}}] = 0\,, \end{cases} \tag{7}$$

for all $i, j \in \mathscr{S}$ with $j \neq i$ and for all $s \geqq 0$. Solving these equations gives the Laplace transform of the distribution of the coalescence time of

a pair of genes from specified locations. The probabilities of identity given by equations (4) and (5) are then obtained by taking the expectation over the locations of a pair of genes drawn at random from the same subpopulation or from the collection of subpopulations considered.

3.2 *Slatkin's approximation* $F_{ST}^{(0)}$

Slatkin (1991) suggested that F_{ST} might be approximated by its limit as the mutation rate tends to zero, which, by applying l'Hôpital's rule, can be expressed in terms of mean coalescence times as

$$\begin{aligned} F_{ST}^{(0)} &:= \lim_{\theta \downarrow 0} F_{ST} \\ &= \frac{ET - ET_0}{ET}, \end{aligned} \tag{8}$$

provided these mean coalescence times are finite.

By differentiating equations (7) with respect to s, at $s = 0$, or by conditioning on the first jump of the structured coalescent, the following equations are obtained for the mean coalescence time of a pair of genes from specified locations:

$$\begin{cases} \left(\frac{1}{c_i} + M_i\right) E[T_{ii}] - \sum_{k \neq i} M_{ik} E[T_{ik}] = 1 \\ \left(\frac{M_i}{2} + \frac{M_j}{2}\right) E[T_{ij}] - \sum_{k \neq i} \frac{M_{ik}}{2} E[T_{jk}] - \sum_{k \neq j} \frac{M_{jk}}{2} E[T_{ik}] = 1, \end{cases} \tag{9}$$

for all $i, j \in \mathscr{S}$ with $j \neq i$, provided all ET_{ij} are finite (Notohara 1990). Solving these equations, ET_0 and ET can be calculated. If f_0 and $\bar{f}$ are known as a function of θ, ET_0 and ET can also be obtained by differentiating f_0 and $\bar{f}$ with respect to θ, at $\theta = 0$.

For a range of models of population structure, $ET_0 = \sum_{i \in \mathscr{S}} c_i$ (in continuous time), independent of the precise migration pattern and migration rates (Li 1976; Strobeck 1987; Slatkin 1987; Herbots 1994; Nagylaki 1998b; see also Sect. 6). For those models the calculation of (and the resulting expression for) Slatkin's approximation $F_{ST}^{(0)}$ is relatively simple, compared to the exact F_{ST} value. Another advantage of Slatkin's approximation is the fact that it is, of course, independent of the mutation rate (which is useful since in most cases, mutation rates are unknown, but known to be very small). Furthermore, it is commonly believed that for small mutation rates, F_{ST} is nearly independent

of the mutation rate. However, as we will show in the next sections, the latter is not generally true. Nevertheless, at loci for which the infinite alleles model is reasonable, Slatkin's approximation should be accurate in many realistic situations; for criteria on and some discussion of the accuracy of $F_{ST}^{(0)}$ as an approximation for F_{ST}, see Slatkin (1993) and Herbots (1994).

3.3 The derivative of F_{ST} with respect to the mutation rate

To gain some idea of the strength of the dependence of F_{ST} on the mutation rate we also consider the derivative of F_{ST} with respect to θ, in the limit as $\theta \downarrow 0$. By repeatedly applying l'Hôpital's rule, this limit can be expressed in terms of first and second moments (or means and variances) of coalescence times of pairs of genes:

$$\lim_{\theta \downarrow 0} \frac{\partial}{\partial \theta} F_{ST} = -\frac{E[T^2]\, ET_0 - E[T_0^2]\, ET}{2(ET)^2}, \tag{10}$$

provided these second moments are finite.

Differentiating equations (7) twice with respect to s, at $s = 0$, or conditioning on the first jump of the structured coalescent, the following system of linear equations is found:

$$\begin{cases} \left(\frac{1}{c_i} + M_i\right) E[T_{ii}^2] - \sum_{k \neq i} M_{ik}\, E[T_{ik}^2] = 2E[T_{ii}] \\ \left(\frac{M_i}{2} + \frac{M_j}{2}\right) E[T_{ij}^2] - \sum_{k \neq i} \frac{M_{ik}}{2} E[T_{jk}^2] - \sum_{k \neq j} \frac{M_{jk}}{2} E[T_{ik}^2] = 2E[T_{ij}] \end{cases} \tag{11}$$

for all $i, j \in \mathscr{S}$ with $j \neq i$, provided all $E[T_{ij}^2]$ are finite. The quantities $E[T_0^2]$ and $E[T^2]$ can either be calculated from these equations, or by differentiating f_0 and $\bar{f}$ twice with respect to θ at $\theta = 0$.

4 Some symmetric models of population structure

In this section we consider the finite island model and the finite and infinite stepping-stone models in one and two dimensions. In each of these models all subpopulations have the same size. The island model assumes the same migration rate between all pairs of subpopulations. In the stepping-stone models the subpopulations are situated at the points of a finite or infinite rectangular lattice in one or more dimensions and migration occurs only between adjacent subpopulations. For

the finite stepping-stone model in one dimension, we assume that the two subpopulations at the ends are connected by migration, giving a circular stepping-stone model. Similarly, we assume that in the finite two-dimensional stepping-stone model, the subpopulations are situated on a torus. These assumptions are commonly made, in order to avoid edge-effects.

Within each of the models considered in this section, all subpopulations are identical with respect to size and migration pattern. We assume that time-scaling is such that two genes in a single subpopulation have coalescence rate 1 (working backwards in time), and we denote by M the total scaled migration rate out of each subpopulation, while θ is the scaled mutation rate. Each gene leaves its subpopulation at rate $M/2$ and mutates at rate $\theta/2$. If the subpopulations each contain $2N$ genes and evolve according to the neutral Wright–Fisher model, and if a fixed proportion m of each subpopulation is replaced by immigrants every generation (the specific migration pattern determines which subpopulations the immigrant genes come from), time-scaling is in units of $2N$ generations, $M = \lim_{N\to\infty}(4Nm)$ and $\theta = \lim_{N\to\infty}(4N\mu)$, where μ is the probability of mutation per gene per generation. So in this case, M and θ are twice the number of migrant genes and twice the expected number of mutant genes, respectively, per subpopulation[2] per generation. Under other models for reproduction, the time-scaling and hence the relationship between M and m and between θ and μ may be different.

Note that an F_{ST} value is specific to the collection of subpopulations considered (see also the definition of F_{ST}, equation (1)), which does not necessarily include all subpopulations. There are different types of F_{ST} value that appear in the literature, and the distinction is not always clearly made. For the finite island model, an F_{ST} value is usually a *global* value, in which $\bar{f}$, ET and $E[T^2]$ (see equations (1), (8) and (10)) assume sampling of pairs of genes from the *total* population. The collection of subpopulations considered is in this case the collection of all subpopulations. For stepping-stone models, F_{ST} values are often *pairwise* values, in which $\bar{f}$, ET and $E[T^2]$ are averages over a *pair* of subpopulations a specified distance apart. Because in pairwise F_{ST} values the contribution of genes from a single subpopulation

[2] For a panmictic population under the Wright–Fisher model, time is generally scaled in terms of the population size, and the scaled mutation rate, also denoted by θ, is twice the expected number of mutations in the population per generation. For a subdivided population, it is standard (see, for example, Griffiths 1981, Hudson 1990 and Notohara 1990) to scale time by subpopulation rather than population size, with θ and M defined as in this paper

(which tend to be more similar) has more weight than in global ones, pairwise F_{ST} values tend to be lower than the corresponding global F_{ST} values.

4.1 The finite island model

Under this model the population is divided into n equal-sized subpopulations ($n \geqq 2$ and finite) with the same migration rate between any two subpopulations. We assume that the genealogy of a sample from the population is well described by the structured coalescent $\{\boldsymbol{\alpha}(t): t \geqq 0\}$ with infinitesimal generator $\mathbf{Q}$ given by equation (3), where $\mathscr{S} = \{1, \ldots, n\}$ and where $c_i = 1$, $M_i = M$ and $M_{ij} = M/(n-1)$ for $i, j = 1, \ldots, n$ with $j \neq i$. So the scaled migration rate of a gene to any specific other subpopulation is $M/(2(n-1))$.

As before, we denote by T_0 the coalescence time of a pair of genes from a single subpopulation. Let T_1 be the coalescence time of a pair of genes chosen at random from different subpopulations. Equations (7) reduce to the following equations for the Laplace transforms of the distributions of T_0 and T_1:

$$\begin{cases}(1 + M + s)\, E[\mathrm{e}^{-sT_0}] - ME[\mathrm{e}^{-sT_1}] = 1 \\ \{M + (n-1)s\} E[\mathrm{e}^{-sT_1}] - ME[\mathrm{e}^{-sT_0}] = 0\,. \end{cases}$$

Their solution is given by

$$E[\mathrm{e}^{-sT_0}] = \frac{M + (n-1)s}{M + (nM + n - 1)s + (n-1)s^2} \tag{12}$$

$$E[\mathrm{e}^{-sT_1}] = \frac{M}{M + (nM + n - 1)s + (n-1)s^2} \tag{13}$$

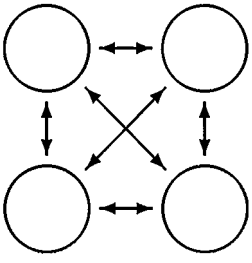

Fig. 1. Island model with $n = 4$ subpopulations. Each circle stands for a subpopulation and the arrows indicate migration

(Hudson 1990). These or similar results have been found before without consideration of the structured coalescent, by numerous authors, including Maruyama (1970; in which a small correction is needed, as was pointed out by Latter 1973), Nei (1975), Griffiths (1981), Nagylaki (1983) and Crow and Aoki (1984). The means and variances of T_0 and T_1 are obtained by differentiation of results (12) and (13) or by solving equations (9) and (11):

$$ET_0 = n$$

$$ET_1 = n + \frac{n-1}{M}$$

(Hudson 1990; Notohara 1990; Hey 1991) and

$$\mathrm{Var}(T_0) = n^2 + 2\,\frac{(n-1)^2}{M}$$

$$\mathrm{Var}(T_1) = n^2 + 2\,\frac{(n-1)^2}{M} + \frac{(n-1)^2}{M^2}$$

(Hey 1991). In Herbots (1997) we have also calculated the probability density functions of T_0 and T_1.

Recalling the notation f_0 for the probability of identity of a pair of genes from a single subpopulation and denoting by f_1 the probability of identity of two genes from different subpopulations, f_0 and f_1 are given by results (12) and (13), respectively, with $s = \theta$. Two genes chosen uniformly at random from the total population are from the same subpopulation with probability $1/n$ and from different subpopulations with probability $1 - 1/n$. Hence the probability of identity of a pair of genes sampled at random from the total population is

$$\begin{aligned}\bar{f} &= \frac{1}{n} f_0 + \left(1 - \frac{1}{n}\right) f_1 \\ &= \frac{nM + (n-1)\,\theta}{n\{M + (nM + n - 1)\,\theta + (n-1)\,\theta^2\}}\,. \end{aligned} \tag{14}$$

Substituting (12) and (14) into the definition of F_{ST}, equation (1), we find the (global) value of F_{ST}:

$$F_{ST} = \frac{1}{1 + Mn^2/(n-1)^2 + \theta n/(n-1)} \tag{15}$$

which was obtained earlier by Nei (1975) and Takahata (1983) by means of classical techniques. The value of Slatkin's approximation

$F_{ST}^{(0)}$ is found by letting $\theta \downarrow 0$ in result (15), or from equation (8):

$$F_{ST}^{(0)} = \frac{1}{1 + Mn^2/(n-1)^2}$$

(Slatkin 1991 and references therein). The derivative of F_{ST} with respect to θ, in the limit as $\theta \downarrow 0$, is calculated directly from result (15):

$$\lim_{\theta \downarrow 0} \frac{\partial}{\partial \theta} F_{ST} = -\frac{n/(n-1)}{\{1 + Mn^2/(n-1)^2\}^2} .$$

This result can also be obtained from equation (10).

For the sake of completeness, we also calculate the exact and approximate pairwise F_{ST} values. Whilst the population is made up of n subpopulations, the "collection of subpopulations considered" in the definition of F_{ST} (equation (1)) is in that case a pair of subpopulations. Two genes sampled uniformly at random from the union of these two subpopulations are from the same subpopulation with probability 1/2. Hence

$$\bar{f} = \frac{f_0 + f_1}{2} .$$

Denoting the F_{ST} value of a pair of subpopulations by $F_{ST}(1)$, equation (1) gives

$$F_{ST}(1) = \frac{1}{1 + 2Mn/(n-1) + 2\theta} .$$

The pairwise value of Slatkin's approximation is

$$F_{ST}^{(0)}(1) = \frac{1}{1 + 2Mn/(n-1)} .$$

The derivative of the pairwise value of F_{ST} with respect to the scaled mutation rate, in the limit of zero mutation rate, is

$$\lim_{\theta \downarrow 0} \frac{\partial}{\partial \theta} F_{ST}(1) = -\frac{2}{\{1 + 2Mn/(n-1)\}^2} .$$

4.2 The circular stepping-stone model

In this model the population is divided into n equal-sized subpopulations (where n is finite) which are located on a circle, and migration

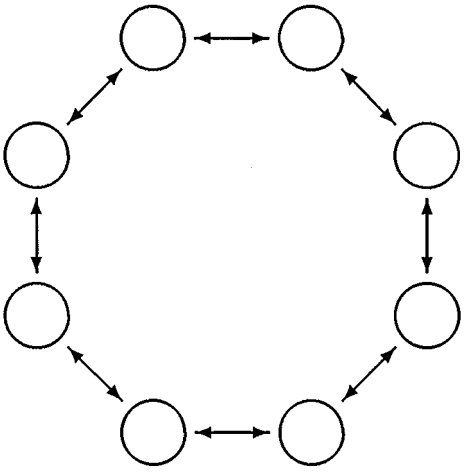

Fig. 2. Circular stepping-stone model with $n = 8$ subpopulations

occurs only between adjacent subpopulations. We assume that with the appropriate time-scale, the genealogy of a sample from the population is well described by the structured coalescent with Q-matrix given by equation (3), where $\mathscr{S} = \{1, \ldots, n\}$ and where for $i, j = 1, \ldots, n$ with $j \neq i$:

$$c_i = 1$$

$$M_{ij} = \begin{cases} aM & \text{if } j = (i-1) \bmod n \\ (1-a)\,M & \text{if } j = (i+1) \bmod n \\ 0 & \text{otherwise}, \end{cases}$$

where a is a constant such that $0 \leqq a \leqq 1$. Numbering the subpopulations in clockwise direction and working backward in time, each gene has a scaled migration rate of $aM/2$ in anti-clockwise direction and of $(1-a)M/2$ in clockwise direction. This model could resemble colonies around a lake or a mountain, or along the edge of a forest or the shore of an island.

The distribution of the coalescence time of a pair of genes under this model depends on their locations only through their distance, d, defined as the number of subpopulations separating the two genes; d ranges from zero to $[n/2]$, the largest integer not larger than $n/2$. Denoting by T_d the coalescence time of two genes at distance d,

equations (7) reduce to recursive equations in d:

$$(1 + M + s)E[e^{-sT_0}] - ME[e^{-sT_1}] = 1, \tag{16}$$

$$(M + s)E[e^{-sT_d}] - \frac{M}{2}E[e^{-sT_{d-1}}] - \frac{M}{2}E[e^{-sT_{d+1}}] = 0 \tag{17}$$

$$\text{for } d = 1, \ldots, [\tfrac{n}{2}] - 1,$$

$$(M + s)E[e^{-sT_{n/2}}] - ME[e^{-sT_{n/2-1}}] = 0 \quad \text{if } n \text{ is even}, \tag{18}$$

$$\left(\frac{M}{2} + s\right)E[e^{-sT_{(n-1)/2}}] - \frac{M}{2}E[e^{-sT_{(n-1)/2-1}}] = 0 \quad \text{if } n \text{ is odd}. \tag{19}$$

(Note that these equations are independent of the parameter a.) The coefficients in this system of recursive equations are constant in d. The general solution of equation (17) is

$$A_+ \cdot \left(\frac{\lambda_+(s)}{M}\right)^d + A_- \cdot \left(\frac{\lambda_-(s)}{M}\right)^d, \tag{20}$$

where $\lambda_+(s)/M$ and $\lambda_-(s)/M$ are the solutions of the characteristic equation

$$\frac{M}{2}\lambda^2 - (M + s)\lambda + \frac{M}{2} = 0,$$

that is,

$$\begin{aligned} \lambda_+(s) &= M + s + \sqrt{(2M + s)s} \\ \lambda_-(s) &= M + s - \sqrt{(2M + s)s}. \end{aligned} \tag{21}$$

The constants A_+ and A_- are found from the boundary conditions, equations (16) and (18) or (19). Denoting

$$\begin{aligned} &a_+(s) = a_-(s) = 1 \quad \text{if } n \text{ is even} \\ &\left.\begin{aligned} a_+(s) &= \sqrt{(2M + s)s} + s \\ a_-(s) &= \sqrt{(2M + s)s} - s \end{aligned}\right\} \quad \text{if } n \text{ is odd}, \end{aligned} \tag{22}$$

we obtain the following solution for the system of equations (16) to (19): for $d = 0, \ldots, [n/2]$,

$$E[e^{-sT_d}] = \frac{M^d\{a_+(s)\lambda_+(s)^{[n/2]-d} + a_-(s)\lambda_-(s)^{[n/2]-d}\}}{(1 + \sqrt{(2M + s)s})\, a_+(s)\lambda_+(s)^{[n/2]} + (1 - \sqrt{(2M + s)s})\, a_-(s)\lambda_-(s)^{[n/2]}}. \tag{23}$$

Maruyama (1970) performed an eigenvector analysis to solve a similar system of recursive equations in the case of a circular stepping-stone model with finite population size. The limit of Maruyama's result (with some minor corrections: $[(n+1)/2]$ in Maruyama's formulae for f_i and S should be $[n/2]$; in his formula for S the term for $k=0$ should also be included in the summation) as the subpopulation size tends to infinity gives an alternative expression for the Laplace transform of the distribution of T_d:

$$E[\mathrm{e}^{-sT_d}] = \frac{\frac{1}{n}\sum_{k=0}^{n-1} \frac{\cos\frac{2\pi kd}{n}}{s+M(1-\cos\frac{2\pi k}{n})}}{1+\frac{1}{n}\sum_{k=0}^{n-1} \frac{1}{s+M(1-\cos\frac{2\pi k}{n})}} \tag{24}$$

for $d=0,\ldots,[n/2]$. (It is easily verified that (24) indeed solves equations (16) to (19). As the solution of this system of recursive equations is unique, results (23) and (24) are equal.) Malécot (1975) found a (slightly simpler) approximation of (23), valid when $ns \ll M$.

By differentiating result (23) with respect to s, or by solving equations (9) and (11), we find the mean and the variance of the coalescence time of a pair of genes d steps apart:

$$ET_d = n + \frac{d(n-d)}{M} \tag{25}$$

$$\mathrm{Var}(T_d) = n^2 + \frac{n(n^2-1)}{3M} + \frac{d(n-d)\{n^2+1-2d(n-d)\}}{3M^2} \tag{26}$$

for $d=0,\ldots,[n/2]$. The mean coalescence time of two genes at distance d, equation (25), was calculated earlier by Slatkin (1991) using standard results on random walks. In fact, all results obtained for this model could also have been found using results on first passage times of one-dimensional random walks.

Calculating F_{ST} values for a stepping-stone model, sampling is often restricted to pairs of subpopulations at distance d, i.e. the collection of subpopulations considered is a pair of subpopulations d steps apart. The probability of identity of a pair of genes sampled at random from the union of these two subpopulations is

$$\bar{f} = \frac{f_0 + f_d}{2}, \tag{27}$$

where f_d is the probability of identity of two genes at distance d. Substituting this into the definition of F_{ST}, equation (1), the F_{ST} value

of a pair of subpopulations at distance d is

$$F_{ST}(d) = \frac{f_0 - f_d}{2 - (f_0 + f_d)} . \tag{28}$$

The value of f_d is result (23) with s replaced by θ. Substituting this result into (28), we find for $d = 1, \ldots, [n/2]$:

$$F_{ST}(d) = \frac{a_+(\theta)\lambda_+(\theta)^{[n/2]-d}\{\lambda_+(\theta)^d - M^d\} + a_-(\theta)\,\lambda_-(\theta)^{[n/2]-d}\{\lambda_-(\theta)^d - M^d\}}{a_+(\theta)\lambda_+(\theta)^{[n/2]-d}\{b_+(\theta)\,\lambda_+(\theta)^d - M^d\} + a_-(\theta)\,\lambda_-(\theta)^{[n/2]-d}\{b_-(\theta)\lambda_-(\theta)^d - M^d\}} \tag{29}$$

where $a_+(\cdot)$, $a_-(\cdot)$, $\lambda_+(\cdot)$ and $\lambda_-(\cdot)$ are defined by equations (22) and (21) and where in addition

$$b_+(\theta) = 1 + 2\sqrt{(2M + \theta)\,\theta}$$

$$\mathrm{b}_-(\theta) = 1 - 2\sqrt{(2M + \theta)\,\theta} .$$

Alternatively, substitution of (24) into (28) gives the following equivalent expression for the F_{ST} value of a pair of subpopulations at distance d:

$$F_{ST}(d) = \frac{\frac{1}{n}\sum_{k=0}^{n-1} \frac{1 - \cos\frac{2\pi kd}{n}}{\theta + M(1 - \cos\frac{2\pi k}{n})}}{2 + \frac{1}{n}\sum_{k=0}^{n-1} \frac{1 - \cos\frac{2\pi kd}{n}}{\theta + M(1 - \cos\frac{2\pi k}{n})}}$$

for $d = 1, \ldots, [n/2]$. Taking averages as in (27), we have for Slatkin's approximation $F_{ST}^{(0)}$, calculated for a pair of subpopulations d steps apart:

$$F_{ST}^{(0)}(d) = \frac{ET_d - ET_0}{ET_d + ET_0} \tag{30}$$

(Slatkin 1991), while the derivative of $F_{ST}(d)$ with respect to θ, in the limit as $\theta \downarrow 0$, is given by

$$\lim_{\theta \downarrow 0} \frac{\partial}{\partial \theta} F_{ST}(d) = -\frac{E[T_d^2]ET_0 - E[T_0^2]ET_d}{(ET_d + ET_0)^2} . \tag{31}$$

Substituting results (25) and (26) into equations (30) and (31), we obtain for $d = 1, \ldots, [n/2]$:

$$F_{ST}^{(0)}(d) = \frac{1}{1 + 2Mn/[d(n-d)]}$$

(Slatkin 1991) and

$$\lim_{\theta \downarrow 0} \frac{\partial}{\partial \theta} F_{ST}(d) = -\frac{n}{3} \cdot \frac{1 + 2/[d(n-d)]}{\{1 + 2Mn/[d(n-d)]\}^2} .$$

We also calculate the global value of F_{ST}, in which $\bar{f}$ involves sampling from the total population rather than from a pair of subpopulations d steps apart. The probability of identity of a pair of genes sampled uniformly at random from the total population is

$$\bar{f} = \begin{cases} \frac{1}{n} f_0 + \frac{2}{n} \sum_{d=1}^{n/2-1} f_d + \frac{1}{n} f_{n/2} & \text{if } n \text{ is even} \\ \frac{1}{n} f_0 + \frac{2}{n} \sum_{d=1}^{(n-1)/2} f_d & \text{if } n \text{ is odd} . \end{cases} \tag{32}$$

From expression (24) for the Laplace transform of the distribution of T_d, and using that

$$\sum_{d=0}^{n-1} \cos \frac{2\pi k d}{n} = \begin{cases} 0 & \text{for } k = 1, \ldots, n-1 \\ n & \text{for } k = 0 , \end{cases}$$

we find:

$$\bar{f} = \frac{1}{\theta \{ n + \sum_{k=0}^{n-1} \frac{1}{\theta + M(1 - \cos 2\pi k/n)} \}} \tag{33}$$

for n even or odd. This result is also the limit of infinite subpopulation size of a result in Maruyama (1970). The global value of F_{ST} follows by substituting (33) and (24) with $d = 0$ and $s = \theta$ into equation (1):

$$F_{ST} = \frac{\sum_{k=1}^{n-1} \frac{1}{\theta + M(1 - \cos \frac{2\pi k}{n})}}{n + \sum_{k=1}^{n-1} \frac{1}{\theta + M(1 - \cos \frac{2\pi k}{n})}} . \tag{34}$$

From the equality of expressions (23) and (24) with $d = 0$ we have the following identity:

$$\frac{1}{n} \sum_{k=0}^{n-1} \frac{1}{\theta + M(1 - \cos \frac{2\pi k}{n})}$$

$$= \frac{1}{\sqrt{(2M + \theta)\theta}} \cdot \frac{a_+(\theta)\, \lambda_+(\theta)^{[n/2]} + a_-(\theta) \lambda_-(\theta)^{[n/2]}}{a_+(\theta)\, \lambda_+(\theta)^{[n/2]} - a_-(\theta) \lambda_-(\theta)^{[n/2]}}$$

where $a_+(\cdot)$, $a_-(\cdot)$, $\lambda_+(\cdot)$ and $\lambda_-(\cdot)$ are given by equations (22) and (21). Substituting this identity into result (34), we find the following explicit expression for the global value of F_{ST}:

$$\begin{aligned} F_{ST} = [&(n\theta - \sqrt{(2M+\theta)\,\theta})\, a_+(\theta)\, \lambda_+(\theta)^{[n/2]} \\ &+ (n\theta + \sqrt{(2M+\theta)\,\theta})\, a_-(\theta) \lambda_-(\theta)^{[n/2]}]/ \\ &[(n\theta - (1-n\theta)\sqrt{(2M+\theta)\,\theta})\, a_+(\theta)\, \lambda_+(\theta)^{[n/2]} \\ &+ (n\theta + (1-n\theta)\sqrt{(2M+\theta)\,\theta})\, a_-(\theta)\, \lambda_-(\theta)^{[n/2]}] \,. \end{aligned} \tag{35}$$

The global values of Slatkin's approximation for F_{ST} and of $\lim_{\theta \downarrow 0} \partial F_{ST}/\partial\theta$ are found from equations (8) and (10), where ET and $E[T^2]$ are calculated analogously to equation (32). Alternatively they can be obtained directly from (35). The results are:

$$F_{ST}^{(0)} = \frac{n^2 - 1}{n^2 - 1 + 6nM}$$

$$\lim_{\theta \downarrow 0} \frac{\partial}{\partial\theta} F_{ST} = -\frac{n}{5} \cdot \frac{n^4 + 10n^2 - 11}{(n^2 - 1 + 6nM)^2} \,.$$

4.3 The infinite linear stepping-stone model

Here the population consists of an infinite line of equal-sized subpopulations. Migration is possible only to neighbouring subpopulations. We assume the structured coalescent with Q-matrix given by (3) where $\mathscr{S} = \mathbb{Z}$ and where for every $i, j \in \mathscr{S}$ with $j \neq i$,

$$c_i = 1$$

$$M_{ij} = \begin{cases} aM & \text{if } j = i - 1 \\ (1-a)\, M & \text{if } j = (i+1) \\ 0 & \text{otherwise} \end{cases}$$

where a is a constant such that $0 \leqq a \leqq 1$. Working backwards in time, every gene has a scaled migration rate of $aM/2$ in negative direction

Fig. 3. The infinite linear stepping-stone model

and of $(1-a)M/2$ in positive direction. This model might be appropriate for a very long array of colonies along a shore or a river-bed.

We denote again by T_d the coalescence time of a pair of genes at distance d. Equations (7) reduce to equations (16) and (17) in the previous subsection, where (17) now holds for every $d \in \mathbb{N}\backslash\{0\}$. The general solution of equation (17) is (20). Equation (16) provides one boundary condition. As a second boundary condition, we know that as $d \to \infty$, $E[e^{-sT_d}]$ remains bounded by 1. Because $\lambda_+(s)/M > 1$ for $s > 0$, this implies that $A_+ = 0$. The value of A_- is found subsequently by substituting $E[e^{-sT_d}] = A_-(\lambda_-(s)/M)^d$ into equation (16). The resulting expression for the Laplace transform of the distribution of T_d is

$$E[e^{-sT_d}] = \frac{(M + s - \sqrt{(2M + s)s})^d}{M^d(1 + \sqrt{(2M + s)s})} \tag{36}$$

for $d = 0, 1, 2, \ldots$. Malécot (1948; English translation: 1969) and Maruyama (1970) earlier obtained approximations of this result, valid when both $s \ll M$ and $ds \ll M$. We note that (36) is the limit of the corresponding result for the circular stepping-stone model, (23), as the number of subpopulations on the circle becomes infinitely large. Taking the limit of (24) as $n \to \infty$ we obtain an alternative expression for (36):

$$E[e^{-sT_d}] = \frac{\int_{-\pi}^{\pi} \frac{\cos(xd)}{M + s - M\cos x}\, dx}{2\pi + \int_{-\pi}^{\pi} \frac{1}{M + s - M\cos x}\, dx} \tag{37}$$

for $d = 0, 1, 2, \ldots$. This is also the limit of infinite subpopulation size of a result Maruyama (1970) gave as an approximation valid for a circular stepping-stone model with many subpopulations. The above results for the Laplace transform of the distribution of the coalescence time of a pair of genes d steps apart in the infinite linear stepping-stone model could also have been found by noting that the distance between the two genes, until it first becomes zero, performs a one-dimensional symmetric random walk and by using results on first passage times (see for example Feller 1966, Sect. XIV.6). Differentiating (36) with respect to s and taking the limit as $s \downarrow 0$, we find that the mean, and hence also the second moment, of the coalescence time of any pair of genes are infinite:

$$ET_d = \infty$$

$$E[T_d^2] = \infty$$

for $d = 0, 1, 2, \ldots$ (see also Griffiths 1981). This is effectively a consequence of the null-recurrence of the symmetric random walk in one

dimension. Notohara (1990) derived a system of differential equations for the cumulative distribution function of the coalescence time of a pair of genes from specified locations. He also studied the probability of coalescence in a particular subpopulation, given the initial locations of the two genes.

The F_{ST} value of a pair of subpopulations at distance d, denoted by $F_{ST}(d)$, is calculated by substituting results (36) or (37), with $s = \theta$, into equation (28):

$$F_{ST}(d) = \frac{M^d - (M + \theta - \sqrt{(2M + \theta)\,\theta})^d}{M^d - (M + \theta - \sqrt{(2M + \theta)\,\theta})^d + 2M^d \cdot \sqrt{(2M + \theta)\,\theta}} \tag{38}$$

$$= \frac{\int_{-\pi}^{\pi} \frac{1 - \cos(xd)}{M + \theta - M\cos x}\, dx}{4\pi + \int_{-\pi}^{\pi} \frac{1 - \cos(xd)}{M + \theta - M\cos x}\, dx}$$

for $d = 1, 2, \ldots$. The value of its approximation $F_{ST}^{(0)}(d)$ is found taking the limit of (38) as $\theta \downarrow 0$, using l'Hôpital's rule:

$$F_{ST}^{(0)}(d) = \frac{1}{1 + 2M/d} \tag{39}$$

for $d = 1, 2, \ldots$. (This value cannot be found from equation (30), since the mean coalescence time of any two genes under this model is infinite.) Slatkin (1991) obtained expression (39) as an approximation valid for a circular stepping-stone model with a large number of subpopulations. From result (38) we also find (using l'Hôpital's rule) that

$$\lim_{\theta \downarrow 0} \frac{\partial}{\partial \theta} F_{ST}(d) = -\infty \tag{40}$$

for $d = 1, 2, \ldots$. This means that under the infinite linear stepping-stone model, $F_{ST}(d)$ depends very strongly on the mutation rate, when this is small.

4.4 Two-dimensional stepping-stone model on a torus

A two-dimensional array of equal-sized subpopulations is situated on a torus, and migration is possible only from each subpopulation to its four neighbouring subpopulations. Denoting by K the number of subpopulations in one dimension and by L that in the other ($K, L \geqq 2$), the total number of subpopulations is $n = K \times L$. The assumption that

corresponding subpopulations on opposite sides of the two-dimensional array are linked by migration, although standard (Maruyama 1970; Malécot 1975; Crow and Aoki 1984; Slatkin 1991, 1993), may be unrealistic. Without this assumption however the model is intractible. This model may still provide a good approximation for a large two-dimensional lattice of colonies. One interesting feature of this finite two-dimensional stepping-stone model is that it makes it possible to investigate the dependence of F_{ST} on habitat shape (see Crow and Aoki 1984 and Herbots 1994).

Denoting by $M_{(1)}$ and $M_{(2)}$ the scaled migration rates in the respective dimensions of the torus ($M_{(1)} > 0$, $M_{(2)} > 0$) and by $M = M_{(1)} + M_{(2)}$ the total scaled migration rate from each subpopulation, we assume that the genealogy of a sample from the population is well described by the structured coalescent with Q-matrix (3), where $\mathscr{S} = \{0, \ldots, K-1\} \times \{0, \ldots, L-1\}$ and where for every (i, j), $(k, l) \in \mathscr{S}$ with $(i, j) \neq (k, l)$:

$$c_{(i,j)} = 1$$

and

$$M_{(i,j)(k,l)} = \begin{cases} aM_{(1)} & \text{if } l = j \text{ and } k = (i-1) \bmod K \\ (1-a)M_{(1)} & \text{if } l = j \text{ and } k = (i+1) \bmod K \\ bM_{(2)} & \text{if } k = i \text{ and } l = (j-1) \bmod L \\ (1-b)M_{(2)} & \text{if } k = i \text{ and } l = (j+1) \bmod L \\ 0 & \text{otherwise} \end{cases}$$

where a and b are constants such that $0 \leqq a \leqq 1$ and $0 \leqq b \leqq 1$. Every gene has scaled rate $M/2$ of leaving its subpopulation; it moves in the first dimension of the torus at scaled rate $M_{(1)}/2$ and in the second dimension of the torus at scaled rate $M_{(2)}/2$.

The distribution of the coalescence time of a pair of genes from this population is a function of the numbers of subpopulations, d_1 and d_2, separating the two genes in the respective dimensions of the torus ($d_1 = 0, \ldots, [K/2]$; $d_2 = 0, \ldots, [L/2]$). We denote by $T_{(d_1,d_2)}$ the coalescence time of a pair of genes at "distance" (d_1, d_2). Equations (7) for the Laplace transform of the distribution of the coalescence time of a pair of genes reduce to a system of recursive equations in the *two* recursion indices d_1 and d_2. Rather than solving these equations directly, we found it easier to obtain results using the theory of random walks. Note however that these recursive equations are independent of the parameters a and b.

The coalescence time of two genes d_1 steps apart in the first dimension of the torus and d_2 steps apart in the other dimension is the

time $T^{(r)}_{(d_1, d_2)}$ until (working backward in time) the ancestors of the two genes are present in a single subpopulation for the first time, plus the coalescence time of two genes in a single subpopulation:

$$T_{(d_1, d_2)} \stackrel{\mathrm{d}}{=} T^{(r)}_{(d_1, d_2)} + T_{(0,0)} \,.$$

Because of the Markov character of the structured coalescent, the two times on the right-hand side are independent, so that

$$E[\mathrm{e}^{-sT_{(d_1,d_2)}}] = E[\mathrm{e}^{-sT^{(r)}_{(d_1,d_2)}}]\, E[\mathrm{e}^{-sT_{(0,0)}}] \,. \tag{41}$$

For $d_1 = d_2 = 0$: $E[\mathrm{e}^{-sT^{(r)}_{(d_1,d_2)}}] = 1$. Assume $(d_1, d_2) \neq (0, 0)$. Until the ancestral lineages of the two genes are present in a single subpopulation for the first time, their "distance" (seen as a bivariate process with one component for each dimension) performs a symmetric[3] random walk on the rectangular lattice $\{0, \ldots, [K/2]\} \times \{0, \ldots, [L/2]\}$. The distribution of $T^{(r)}_{(d_1, d_2)}$ is that of the first passage time through (0, 0) of this random walk, starting at $(d_1, d_2) \neq (0, 0)$. To calculate this distribution, it is easier to label the genes' ancestral lineages as lineages one and two, and to measure the distance between them "clockwise" from lineage one, in both dimensions of the torus. By "clockwise" we mean the direction corresponding to a transition from location 0 to location 1 in that particular dimension of the torus. The distance between the two lineages (measured clockwise from lineage one) thus performs a symmetric random walk on the "$K \times L$ torus" $\{0, \ldots, K-1\} \times \{0, \ldots, L-1\}$, where transitions occur at rate M (at rates $M_{(1)}$ and $M_{(2)}$ in the respective dimensions of the torus). The distribution of $T^{(r)}_{(d_1, d_2)}$ is the same as that of the first passage time through (0, 0) of the latter random walk, starting at $(d_1, d_2) \neq (0, 0)$. We denote by $(D_1, D_2) \equiv \{(D_1, D_2)_v : v = 0, 1, 2, \ldots\}$ the jump chain of the random walk described by the distance between the two genes' lineages, measured clockwise from lineage one, and by $U^{(r)}_{(d_1, d_2)}$ the first passage time of (D_1, D_2) through (0, 0), starting from $(d_1, d_2) \neq (0, 0)$. The relationship between $T^{(r)}_{(d_1, d_2)}$ and $U^{(r)}_{(d_1, d_2)}$ is given by

$$T^{(r)}_{(d_1, d_2)} \stackrel{\mathrm{d}}{=} \sum_{i=1}^{U^{(r)}_{(d_1,d_2)}} X_i$$

where the X_i are mutually independent, exponentially distributed random variables with mean $1/M$, which are also independent of

[3] By a "symmetric" random walk in two dimensions we mean a random walk which is symmetric in each dimension of its statespace (but with possibly different transition rates in the different dimensions)

$U^{(r)}_{(d_1,d_2)}$. Denoting by $F_{(d_1,d_2)\,(0,0)}$ the probability generating function of $U^{(r)}_{(d_1,d_2)}$, that is,

$$F_{(d_1,d_2)\,(0,0)}(z) := E[z^{U^{(r)}_{(d_1,d_2)}}],$$

it follows that

$$\begin{aligned} E[\mathrm{e}^{-sT^{(r)}_{(d_1,d_2)}}] &= E\left[\prod_{i=1}^{U^{(r)}_{(d_1,d_2)}} \mathrm{e}^{-sX_i}\right] \\ &= EE\left[\prod_{i=1}^{U^{(r)}_{(d_1,d_2)}} \mathrm{e}^{-sX_i} \,\middle|\, U^{(r)}_{(d_1,d_2)}\right] \\ &= E\left[\prod_{i=1}^{U^{(r)}_{(d_1,d_2)}} E[\mathrm{e}^{-sX_i}]\right] \\ &= E\left[\left(\frac{M}{M+s}\right)^{U^{(r)}_{(d_1,d_2)}}\right] \\ &= F_{(d_1,d_2)\,(0,0)}\left(\frac{M}{M+s}\right). \end{aligned} \tag{42}$$

In order to find $F_{(d_1,d_2)\,(0,0)}$, we calculate the probability

$$P^{(v)}_{(d_1,d_2)\,(0,0)} := P\{(D_1, D_2)_v = (0,0) \,|\, (D_1, D_2)_0 = (d_1, d_2)\} \tag{43}$$

that the discrete-time random walk (D_1, D_2), starting at (d_1, d_2), is at the origin immediately after the vth transition. Note that K steps in the same direction (clockwise or anti-clockwise) in the first dimension of the torus, or L steps in a single direction in the second dimension of the torus, do not alter the position of (D_1, D_2). Assume that among the first v steps of (D_1, D_2), x steps were in the first dimension of the torus, while $v - x$ steps were in the second dimension. Starting at d_1, D_1 takes the value 0 at step x if among these x steps, there were $d_1 + kK$ more steps anti-clockwise than clockwise, for some $k \in \mathbb{Z}$ (where $-K$ steps anti-clockwise are to be interpreted as $+K$ steps clockwise). Thus

$$\begin{aligned} P^{(v)}_{(d_1,d_2)\,(0,0)} = \frac{1}{2^v} \sum_{x=0}^{v} \binom{v}{x} \left(\frac{M_{(1)}}{M}\right)^x \left(\frac{M_{(2)}}{M}\right)^{v-x} \\ \times \sum_{k=-\infty}^{+\infty} \binom{x}{(x + d_1 + kK)/2} \sum_{l=-\infty}^{+\infty} \binom{v-x}{(v - x + d_2 + lL)/2} \end{aligned} \tag{44}$$

where for $A \in \mathbb{N}$: $\binom{A}{B} \equiv 0$ if $B \notin \{0, 1, \ldots, A\}$. The following lemma follows from a slight extension of Theorem 4.3 in Teugels (1986):

Lemma. *For all* $A, B \in \mathbb{N}$ and $C \in \mathbb{N}_0$:

$$\sum_{k=-\infty}^{+\infty} \binom{A}{(A+B+kC)/2} = \frac{1}{C} \sum_{v=0}^{C-1} \left(2 \cos \frac{2v\pi}{C}\right)^A \cos \frac{2Bv\pi}{C} .$$

Using this lemma, equation (44) gives

$$\begin{aligned} P^{(v)}_{(d_1,d_2)\,(0,0)} &= \frac{1}{2^v KL} \sum_{x=0}^{v} \binom{v}{x} \left(\frac{M_{(1)}}{M}\right)^x \left(\frac{M_{(2)}}{M}\right)^{v-x} \sum_{v=0}^{K-1} \left(2 \cos \frac{2v\pi}{K}\right)^x \\ &\quad \times \cos \frac{2d_1 v\pi}{K} \sum_{w=0}^{L-1} \left(2 \cos \frac{2w\pi}{L}\right)^{v-x} \cos \frac{2d_2 w\pi}{L} \\ &= \frac{1}{(2M)^v KL} \sum_{v=0}^{K-1} \cos \frac{2d_1 v\pi}{K} \sum_{w=0}^{L-1} \cos \frac{2d_2 w\pi}{L} \sum_{x=0}^{v} \binom{v}{x} \\ &\quad \times \left(2M_{(1)} \cos \frac{2v\pi}{K}\right)^x \left(2M_{(2)} \cos \frac{2w\pi}{L}\right)^{v-x} \\ &= \frac{1}{M^v KL} \sum_{v=0}^{K-1} \sum_{w=0}^{L-1} \cos \frac{2d_1 v\pi}{K} \cos \frac{2d_2 w\pi}{L} \\ &\quad \times \left(M_{(1)} \cos \frac{2v\pi}{K} + M_{(2)} \cos \frac{2w\pi}{L}\right)^v . \end{aligned}$$

Introducing the generating function

$$P_{(d_1,d_2)\,(0,0)}(z) := \sum_{v=0}^{\infty} P^{(v)}_{(d_1,d_2)\,(0,0)}\, z^v ,$$

we have for $0 \leqq z < 1$:

$$\begin{aligned} P_{(d_1,d_2)\,(0,0)}(z) &= \frac{1}{KL} \sum_{v=0}^{K-1} \sum_{w=0}^{L-1} \cos \frac{2d_1 v\pi}{K} \cos \frac{2d_2 w\pi}{L} \\ &\quad \times \sum_{v=0}^{\infty} \left(M_{(1)} \cos \frac{2v\pi}{K} + M_{(2)} \cos \frac{2w\pi}{L}\right)^v \left(\frac{z}{M}\right)^v \\ &= \frac{1}{KL} \sum_{v=0}^{K-1} \sum_{w=0}^{L-1} \frac{\cos \frac{2d_1 v\pi}{K} \cos \frac{2d_2 w\pi}{L}}{1 - z\,(M_{(1)} \cos \frac{2v\pi}{K} + M_{(2)} \cos \frac{2w\pi}{L})/M} . \end{aligned} \quad (45)$$

The probability generating function of $U^{(r)}_{(d_1,d_2)}$ is, for $(d_1, d_2) \neq (0, 0)$, given by

$$F_{(d_1,d_2)\,(0,0)}(z) = \frac{P_{(d_1,d_2)\,(0,0)}(z)}{P_{(0,0)\,(0,0)}(z)}$$

(see for example equation (5.3) in Chapter XV of Feller 1968). Combining this with (42) and (45), we obtain:

$$E[\mathrm{e}^{-sT^{(r)}_{(d_1,d_2)}}] = \frac{P_{(d_1,d_2)\,(0,0)}\left(\frac{M}{M+s}\right)}{P_{(0,0)\,(0,0)}\left(\frac{M}{M+s}\right)} \tag{46}$$

$$= \frac{\sum_{v=0}^{K-1}\sum_{w=0}^{L-1}\frac{\cos(2d_1v\pi/K)\cos(2d_2w\pi/L)}{M+s-[M_{(1)}\cos(2v\pi/K)+M_{(2)}\cos(2w\pi/L)]}}{\sum_{v=0}^{K-1}\sum_{w=0}^{L-1}\frac{1}{M+s-[M_{(1)}\cos(2v\pi/K)+M_{(2)}\cos(2w\pi/L)]}} \tag{47}$$

which is valid for $d_1 = 0, \ldots, [K/2]$ and $d_2 = 0, \ldots, [L/2]$, and for $s > 0$.

For the Laplace transform of the distribution of the coalescence time of two genes from a single subpopulation, (7) gives the following equation:

$$(1 + M + s)\,E[\mathrm{e}^{-sT_{(0,0)}}] - M_{(1)}\,E[\mathrm{e}^{-sT_{(1,0)}}] - M_{(2)}\,E[\mathrm{e}^{-sT_{(0,1)}}] = 1\,. \tag{48}$$

Substituting (41) into this equation, it follows that

$$E[\mathrm{e}^{-sT_{(0,0)}}] = \frac{1}{1 + M + s - M_{(1)}\,E[\mathrm{e}^{-sT^{(r)}_{(1,0)}}] - M_{(2)}\,E[\mathrm{e}^{-sT^{(r)}_{(0,1)}}]}\,. \tag{49}$$

Substituting this and (47) into equation (41), we find for $d_1 = 0, \ldots, [K/2]$, $d_2 = 0, \ldots, [L/2]$ and $s > 0$:

$$E[\mathrm{e}^{-sT_{(d_1,d_2)}}] = \frac{E[\mathrm{e}^{-sT^{(r)}_{(d_1,d_2)}}]}{1 + M + s - M_{(1)}\,E[\mathrm{e}^{-sT^{(r)}_{(1,0)}}] - M_{(2)}\,E[\mathrm{e}^{-sT^{(r)}_{(0,1)}}]} \tag{50}$$

$$= \frac{\sum_{v=0}^{K-1}\sum_{w=0}^{L-1}\frac{\cos(2d_1v\pi/K)\cos(2d_2w\pi/L)}{M+s-[M_{(1)}\cos(2v\pi/K)+M_{(2)}\cos(2w\pi/L)]}}{KL + \sum_{v=0}^{K-1}\sum_{w=0}^{L-1}\frac{1}{M+s-[M_{(1)}\cos(2v\pi/K)+M_{(2)}\cos(2w\pi/L)]}}\,. \tag{51}$$

For a stepping-stone model on a square torus ($K = L$) with finite population size, Maruyama (1970) earlier calculated the probability of identity of a pair of genes, without the use of coalescent techniques or the theory of random walks. Apart from a minor error in Maruyama's result ($[(n+1)/2]$ there should be $[n/2]$), Maruyama's value, in the limit of infinite subpopulation size, agrees with our result.

The mean and the variance of the coalescence time of two genes at distance (d_1, d_2) are calculated by differentiation of equation (51). Denoting $\mathscr{S}_0 := \mathscr{S}\backslash\{(0,0)\}$, the results are, for $d_1 = 0, \ldots, [K/2]$ and $d_2 = 0, \ldots, [L/2]$:

$$ET_{(d_1,d_2)}$$

$$= KL + \sum_{(v,w)\in\mathscr{S}_0}\frac{1 - \cos(2v\pi d_1/K)\cos(2w\pi d_2/L)}{M - [M_{(1)}\cos(2v\pi/K) + M_{(2)}\cos(2w\pi/L)]} \tag{52}$$

and

$$\begin{aligned}
&\mathrm{Var}(T_{(d_1,d_2)}) \\
&= (KL)^2 + 2KL \sum_{(v,w)\in\mathscr{S}_0} \frac{1}{M - [M_{(1)}\cos(2v\pi/K) + M_{(2)}\cos(2w\pi/L)]} \\
&\quad + \left(\sum_{(v,w)\in\mathscr{S}_0} \frac{1}{M - [M_{(1)}\cos(2v\pi/K) + M_{(2)}\cos(2w\pi/L)]} \right)^2 \\
&\quad - \left(\sum_{(v,w)\in\mathscr{S}_0} \frac{\cos(2v\pi d_1/K)\cos(2w\pi d_2/L)}{M - [M_{(1)}\cos(2v\pi/K) + M_{(2)}\cos(2w\pi/L)]} \right)^2 \\
&\quad + 2 \sum_{(v,w)\in\mathscr{S}_0} \frac{1 - \cos(2v\pi d_1/K)\cos(2w\pi d_2/L)}{\{M - [M_{(1)}\cos(2v\pi/K) + M_{(2)}\cos(2w\pi/L)]\}^2} .
\end{aligned}$$

In the special case of $M_{(1)} = M_{(2)}$, Slatkin (1991, 1993) obtained for the mean coalescence time (up to minor typographical errors) the same value as (52) by solving a matrix equation (using the technique of Maruyama 1970).

The exact F_{ST} value, $F_{ST}(d_1, d_2)$, of a pair of subpopulations at "distance" (d_1, d_2) on a $K \times L$ torus is calculated from result (51) according to equation (28) with 0 replaced by (0, 0) and d by (d_1, d_2):

$$F_{ST}(d_1, d_2) = \frac{\sum_{v=0}^{K-1}\sum_{w=0}^{L-1} \frac{1 - \cos(2v\pi d_1/K)\cos(2w\pi d_2/L)}{M + \theta - [M_{(1)}\cos(2v\pi/K) + M_{(2)}\cos(2w\pi/L)]}}{2KL + \sum_{v=0}^{K-1}\sum_{w=0}^{L-1} \frac{1 - \cos(2v\pi d_1/K)\cos(2w\pi d_2/L)}{M + \theta - [M_{(1)}\cos(2v\pi/K) + M_{(2)}\cos(2w\pi/L)]}} \tag{53}$$

for $d_1 = 0, \ldots, [K/2]$ and $d_2 = 0, \ldots, [L/2]$ with $(d_1, d_2) \neq (0, 0)$. The values of Slatkin's approximation $F_{ST}^{(0)}(d_1, d_2)$ and of $\lim_{\theta \downarrow 0} \partial F_{ST}(d_1, d_2)/\partial\theta$ are found analogously according to equations (30) and (31), or are found directly from result (53). We obtain for $d_1 = 0, \ldots, [K/2]$ and $d_2 = 0, \ldots, [L/2]$ with $(d_1, d_2) \neq (0, 0)$:

$$F_{ST}^{(0)}(d_1, d_2) = \frac{\sum_{(v,w)\in\mathscr{S}_0} \frac{1 - \cos(2v\pi d_1/K)\cos(2w\pi d_2/L)}{M - [M_{(1)}\cos(2v\pi/K) + M_{(2)}\cos(2w\pi/L)]}}{2KL + \sum_{(v,w)\in\mathscr{S}_0} \frac{1 - \cos(2v\pi d_1/K)\cos(2w\pi d_2/L)}{M - [M_{(1)}\cos(2v\pi/K) + M_{(2)}\cos(2w\pi/L)]}}$$

(see also Slatkin 1991, 1993 in the special case of $M_{(1)} = M_{(2)}$) and

$$\begin{aligned}
&\lim_{\theta\downarrow 0} \frac{\partial}{\partial\theta} F_{ST}(d_1, d_2) \\
&= -\frac{2KL \sum_{(v,w)\in\mathscr{S}_0} \frac{1 - \cos(2v\pi d_1/K)\cos(2w\pi d_2/L)}{\{M - [M_{(1)}\cos(2v\pi/K) + M_{(2)}\cos(2w\pi/L)]\}^2}}{\{2KL + \sum_{(v,w)\in\mathscr{S}_0} \frac{1 - \cos(2v\pi d_1/K)\cos(2w\pi d_2/L)}{M - [M_{(1)}\cos(2v\pi/K) + M_{(2)}\cos(2w\pi/L)]}\}^2} .
\end{aligned}$$

In order to find the global value of F_{ST} under this model, we need to calculate the probability of identity of a pair of genes drawn at random from the total population. From result (51) we obtain:

$$\bar{f} = \frac{\frac{1}{KL}\sum_{v=0}^{K-1}\sum_{w=0}^{L-1}\frac{\sum_{d_1=0}^{K-1}\cos(2d_1v\pi/K)\ \sum_{d_2=0}^{L-1}\cos(2d_2w\pi/L)}{M+\theta-[M_{(1)}\cos(2v\pi/K)+M_{(2)}\cos(2w\pi/L)]}}{KL+\sum_{v=0}^{K-1}\sum_{w=0}^{L-1}\frac{1}{M+\theta-[M_{(1)}\cos(2v\pi/K)+M_{(2)}\cos(2w\pi/L)]}}$$

$$= \frac{1}{\theta\left\{KL+\sum_{v=0}^{K-1}\sum_{w=0}^{L-1}\frac{1}{M+\theta-[M_{(1)}\cos(2v\pi/K)+M_{(2)}\cos(2w\pi/L)]}\right\}}.$$

Substituting this and result (51) with $d_1 = d_2 = 0$ and $s = \theta$ into the definition of F_{ST}, equation (1), we find the global value of F_{ST}:

$$F_{ST} = \frac{\sum_{(v,w)\in\mathscr{S}_0}\frac{1}{M+\theta-[M_{(1)}\cos(2v\pi/K)+M_{(2)}\cos(2w\pi/L)]}}{KL+\sum_{(v,w)\in\mathscr{S}_0}\frac{1}{M+\theta-[M_{(1)}\cos(2v\pi/K)+M_{(2)}\cos(2w\pi/L)]}}.$$

The limit of this result as $\theta\downarrow 0$ is the global value of Slatkin's approximation for F_{ST}:

$$F_{ST}^{(0)} = \frac{\sum_{(v,w)\in\mathscr{S}_0}\frac{1}{M-[M_{(1)}\cos(2v\pi/K)+M_{(2)}\cos(2w\pi/L)]}}{KL+\sum_{(v,w)\in\mathscr{S}_0}\frac{1}{M-[M_{(1)}\cos(2v\pi/K)+M_{(2)}\cos(2w\pi/L)]}}.$$

We also find that

$$\lim_{\theta\downarrow 0}\frac{\partial}{\partial\theta}F_{ST} = -\frac{KL\sum_{(v,w)\in\mathscr{S}_0}\frac{1}{\{M-[M_{(1)}\cos(2v\pi/K)+M_{(2)}\cos(2w\pi/L)]\}^2}}{\left\{KL+\sum_{(v,w)\in\mathscr{S}_0}\frac{1}{M-[M_{(1)}\cos(2v\pi/K)+M_{(2)}\cos(2w\pi/L)]}\right\}^2}.$$

4.5 The infinite two-dimensional stepping-stone model

The population consists of a two-dimensional rectangular lattice of equal-sized subpopulations. The number of subpopulations in each dimension is infinite. From each subpopulation, genes can migrate only to the four neighbouring subpopulations.

We assume the structured coalescent with Q-matrix given by (3) with $\mathscr{S} = \mathbb{Z}\times\mathbb{Z}$ and where for every $(i,j),(k,l)\in\mathscr{S}$ with $(i,j)\neq(k,l)$:

$$c_{(i,j)} = 1$$

and

$$M_{(i,j)(k,l)} = \begin{cases} aM_{(1)} & \text{if } l=j \text{ and } k=i-1 \\ (1-a)M_{(1)} & \text{if } l=j \text{ and } k=i+1 \\ bM_{(2)} & \text{if } k=i \text{ and } l=j-1 \\ (1-b)M_{(2)} & \text{if } k=i \text{ and } l=j+1 \\ 0 & \text{otherwise ,} \end{cases}$$

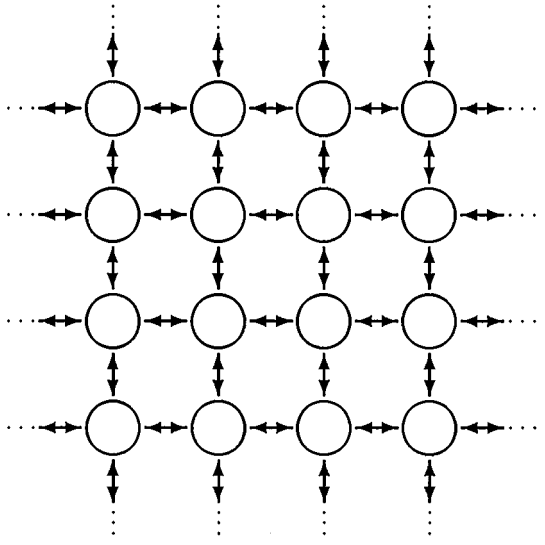

Fig. 4. The infinite two-dimensional stepping-stone model

where $M_{(1)}$ and $M_{(2)}$ are the scaled migration rates in the respective dimensions of the lattice ($M_{(1)} > 0$, $M_{(2)} > 0$ and $M = M_{(1)} + M_{(2)}$) and $a, b \in [0, 1]$ are constants.

We denote by $T_{(d_1, d_2)}$ the coalescence time of a pair of genes d_1 subpopulations apart in the first dimension of the lattice and d_2 subpopulations apart in the second dimension ($d_1, d_2 \in \mathbb{N}$). Rather than solving equations (7), we show that the Laplace transform of the distribution of $T_{(d_1, d_2)}$ under this model is the limit of result (51) for the stepping-stone model on the torus as, in both dimensions of the torus, the number of subpopulations tends to infinity. Denoting by $T^{(r)}_{(d_1, d_2)}$ the time until the ancestral lineages of a pair of genes initially at "distance" (d_1, d_2) in the infinite two-dimensional stepping-stone model are present in a single subpopulation for the first time, we have again that

$$T_{(d_1, d_2)} \stackrel{\mathrm{d}}{=} T^{(r)}_{(d_1, d_2)} + T_{(0,0)} ,$$

with $T^{(r)}_{(d_1, d_2)}$ and $T_{(0,0)}$ independent. Until the two genes' lineages are present in a single subpopulation for the first time, their distance (seen as a bivariate process with a component for each dimension of the lattice) performs a symmetric random walk on $\mathbb{N} \times \mathbb{N}$, where transitions occur at rates $M_{(1)}$ and $M_{(2)}$ in the respective dimensions of $\mathbb{N} \times \mathbb{N}$, and at rate M in total. For $(d_1, d_2) \neq (0, 0)$, the distribution of $T^{(r)}_{(d_1, d_2)}$ is that of the first passage time through $(0, 0)$ of this random

walk, starting at (d_1, d_2). We denote by $(D_1, D_2) \equiv \{(D_1, D_2)_v : v = 0, 1, 2, \ldots\}$ the jump chain of this random walk ((D_1, D_2) is a discrete-time symmetric random walk on $\mathbb{N} \times \mathbb{N}$) and by $P^{(v)}_{(d_1, d_2)\,(0,0)}(\infty, \infty)$ the probability that (D_1, D_2), starting from (d_1, d_2), is at the origin immediately after the vth step:

$$P^{(v)}_{(d_1, d_2)\,(0,0)}(\infty, \infty) := P\{(D_1, D_2)_v = (0,0) \,|\, (D_1, D_2)_0 = (d_1, d_2)\} .$$

The corresponding probability for the $K \times L$ torus, introduced in equation (43), is now denoted by $P^{(v)}_{(d_1, d_2)\,(0,0)}(K, L)$. For fixed $d_1, d_2, v \in \mathbb{N}$ we have that for $K > d_1 + v$ and $L > d_2 + v$:

$$P^{(v)}_{(d_1, d_2)\,(0,0)}(K, L) = P^{(v)}_{(d_1, d_2)\,(0,0)}(\infty, \infty) ,$$

because for such K and L the random walk cannot go round the torus in $d_1 + v$ or $d_2 + v$ steps in its respective dimensions. Hence

$$\lim_{K \to \infty} \lim_{L \to \infty} P^{(v)}_{(d_1, d_2)\,(0,0)}(K, L) = P^{(v)}_{(d_1, d_2)\,(0,0)}(\infty, \infty)$$

for every $d_1, d_2, v \in \mathbb{N}$. By the dominated convergence theorem we also have convergence of the generating functions, for $0 \leqq z < 1$:

$$\lim_{K \to \infty} \lim_{L \to \infty} \sum_{v=0}^{\infty} P^{(v)}_{(d_1, d_2)\,(0,0)}(K, L)\, z^v = \sum_{v=0}^{\infty} P^{(v)}_{(d_1, d_2)\,(0,0)}(\infty, \infty)\, z^v \quad (54)$$

for $d_1, d_2 \in \mathbb{N}$. By a similar argument as in the previous subsection, equation (46) still holds (where the generating functions $P_{(d_1, d_2)\,(0,0)}$ and $P_{(0,0)\,(0,0)}$ now refer to the discrete-time random walk on $\mathbb{N} \times \mathbb{N}$ described above). Since equation (7) gives for the Laplace transform of the distribution of $T_{(0,0)}$ under the infinite two-dimensional stepping-stone model the same equation as (48), equation (50) is still valid as well. Thus it follows from result (54) that the Laplace transform of the distribution of the coalescence time of a pair of genes at distance (d_1, d_2) in the infinite two-dimensional stepping-stone model is obtained by letting $K, L \to \infty$ in result (51) for the stepping-stone model on the $K \times L$ torus. So we find for the infinite two-dimensional stepping-stone model:

$$E[e^{-sT_{(d_1, d_2)}}] = \frac{\int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \frac{\cos(d_1 x)\cos(d_2 y)}{M + s - (M_{(1)} \cos x + M_{(2)} \cos y)}\, dx\, dy}{(2\pi)^2 + \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \frac{1}{M + s - (M_{(1)} \cos x + M_{(2)} \cos y)}\, dx\, dy} \quad (55)$$

for $d_1, d_2 \in \mathbb{N}$ and $s > 0$. This result is also the limit of infinite subpopulation size of an expression Maruyama (1970) suggested as an approximation for the probability of identity of a pair of genes under

a stepping-stone model on a large square torus. The double integrals in result (55) can be reduced to single integrals, which are easier to evaluate by computer using numerical integration. To do so, the results for the infinite linear stepping-stone model prove very helpful. From the equality of expressions (36) and (37) for the Laplace transform of the distribution of the coalescence time of a pair of genes d steps apart under the infinite linear stepping-stone model, we find the following identity (valid for every $d \in \mathbb{N}$ and for $s > 0$):

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\cos(xd)}{M+s-M\cos x}\,dx = \frac{(M+s-\sqrt{(2M+s)s})^d}{M^d\sqrt{(2M+s)s}}\,. \quad (56)$$

Using this identity with $M_{(1)}+s-M_{(1)}\cos x$ instead of s, and with $M_{(2)}$ instead of M, we can re-write the double integral in the numerator of (55) as

$$\begin{aligned}
&\frac{1}{(2\pi)^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\frac{\cos(d_1x)\cos(d_2y)}{M+s-(M_{(1)}\cos x+M_{(2)}\cos y)}\,dx\,dy\\
&\quad=\frac{1}{2\pi}\int_{-\pi}^{\pi}\cos(d_1x)\\
&\qquad\times\left[\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\cos(d_2y)}{M_{(2)}+(M_{(1)}+s-M_{(1)}\cos x)-M_{(2)}\cos y}\,dy\right]dx\\
&\quad=\frac{1}{\pi}\int_{0}^{\pi}\cos(d_1x)\\
&\qquad\times\frac{(M+s-M_{(1)}\cos x-\sqrt{(M+s-M_{(1)}\cos x)^2-M_{(2)}^2})^{d_2}}{M_{(2)}^{d_2}\sqrt{(M+s-M_{(1)}\cos x)^2-M_{(2)}^2}}\,dx
\end{aligned}$$

so that

$$E[e^{-sT_{(d_1,d_2)}}]=\frac{\int_0^{\pi}\cos(d_1x)\,\frac{(M+s-M_{(1)}\cos x-\sqrt{(M+s-M_{(1)}\cos x)^2-M_{(2)}^2})^{d_2}}{M_{(2)}^{d_2}\sqrt{(M+s-M_{(1)}\cos x)^2-M_{(2)}^2}}\,dx}{\pi+\int_0^{\pi}\frac{1}{\sqrt{(M+s-M_{(1)}\cos x)^2-M_{(2)}^2}}\,dx} \quad (57)$$

for $d_1, d_2 \in \mathbb{N}$ and $s > 0$. Malécot (1950, 1975) obtained a similar expression for the probability of identity of a pair of genes under a more general two-dimensional migration model. It can be shown that for the stepping-stone model described here, Malécot's result gives an approximation of result (57) (with little actual simplification) by focussing on the value of the integrands in the neighbourhood of $x = 0$, valid when both s/M_2 and $d_2\,s/M_2$ are negligible relative to 1.

Because the unbounded two-dimensional symmetric random walk is null-recurrent, the mean and the second moment of the coalescence time of any pair of genes are infinite:

$$ET_{(d_1,d_2)} = \infty \tag{58}$$

$$E[T^2_{(d_1,d_2)}] = \infty \tag{59}$$

for every $d_1, d_2 \in \mathbb{N}$.

The exact F_{ST} value of a pair of subpopulations d_1 and d_2 steps apart in the respective dimensions of the lattice, denoted by $F_{ST}(d_1, d_2)$, is calculated from results (55) or (57) according to equation (28) with d replaced by (d_1, d_2) and 0 by (0, 0):

$$F_{ST}(d_1,d_2) = \frac{\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} \frac{1-\cos(d_1x)\cos(d_2y)}{M+\theta-(M_{(1)}\cos x + M_{(2)}\cos y)}\,dx\,dy}{8\pi^2 + \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} \frac{1-\cos(d_1x)\cos(d_2y)}{M+\theta-(M_{(1)}\cos x + M_{(2)}\cos y)}\,dx\,dy} \tag{60}$$

$$= \frac{\int_0^{\pi} \frac{M_{(2)}^{d_2} - \cos(d_1x)(M+\theta-M_{(1)}\cos x - \sqrt{(M+\theta-M_{(1)}\cos x)^2 - M_{(2)}^2})^{d_2}}{M_{(2)}^{d_2}\sqrt{(M+\theta-M_{(1)}\cos x)^2 - M_{(2)}^2}}\,dx}{2\pi + \int_0^{\pi} \frac{M_{(2)}^{d_2} - \cos(d_1x)(M+\theta-M_{(1)}\cos x - \sqrt{(M+\theta-M_{(1)}\cos x)^2 - M_{(2)}^2})^{d_2}}{M_{(2)}^{d_2}\sqrt{(M+\theta-M_{(1)}\cos x)^2 - M_{(2)}^2}}\,dx} \tag{61}$$

for $d_1, d_2 \in \mathbb{N}$ with $(d_1, d_2) \neq (0, 0)$.

Because the mean and the second moment of the coalescence time of any two genes are infinite, the value of the approximation $F_{ST}^{(0)}(d_1, d_2)$ and that of $\lim_{\theta\downarrow 0} \partial F_{ST}(d_1, d_2)/\partial\theta$ cannot be obtained from equations (30) and (31), but have to be calculated directly from $F_{ST}(d_1, d_2)$. The integrand of the double integral in result (60) is non-negative and increases as $\theta > 0$ decreases. The same is true for the integrand of the single integral in result (61), as

$$\frac{M_{(2)}^{d_2} - \cos(d_1x)(M+\theta-M_{(1)}\cos x - \sqrt{(M+\theta-M_{(1)}\cos x)^2 - M_{(2)}^2})^{d_2}}{M_{(2)}^{d_2}\sqrt{(M+\theta-M_{(1)}\cos x)^2 - M_{(2)}^2}}$$

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{1-\cos(d_1x)\cos(d_2y)}{M+\theta-(M_{(1)}\cos x + M_{(2)}\cos y)}\,dy \tag{62}$$

(see (56) with M replaced by $M_{(2)}$ and with $s = M_{(1)} + \theta - M_{(1)}\cos x$), for $\theta > 0$ and $x \in \mathbb{R}$. Hence letting $\theta\downarrow 0$ in results (60) and (61) it follows by the monotone convergence theorem that for $d_1, d_2 \in \mathbb{N}$ with $(d_1, d_2) \neq (0, 0)$:

$$F_{ST}^{(0)}(d_1,d_2) = \frac{\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} \frac{1-\cos(d_1x)\cos(d_2y)}{M-(M_{(1)}\cos x + M_{(2)}\cos y)}\,dx\,dy}{8\pi^2 + \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} \frac{1-\cos(d_1x)\cos(d_2y)}{M-(M_{(1)}\cos x + M_{(2)}\cos y)}\,dx\,dy} \tag{63}$$

$$= \frac{\int_0^{\pi} \frac{M_{(2)}^{d_2} - \cos(d_1x)(M-M_{(1)}\cos x - \sqrt{(M-M_{(1)}\cos x)^2 - M_{(2)}^2})^{d_2}}{M_{(2)}^{d_2}\sqrt{(M-M_{(1)}\cos x)^2 - M_{(2)}^2}}\,dx}{2\pi + \int_0^{\pi} \frac{M_{(2)}^{d_2} - \cos(d_1x)(M-M_{(1)}\cos x - \sqrt{(M-M_{(1)}\cos x)^2 - M_{(2)}^2})^{d_2}}{M_{(2)}^{d_2}\sqrt{(M-M_{(1)}\cos x)^2 - M_{(2)}^2}}\,dx} \tag{64}$$

where we note that although the integrand in result (64) has a singularity at $x = 0$, its limit as $x \downarrow 0$ is finite, so that the integral in result (64) is finite (which implies that also the double integral in result (63) is finite). In the special case of $M_{(1)} = M_{(2)}$ and $d_2 = 0$, Slatkin (1991) gave an expression similar to (64) as an approximation valid for a stepping-stone model on a large square torus.

We now calculate the value of

$$\lim_{\theta \downarrow 0} \frac{\partial}{\partial \theta} F_{ST}(d_1, d_2)$$

$$= \frac{2\pi \lim_{\theta \downarrow 0} \int_0^\pi \frac{\partial}{\partial \theta} \frac{M_{(2)}^{d_2} - \cos(d_1 x)(M + \theta - M_{(1)} \cos x - \sqrt{(M + \theta - M_{(1)} \cos x)^2 - M_{(2)}^2})^{d_2}}{M_{(2)}^{d_2} \sqrt{M + \theta M_{(1)} \cos x)^2 - M_{(2)}^2}} dx}{\left\{ 2\pi + \int_0^\pi \frac{M_{(2)}^{d_2} - \cos(d_1 x)(M - M_{(1)} \cos x - \sqrt{(M - M_{(1)} \cos x)^2 - M_{(2)}^2})^{d_2}}{M_{(2)}^{d_2} \sqrt{(M - M_{(1)} \cos x)^2 - M_{(2)}^2}} dx \right\}^2} \tag{65}$$

where we have applied Leibniz's rule. The integral in the denominator of (65) is the same as that in (64) and is finite. Using (62) and Leibniz's rule, we have for $\theta > 0$:

$$-\frac{\partial}{\partial \theta} \frac{M_{(2)}^{d_2} - \cos(d_1 x)(M + \theta - M_{(1)} \cos x - \sqrt{(M + \theta - M_{(1)} \cos x)^2 - M_{(2)}^2})^{d_2}}{M_{(2)}^{d_2} \sqrt{(M + \theta - M_{(1)} \cos x)^2 - M_{(2)}^2}}$$

$$= \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{1 - \cos(d_1 x) \cos(d_2 y)}{\{M + \theta - (M_{(1)} \cos x + M_{(2)} \cos y)\}^2} dy$$

which is non-negative and monotonically increasing with decreasing θ. By the monotone convergence theorem,

$$\lim_{\theta \downarrow 0} \int_0^\pi \frac{\partial}{\partial \theta} \{[M_{(2)}^{d_2} - \cos(d_1 x)(M + \theta - M_{(1)} \cos x$$

$$- \sqrt{(M + \theta - M_{(1)} \cos x)^2 - M_{(2)}^2})^{d_2}]/$$

$$[M_{(2)}^{d_2} \sqrt{(M + \theta - M_{(1)} \cos x)^2 - M_{(2)}^2}]\} dx$$

$$= \int_0^\pi \lim_{\theta \downarrow 0} \frac{\partial}{\partial \theta} \{[M_{(2)}^{d_2} - \cos(d_1 x)(M + \theta - M_{(1)} \cos x$$

$$- \sqrt{(M + \theta - M_{(1)} \cos x)^2 - M_{(2)}^2})^{d_2}]/$$

$$[M_{(2)}^{d_2} \sqrt{(M + \theta - M_{(1)} \cos x)^2 - M_{(2)}^2}]\} dx .$$

Once the derivative and the limit in the latter integral are calculated explicitly, a Taylor expansion about $x = 0$ of all the cosines in the resulting integrand (which is continuous on $(0, \pi]$ but has a singularity at $x = 0$) shows that

$$\lim_{x \downarrow 0} \frac{-\lim_{\theta \downarrow 0} \frac{\partial}{\partial \theta} \frac{M_{(2)}^{d_2} - \cos(d_1 x)(M + \theta - M_{(1)} \cos x - \sqrt{(M + \theta - M_{(1)} \cos x)^2 - M_{(2)}^2})^{d_2}}{M_{(2)}^{d_2} \sqrt{(M + \theta - M_{(1)} \cos x)^2 - M_{(2)}^2}}}{1/x}$$

$$= \frac{d_1^2 M_{(2)} + d_2^2 M_{(1)}}{2(M_{(1)} M_{(2)})^{3/2}}.$$

Hence, because

$$\int_0^\pi \frac{1}{x} dx = +\infty$$

it follows by the quotient test for improper integrals that for $d_1, d_2 \in \mathbb{N}$ with $(d_1, d_2) \neq (0, 0)$:

$$\int_0^\pi \lim_{\theta \downarrow 0} \frac{\partial}{\partial \theta} \{[M_{(2)}^{d_2} - \cos(d_1 x)(M + \theta - M_{(1)} \cos x - \sqrt{(M + \theta - M_{(1)} \cos x)^2 - M_{(2)}^2})^{d_2}] / [M_{(2)}^{d_2} \sqrt{(M + \theta - M_{(1)} \cos x)^2 - M_{(2)}^2}]\} dx = -\infty.$$

Thus for $d_1, d_2 \in \mathbb{N}$ with $(d_1, d_2) \neq (0, 0)$:

$$\lim_{\theta \downarrow 0} \frac{\partial}{\partial \theta} F_{ST}(d_1, d_2) = -\infty, \tag{66}$$

as was also the case for the infinite one-dimensional stepping-stone model.

5 Some less symmetric structures

The island model and the stepping-stone models studied in the previous section are the models traditionally dealt with in the literature. The structure of natural populations is, of course, far less symmetric. In particular, most real populations have subpopulations of unequal sizes. However, most published theoretical studies of genetic differentiation in subdivided populations have been restricted to models in which all subpopulations are identical with respect to size and migration pattern (exceptions include studies by Nagylaki (1988), Nagylaki and Barcilon (1988), and Nagylaki et al. (1993) of the effect of spatial

inhomogeneities and geographical barriers on a one-dimensional stepping-stone model). In this section two specific models of population structure are considered which allow for unequal subpopulation sizes and/or different migration patterns from different subpopulations. In the first subsection we consider a two-population model (previously studied also by Takahata (1988) and Notohara (1990)), in which the population consists of two subpopulations of possibly different sizes. In Subsect. 2 we introduce the "continental island model", where one subpopulation (the "continent") has a migration pattern different from that of the other subpopulations (the "islands"). Although still unrealistic in most cases, these models may be used to get some idea of what effect such asymmetries in the structure of the population may have on the genealogy of and the level of subpopulation differentiation in the population (see Herbots 1994).

In this section we will be concerned with the *global* (and not with the pairwise) values of F_{ST} and its approximation $F_{ST}^{(0)}$. In Sect. 3, F_{ST} was expressed in terms of coalescence times of pairs of genes sampled from the population. When a gene is sampled from the population and subpopulation sizes are unequal, there seem two distinct, natural, sampling schemes. If the population subdivision is geographic and genes are sampled at specified locations, it is known which subpopulation each gene is taken from, but the relative sizes of the different subpopulations are, in most cases, unknown. In those cases, *equal weighting* of the subpopulations may be appropriate, that is, to choose a gene at random from the population, we first choose one of the subpopulations, with each choice being equally likely, regardless of its size, and then choose a gene uniformly at random from the chosen subpopulation. Assuming that the population consists of n subpopulations (where n is finite) and labelling the subpopulations as $1, 2, \ldots, n$, the probabilities of identity, f_0 and $\bar{f}$, of, respectively, two genes sampled from a single subpopulation and two genes sampled from the total population are in this context given by

$$f_0 = \frac{1}{n} \sum_{i=1}^{n} f_{ii} \quad \text{and} \quad \bar{f} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} f_{ij} \tag{67}$$

where f_{ij} is the probability of identity of a gene from subpopulation i and a gene from subpopulation j. In other settings (perhaps for example human populations), one may suspect population subdivision, but it may be difficult to outline the different subpopulations and one may have a random sample from the population, not knowing which individuals belong to which subpopulations. In such cases, it may be realistic to assume that all individuals in the population are equally likely to be sampled, so that the various subpopulations are naturally

weighted by their relative sizes:

$$f_0 = \sum_{i=1}^{n} P_i f_{ii} \quad \text{and} \quad \bar{f} = \sum_{i=1}^{n} \sum_{j=1}^{n} P_i P_j f_{ij} \tag{68}$$

where P_i is the proportion of the total population that belongs to subpopulation i.

As before it is assumed throughout this section that, with the appropriate scaling of time, the genealogy of a sample from the population is well described by the structured coalescent, given by equation (3). We restrict to the case where migration is "conservative" (Nagylaki 1980) in the sense that

$$\forall i \in \mathscr{S}: \sum_{j \neq i} c_i M_{ij} = \sum_{j \neq i} c_j M_{ji} \,. \tag{69}$$

(Note that this assumption was also fulfilled by all the models considered in Sect. 4.) As in the previous sections, θ denotes the scaled mutation rate (every gene mutates at scaled rate $\theta/2$). We also denote $c := \sum_{i=1}^{n} c_i$ and $U := c\theta$. As an example, consider the following model in discrete time: the population is divided into a finite number of subpopulations, each subpopulation i contains $2c_i N$ genes (i.e. the c_i correspond to the relative subpopulation sizes), reproduction in each subpopulation follows the neutral Wright–Fisher model and a constant proportion q_{ij} of the individuals born in subpopulation i migrate to subpopulation j every generation ($i, j \in \mathscr{S}$), where

$$\forall i \in \mathscr{S}: c_i \sum_{j \neq i} q_{ij} = \sum_{j \neq i} c_j q_{ji} \,. \tag{70}$$

The latter assumption means that the size of each subpopulation is maintained under migration. Measuring time in units of $2N$ generations, the genealogy of a sample from this population is well described by the structured coalescent with Q-matrix (3), where for all $i, j \in \mathscr{S}$ with $j \neq i$: $M_{ij} = \lim_{N \to \infty}(4N c_j q_{ji}/c_i)$ (see Herbots 1994, 1997). Assumption (70) implies that equation (69) holds, so that migration is conservative. Denoting by μ the probability of mutation per gene per generation, we have that $\theta = \lim_{N \to \infty}(4N\mu)$ under this model, so that (in the coalescent approximation) U is twice the number of mutations expected in the total population per generation.

5.1 The two-population model

The population is divided into two subpopulations ($n = 2$) of possibly different sizes, and there is (conservative) migration between the two subpopulations.

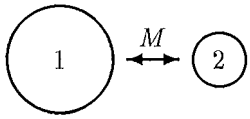

Fig. 5. An example of the two-population model

We assume the structured coalescent with $\mathscr{S} = \{1, 2\}$. The assumption of conservative migration, (69), means that $c_1 M_1 = c_2 M_2$, where we recall that M_i is the scaled migration rate from subpopulation i (working backward in time, each gene in subpopulation i has scaled rate $M_i/2$ of migrating to the other subpopulation). For $c_1 = c_2$, the two-population model is, up to a factor of time-scaling, the island model with $n = 2$ subpopulations.

Denoting by T_{ij} the coalescence time of a gene in subpopulation i and a gene in subpopulation j, (7) gives the following system of linear equations for the Laplace transform of the distribution of T_{ij}:

$$\begin{cases} \left(\frac{1}{c_1} + M_1 + s\right) E[\mathrm{e}^{-sT_{11}}] - M_1 E[\mathrm{e}^{-sT_{12}}] = \frac{1}{c_1} \\ \left(\frac{M_1}{2} + \frac{M_2}{2} + s\right) E[\mathrm{e}^{-sT_{12}}] - \frac{M_2}{2} E[\mathrm{e}^{-sT_{11}}] - \frac{M_1}{2} E[\mathrm{e}^{-sT_{22}}] = 0 \\ \left(\frac{1}{c_2} + M_2 + s\right) E[\mathrm{e}^{-sT_{22}}] - M_2 E[\mathrm{e}^{-sT_{12}}] = \frac{1}{c_2} \ . \end{cases}$$

We recall the notation $c = c_1 + c_2$ and we write $P := c_1/c$. In many cases, the c_i correspond to the relative subpopulation sizes so that P is the proportion of the total population that lives in subpopulation 1. We also introduce the "migration rate" $M := c_1 M_1 = c_2 M_2$. As an example, consider again the discrete-time model described just before the beginning of this subsection, in the case of two subpopulations. For that model, $2cN$ is the total number of genes in the population, P is the proportion of genes in subpopulation 1 and M is (in the coalescent approximation) twice the number of genes exchanged between the two subpopulations every generation. We will express all results for the two-population model in terms of the parameters P, c, the "migration rate" M and the "mutation rate" $U = c\theta$. The solution of the above system of linear equations is then given by

$$E[\mathrm{e}^{-sT_{11}}] = \frac{\{M + 2P(1-P)cs\}\{1 + M + (1-P)cs\}}{A} \qquad (71)$$

$$E[e^{-sT_{12}}] = \frac{M\{1 + M + 2P(1-P)cs\}}{A} \tag{72}$$

$$E[e^{-sT_{22}}] = \frac{\{M + 2P(1-P)cs\}\{1 + M + Pcs\}}{A} \tag{73}$$

where

$$A = P(1-P)\{Mcs(4 + 3cs) + 2cs(1 + cs) + 2P(1-P)(cs)^3\} + M(1 + M)(1 + cs) .$$

These results are equivalent to results in Takahata (1988). The mean and the variance of the coalescence time of a pair of genes from specified locations are obtained either by differentiating these results with respect to s or by solving equations (9) and (11). We find:

$$ET_{11} = c - \frac{(1-P)(1-2P)c}{1 + M} \tag{74}$$

$$ET_{12} = c + \frac{2P(1-P)c}{M} \tag{75}$$

$$ET_{22} = c + \frac{P(1-2P)c}{1 + M} \tag{76}$$

(Takahata 1988; Notohara 1990) and

$$\mathrm{Var}(T_{11}) = c^2 + (1-P)c^2\left(\frac{2P}{1 + M} + \frac{8(1-P)P^2}{M(1 + M)} - \frac{(1-P)(1-4P^2)}{(1 + M)^2}\right)$$

$$\mathrm{Var}(T_{12}) = c^2 + 2P(1-P)c^2\left(\frac{1}{1 + M} + \frac{4P(1-P)}{M(1 + M)} + \frac{2P(1-P)}{M^2}\right)$$

$$\mathrm{Var}(T_{22}) = c^2 + Pc^2\left(\frac{2(1-P)}{1 + M} + \frac{8P(1-P)^2}{M(1 + M)} - \frac{P(1-4(1-P)^2)}{(1 + M)^2}\right).$$

Whereas under the symmetric models considered in the previous section, the mean coalescence time of two genes from a single subpopulation is independent of the migration rate, ET_{11} and ET_{22} under the two-population model do depend on M, unless $P = 1/2$.

The value of F_{ST} under *equal weighting* of the subpopulations is obtained from equation (1) where f_0 and $\bar{f}$ are calculated from results (71) to (73) according to equations (67). We find:

$$F_{ST} = \frac{M + 2P(1-P)(2+U)}{M(3+4M) + P(1-P)\{4 + 6U + 8P(1-P)U^2 + 8M + 12MU\}} . \tag{77}$$

Letting $U \downarrow 0$ in this result, we find the value of Slatkin's approximation for F_{ST}:

$$F_{ST}^{(0)} = \frac{4P(1-P) + M}{4P(1-P)(1+2M) + M(3+4M)} .$$

This value could also have been obtained from equation (8), calculating ET_0 and ET from results (74) to (76) by taking averages analogous to (67). Note that under equal weighting, $ET_0 = (ET_{11} + ET_{22})/2$ depends on the migration rate (except in the case of $P = 1/2$). The value of $\lim_{\theta \downarrow 0} \partial F_{ST}/\partial\theta$ under equal weighting of the subpopulations is found directly from result (77), or alternatively, can be obtained from equation (10):

$$\lim_{\theta \downarrow 0} \frac{\partial}{\partial\theta} F_{ST} = -4P(1-P)\,c\,\frac{4P(1-P)(1+2M) + M^2}{\{4P(1-P)(1+2M) + M(3+4M)\}^2} .$$

When calculating the values of F_{ST}, $F_{ST}^{(0)}$ and $\lim_{\theta \downarrow 0} \partial F_{ST}/\partial\theta$ under *weighting by size*, all averages have to be calculated as in equations (68). Assuming that the c_i are proportional to the subpopulation sizes (so that P is the proportion of genes in subpopulation 1), the results are:

$$F_{ST} = \frac{P(1-P)\{M + 4P(1-P) + 2P(1-P)\,U\}}{P(1-P)\{M + 2U + P(1-P)(4 - 2U + 2U^2) + 3MU\} + M(1+M)}$$

$$F_{ST}^{(0)} = \frac{P(1-P)\{M + 4P(1-P)\}}{P(1-P)\{M + 4P(1-P)\} + M(1+M)}$$

$$\lim_{\theta \downarrow 0} \frac{\partial}{\partial\theta} F_{ST} = -P^2(1-P)^2\,c\,\frac{8P(1-P)\{1 - 2P(1-P) + M\} + M^2}{[P(1-P)\{M + 4P(1-P)\} + M(1+M)]^2} .$$

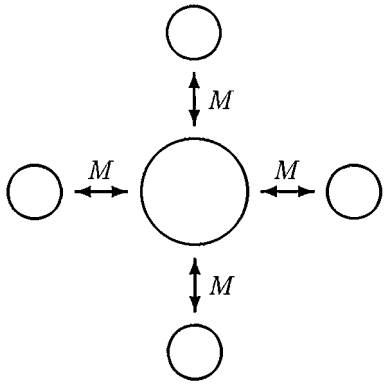

Fig. 6. Continental island model with $n = 5$

5.2 The continental island model

The population is divided into a central subpopulation (the "continent") surrounded by a number of "islands". Migration occurs only between the continent and each island. Individuals cannot directly migrate from one island to another but have to go via the continent. We assume that all the islands are identical with respect to size and with respect to migration rate to the continent.

Let n be the number of subpopulations, including the continent. Labelling the continent as subpopulation 0 and the islands as subpopulations 1 to $n-1$, we assume the structured coalescent (given by equation (3)) with $\mathscr{S} = \{0, 1, \ldots, n-1\}$ and where for $i = 1, \ldots, n-1$: $c_i = c_1$, $M_i = M_{i0} = M_1$ and $M_{0i} = M_0/(n-1)$. Working backward in time, each gene in the continent has scaled rate $M_0/2$ of leaving the continent, every island being equally likely to be its destination, while every gene in the islands has scaled rate $M_1/2$ of migrating to the continent. The assumption of conservative migration, (69), is equivalent to the following relationship between M_0 and M_1:

$$c_0 M_0 = (n-1)\, c_1 M_1 \ .$$

For $n = 2$, the continental island model reduces to the two-population model described in the previous subsection.

We denote again by T_{ij} the coalescence time of a gene from subpopulation i and a gene from subpopulation j. Because all islands are equivalent, there are essentially only four possibilities for the locations of a pair of genes from the population, so that (7) reduces to

the following system of linear equations:

$$\begin{cases} \left(\frac{1}{c_0} + M_0 + s\right) E[\mathrm{e}^{-sT_{00}}] - M_0 E[\mathrm{e}^{-sT_{01}}] = \frac{1}{c_0} \\ \left(\frac{1}{c_1} + M_1 + s\right) E[\mathrm{e}^{-sT_{11}}] - M_1 E[\mathrm{e}^{-sT_{01}}] = \frac{1}{c_1} \\ \left(\frac{M_0}{2} + \frac{M_1}{2} + s\right) E[\mathrm{e}^{-sT_{01}}] - \frac{M_1}{2} E[\mathrm{e}^{-sT_{00}}] \\ \quad - \frac{M_0}{2(n-1)} E[\mathrm{e}^{-sT_{11}}] - \frac{M_0(n-2)}{2(n-1)} E[\mathrm{e}^{-sT_{12}}] = 0 \\ (M_1 + s) E[\mathrm{e}^{-sT_{12}}] - M_1 E[\mathrm{e}^{-sT_{01}}] = 0, \end{cases} \tag{78}$$

where the last equation and the last term in the left-hand side of the third equation are present only if there are at least two islands ($n \geqq 3$). We have that $c = c_0 + (n-1)c_1$ and we denote $P := c_0/c$. If the c_i correspond to the relative subpopulation sizes, P is the proportion of genes in the continent. To facilitate the comparison of results for different relative sizes of continent and islands (see Herbots 1994), it is convenient to define the "migration rate"

$$M := c_1 M_1 = c_0 M_0/(n-1)\,.$$

For example, consider once more the discrete-time model set out just before Subsect. 5.1, where now for $i, j \in \{1, \ldots, n-1\}$ with $j \neq i$: $q_{ij} = 0$, $q_{i0} = q_{10}$ and $q_{0i} = q_{01}$. Assumption (70) means that $c_0 q_{01} = c_1 q_{10}$. As under this discrete-time model, the number of genes in subpopulation i is $2c_i N$, for every $i \in \mathscr{S}$, $2cN$ is the total number of genes in the population and P is the proportion of the total population that lives in the continent. Scaling time in units of $2N$ generations, the genealogy of a sample from this population is well described by the structured coalescent with scaled migration rates $M_{ij} = \lim_{N\to\infty}(4Nc_j q_{ji}/c_i)$ $(j \neq i)$ satisfying the restrictions described above, and M is (in the coalescent approximation) twice the number of genes exchanged between the continent and each of the islands every generation.

Expressed in terms of the parameters n, c, P and M, the solution of equations (78) gives for $i, j = 1, \ldots, n-1$ with $j \neq i$:

$$E[\mathrm{e}^{-sT_{00}}] = [B + 2(1-P^2)(n-1)^2 M^2 cs + (1+4P)(1-P)^2(n-1)M(cs)^2 + 2P(1-P)^3(cs)^3]/D \tag{79}$$

$$E[\mathrm{e}^{-sT_{ii}}] = [B + (n-1+2P-2P^2)(n-1)^2 M^2 cs + 3P(1-P)(n-1)^2 M(cs)^2 + 2(n-1)P^2(1-P)^2(cs)^3]/D \tag{80}$$

$$E[e^{-sT_{0i}}] = (n-1)\,M\{(n-1)\,M + (1-P)\,cs\} \times \{1 + (n-1)\,M + (n-2)\,P + 2P(1-P)\,cs\}/D \tag{81}$$

$$E[e^{-sT_{ij}}] = \frac{(n-1)^2 M^2 \{1 + (n-1)M + (n-2)P + 2P(1-P)\,cs\}}{D} \tag{82}$$

where

$$B = (n-1)^3 M^3 + (1-2P+nP)\,(n-1)^2 M^2 + (1-P)\,(1+2P)(n-1)^2\,Mcs + 2(n-1)\,P(1-P)^2(cs)^2\,,$$

$$D = B + \{(n-1)\,M + n + 3P - 4P^2\}\,(n-1)^2\,M^2\,cs + (1-P)\{1 + 3Pn - 4P^2 + (1+3P)(n-1)\,M\}(n-1)\,M(cs)^2 + P(1-P)^2\{2 - 4P + 2nP + (3+2P)(n-1)\,M\}(cs)^3 + 2P^2(1-P)^3(cs)^4\,.$$

For $n = 2$, results (79) to (81) reduce to the results found in the previous subsection.

The mean and the variance of the coalescence time of a pair of genes are obtained by differentiation of the above results or from equations (9) and (11). We find for $i, j = 1, \ldots, n-1$ with $j \neq i$:

$$ET_{00} = c + \frac{c(1-P)(2P+n-3)}{1 + (n-1)\,M + (n-2)\,P} \tag{83}$$

$$ET_{ii} = c - \frac{cP(2P+n-3)}{1 + (n-1)\,M + (n-2)\,P} \tag{84}$$

$$ET_{0i} = c + \frac{c(1-P)(2P+n-2)}{1 + (n-1)\,M + (n-2)\,P} + \frac{c(1-P)(nP+n-2)}{(n-1)\,M\{1 + (n-1)\,M + (n-2)\,P\}} \tag{85}$$

$$ET_{ij} = c + \frac{c(1-P)(2P+n-1)}{1 + (n-1)\,M + (n-2)P} + \frac{c(1-P)(2P+1)}{M\{1 + (n-1)\,M + (n-2)\,P\}} \tag{86}$$

and

$$\mathrm{Var}(T_{00}) =$$

$$c^2 + (1-P)\,c^2\left[\frac{2(1-P)(n-2+3Pn-6P+4P^2)}{M\{1+(n-1)M+(n-2)P\}^2}\right.$$

$$\left.+\frac{n^2+2n-9-P+Pn^2+24P^2-6P^2n-12P^3+2M(n-1)(n-2+P)}{\{1+(n-1)M+(n-2)P\}^2}\right]$$

$$\mathrm{Var}(T_{ii}) = c^2 + c^2\left[\frac{2(1-P)^2(n-2+3Pn-6P+4P^2)}{(n-1)M\{1+(n-1)M+(n-2)P\}^2}\right.$$

$$+\frac{2(1-P)^2(6P^2-3P-2)+(n-1)(6P^3-9P^2-P^2n+4)}{\{1+(n-1)M+(n-2)P\}^2}$$

$$\left.+\frac{2M(1-P)(n-1)(n-2+P)}{\{1+(n-1)M+(n-2)P\}^2}\right]$$

$$\mathrm{Var}(T_{0i}) =$$

$$c^2 + (1-P)\,c^2\left[\frac{(1-P)\{(n-1)(8P^2-12P+2Pn+2n)-(1-P)^2n^2\}}{(n-1)^2M^2\{1+(n-1)M+(n-2)P\}^2}\right.$$

$$+\frac{2(1-P)(n^2-n-2+3Pn^2-7Pn+2P+4P^2n)}{(n-1)M\{1+(n-1)M+(n-2)P\}^2}$$

$$\left.+\frac{n^2+2n-8+Pn^2-2P-6P^2n+24P^2-12P^3+2M(n-1)(n-2+P)}{\{1+(n-1)M+(n-2)P\}^2}\right]$$

$$\mathrm{Var}(T_{ij}) =$$

$$c^2 + (1-P)\,c^2\left[\frac{(1-P)(n-1+4Pn-8P+4P^2)}{(n-1)M^2\{1+(n-1)M+(n-2)P\}^2}\right.$$

$$+\frac{2(1-P)(n^2-n-1+3Pn^2-6nP+4P^2n)}{(n-1)M\{1+(n-1)M+(n-2)P\}^2}$$

$$\left.+\frac{n^2+2n-7+Pn^2-3P-6P^2n+24P^2-12P^3+2M(n-1)(n-2+P)}{\{1+(n-1)M+(n-2)P\}^2}\right].$$

The global value of F_{ST} under *equal weighting* of the subpopulations follows from equation (1), where f_0 and $\bar{f}$ are calculated from results (79) to (82) according to equations (67):

$$F_{ST} = [\omega + P(n-1)^5M^2 - 6Pn(n-1)^2MV]/$$
$$[\omega + P(n-1)^2M\lambda + n(n-1)^2(1-4P^2)MV$$
$$+ n^2(n-1)^2(1+3P)M^2V + n^2(n-1)(3+2P)MV^2$$
$$+ 2n(n-1)V^2 + 2n^2V^3] \qquad (87)$$

where

$V = P(1-P)\,U$ (recall the definition of the "mutation rate" $U = c\theta$)

$$\lambda = M(n^3 - n^2 - 1 + 2Pn + Pn^2 - 2P^2n^2) + n^2(n-1)\,M^2$$

$$\begin{aligned}\omega = {} & P(n-1)^2 M\kappa + 2P(1-P)\,n(n-1)^2 V \\ & + (n-1)^2(1 + 2P + 3Pn^2)\,MV \\ & + 2(n-1)\,(1 - 2Pn + Pn^2)\,V^2\,,\end{aligned}$$

where in the latter equation

$$\kappa = (1-P)(2Pn^2 - 2 + n^2 - n - 4Pn + 4P)\,.$$

Taking the limit of this result as θ (and hence U) decreases to zero, we find the value of Slatkin's approximation for F_{ST} under equal weighting of the subpopulations:

$$F_{ST}^{(0)} = \frac{\kappa + M(n-1)^3}{\kappa + \lambda}\,.$$

This value could also have been found from equation (8), calculating ET_0 and ET analogous to equations (67) from results (83) to (86). From result (87) we also calculate the value of $\lim_{\theta \downarrow 0} \partial F_{ST}/\partial\theta$ under equal weighting:

$$\lim_{\theta \downarrow 0} \frac{\partial}{\partial\theta} F_{ST} = -\,n(1-P)\,c\,\frac{\psi_0 + \psi_1 M + n(n-1)(n^2 - 2n + P)\,M^2}{(\kappa + \lambda)^2}$$

where κ and λ are as above and where in addition

$$\begin{aligned}\psi_0 = {} & (1-P)(n^2 - n - 2 - 4Pn - 24P^2n + 16P^3n + 4Pn^2 \\ & + 6P^2n^2 - 4P^3n^2 - 8P + 32P^2 - 16P^3)\end{aligned}$$

$$\psi_1 = (1-P)\,(2n^3 + 4Pn^3 - 4n^2 - 5Pn^2 - n - 6Pn + 8P^2n + 2)\,.$$

This result could also have been obtained from equation (10).

The global value of F_{ST} under *weighting by size* is found calculating f_0 and $\bar{f}$ from results (79) to (82) according to equations (68) and substituting the resulting values into equation (1). Assuming that c_0 and c_1 are proportional to the size of the continent and that of a single island (so that P is the proportion of the total population that lives in the continent), the result is:

$$\begin{aligned}F_{ST} = {} & [R + 2P^2(1-P)^3(n-1)\,U^2 + P(1-P)^2(3n + 2P - 5)\,WU \\ & + (n - Pn + 3P - 2)\,W^2]/[R + X_0 + X_1 W \\ & + (n - 3P^2U + 2PU + U + P - 1)\,W^2 + W^3] \qquad (88)\end{aligned}$$

where

$$
\begin{aligned}
W &= (n-1)\,M \\
R &= 2P(1-P)^3(n+Pn-2)\,U \\
&\quad + (1-P)^2(3Pn-6P+n-2+4P^2)\,W - P^2W^2 \\
X_0 &= 2P(1-P)^2\{P(1-P)\,U + P^2 + Pn - 3P + 1\}\,U^2 \\
X_1 &= (1-P)(3PU - P^2U - 2P^3U + 3Pn \\
&\quad - 3P - 3P^2 + 2P^3 + 1)\,U\ .
\end{aligned}
$$

The values of Slatkin's approximation $F_{ST}^{(0)}$ and of $\lim_{\theta\downarrow 0}\partial F_{ST}/\partial\theta$ under weighting by size are obtained either directly from result (88), or from equations (8) and (10) where averages are calculated analogously to equations (68). We find:

$$
F_{ST}^{(0)} = \frac{(1-P)^2(3Pn-6P+n-2+4P^2) + (1-P)(P+n-2)\,W}{(1-P)^2(3Pn-6P+n-2+4P^2) + P(1-P)\,W + W(W+n-1)}
$$

$$
\begin{aligned}
\lim_{\theta\downarrow 0}\frac{\partial}{\partial\theta}F_{ST} = {} & -(1-P)^2\,c\,[Z_0 + Z_1W + (n+P^2-2)\,W^2]/ \\
& \{(1-P)^2(3Pn-6P+n-2+4P^2) \\
& + P(1-P)\,W + W(W+n-1)\}^2
\end{aligned}
$$

where W is as above and where in addition

$$
\begin{aligned}
Z_0 &= (1-P)(n-1)(24P^4 + 7P^3n - 33P^3 + P^2n - 3P^2 + 3P + 1) \\
&\quad + (1-P)^3(16P^3 - 8P^2 - 5P - 1) \\
Z_1 &= 2(1-P)(4P^3 + P^2n - 2P^2 + 2Pn - 4P + n - 2)\ .
\end{aligned}
$$

6 Discussion

It is known that for very symmetric models of population structure with conservative migration, the mean coalescence time of two genes from any single subpopulation is independent of the precise migration pattern and migration rates and, in the notation of this paper, is equal to $c := \sum_{i\in\mathscr{S}} c_i$ (see Strobeck 1987 and Herbots 1994 for precise conditions under which this result holds; see also Li 1976). In the literature, there have been some overstatements of the generality with which this is true (see, for example, Slatkin 1993), and we stress that it is not at all true in general. Whereas for the models considered in Sect. 4 we have

that $ET_{ii} = c$ for every $i \in \mathcal{S}$, under the "less symmetric" models of Sect. 5 the value of ET_{ii} depends on M, for every $i \in \mathcal{S}$ (except in the case of the two-population model with $P = 1/2$). If in the two-population model, c_1 and c_2 are proportional to the sizes of the corresponding subpopulations, the mean coalescence time of a pair of genes from the larger subpopulation is larger than c, while the mean coalescence time of two genes from the smaller subpopulation is smaller than c (see results (74) and (76)). In the continental island model with $n > 2$, the mean coalescence time of a pair of genes from the continent is always larger than c, while that of two genes from the same island is always smaller than c, even if the continent is smaller than the islands (see results (83) and (84)). When in the two-population or continental island models all subpopulations are assigned equal weight, the mean coalescence time of a pair of genes sampled at random from the same subpopulation (ET_0) also depends on the migration rate (except in the case of the symmetric two-population model). Inappropriately assuming in the calculation of Slatkin's approximation $F_{ST}^{(0)}$ that $ET_0 = c$ may lead to highly inaccurate F_{ST} values (see Herbots 1994). Under weighting by size and assuming that the c_i correspond to the relative subpopulation sizes, the mean coalescence time of two genes sampled from the same subpopulation does satisfy $ET_0 = c$ for both the two-population and the continental island models, and in fact for any model of population structure for which the structured coalescent is an appropriate description of the genealogy of a sample, provided the population consists of a finite number of subpopulations all connected by migration and migration is conservative (Strobeck 1987; Herbots 1994; see Nagylaki 1998b for an extension of this result and Slatkin 1987 for a related result). We also note that in many cases (for example, when the size of subpopulation i is $2c_iN$ for every $i \in \mathcal{S}$, all subpopulations follow the same reproductive model and time-scaling is in units of a constant times N generations), c is (on the same time-scale) also equal to the mean coalescence time of a pair of genes sampled at random from a panmictic population of the same total size (and with the same reproductive model) as the subdivided population.

Our results for $\lim_{\theta \downarrow 0} \partial F_{ST}/\partial\theta$ indicate that for small mutation rates, the dependence of F_{ST} on the mutation rate may be more important than is generally believed. In particular, for a pair of subpopulations in the infinite one- or two-dimensional stepping-stone model, $\lim_{\theta \downarrow 0} \partial F_{ST}(d)/\partial\theta$ (or $\lim_{\theta \downarrow 0} \partial F_{ST}(d_1, d_2)/\partial\theta$) takes the value $-\infty$ (results (40) and (66)). Figure 7 shows the pairwise values of F_{ST} (thick lines) as a function of the scaled mutation rate θ under the infinite stepping-stone models in one and two dimensions. The thin lines give the values of the approximation $F_{ST}^{(0)}$. Note that, with the

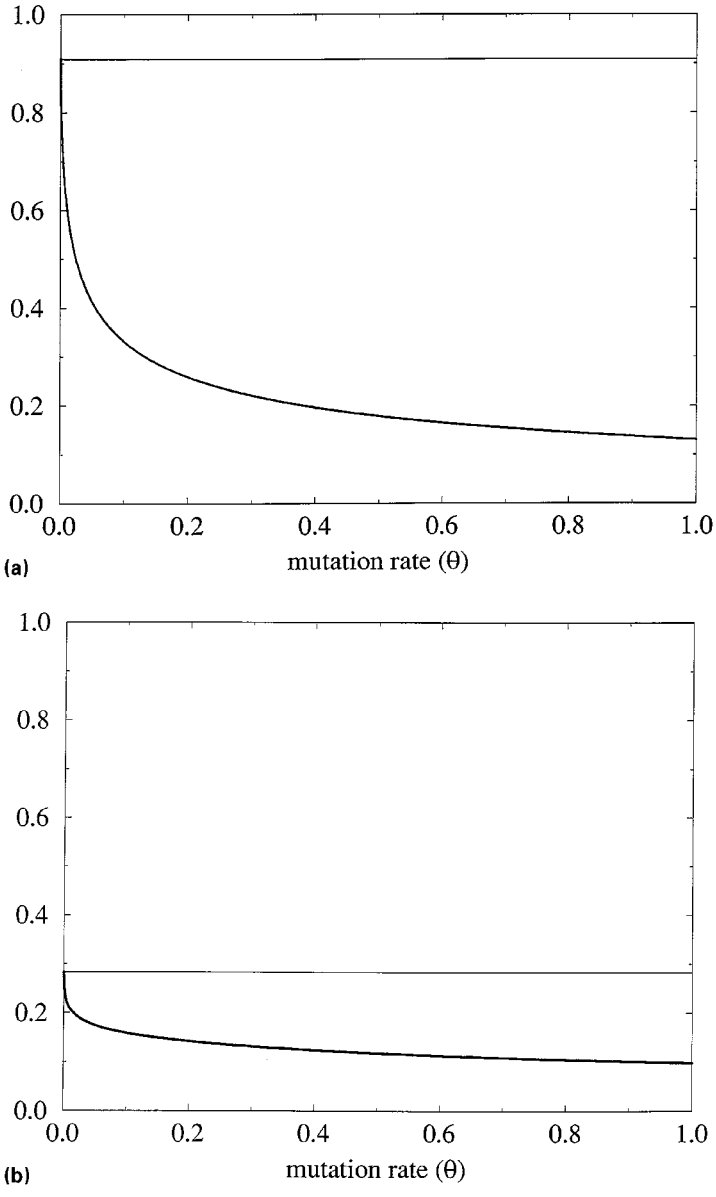

Fig. 7. Pairwise values of F_{ST} (curved lines) and its approximation $F_{ST}^{(0)}$ (straight lines) under an infinite linear stepping-stone model (**a**) and an infinite two-dimensional stepping-stone model (**b**). The distance between the two subpopulations considered is $d = 100$ in the linear stepping-stone model and $(d_1, d_2) = (0, 100)$ in the two-dimensional model. In both cases, a scaled migration rate of $M = 5$ was assumed, with equal migration rates $M_{(1)} = M_{(2)} = 2.5$ in the two dimensions of the two-dimensional model. The values for the two-dimensional model were obtained using the results in terms of single integrals, equations (61) and (64), applying an adaptive Newton–Cotes rule for numerical integration

same scaled migration rate, the amount of genetic differentiation between the pair of subpopulations (as measured by $F_{ST}(d)$ or $F_{ST}(d_1, d_2)$) is much larger in the one-dimensional stepping-stone model than in the two-dimensional model. For the stepping-stone model on the torus with equal migration rates in its two dimensions, the (global or pairwise) F_{ST} values are larger and depend more strongly on the mutation rate as the torus is narrower (i.e. closer to a circle) – see Crow and Aoki (1984) and Herbots (1994). In Herbots (1994) it was also illustrated (for the case of the 2×2 torus or lattice) that unequal migration rates in the two dimensions of a two-dimensional stepping-stone model can substantially increase both the global value of F_{ST} and the strength of its dependence on the mutation rate. From the results calculated in the present paper it can be shown that this is even more the case for pairwise F_{ST} values in two-dimensional stepping-stone models with a larger number of subpopulations. Among the models in Sect. 4, the dependence of F_{ST} on the mutation rate is weakest (and the value of F_{ST} lowest) under the finite island model (Herbots 1994).

Although it seems hard to prove this for an arbitrary model of population structure, the figures in Herbots (1994) suggest as a general fact that F_{ST} is a monotonically decreasing and convex function of the mutation rate (that F_{ST} decreases monotonically with increasing θ is easily proved for the – global or pairwise – values of F_{ST} under all the models in Sect. 4 and under the two-population model). This would imply that Slatkin's approximation $F_{ST}^{(0)}$, which is the limit of F_{ST} as the mutation rate tends to zero, constitutes an upper bound on F_{ST}, and that $\theta|\lim_{\theta \downarrow 0} \partial F_{ST}/\partial\theta|$ is an upper bound on the error made by using $F_{ST}^{(0)}$ as an approximation for F_{ST}.

The coefficient F_{ST} is commonly used to estimate levels of gene flow between subpopulations, assuming the island model of population structure (see Sect. 4.1) with a large number of subpopulations and a small mutation rate. However, it is easily seen from the results presented in this paper that the rate of decrease of F_{ST} with increasing migration rate M varies considerably between different population structures (see also Slatkin 1991, 1993). For example, under the finite island model

$$F_{ST}^{(0)} = \frac{1}{1 + Mn^2/(n-1)^2} \approx \frac{1}{1+M}$$

for large n, while the global value of $F_{ST}^{(0)}$ under the circular stepping-stone model is given by

$$F_{ST}^{(0)} = \frac{1}{1 + \frac{6n}{n^2-1} M}$$

which decreases much more slowly with increasing M. In a real population whose precise structure is not known, it is therefore not clear what relationship the estimated "effective" level of gene flow may bear to the actual frequency of migration (see also Slatkin and Barton 1989).

In this paper we have listed the theoretical values of coalescence times and F_{ST} for a range of models of population structure. Whereas the focus of this paper has been on the calculation of these results using the structured coalescent, Herbots (1994) contains a detailed discussion (based on many of the values presented here) of the dependence of F_{ST} on various parameters of population structure and on the mutation rate, and of the accuracy of Slatkin's approximation for F_{ST}.

Acknowledgements. The author thanks Peter Donnelly for valuable advice on the work presented in this paper, and Thomas Nagylaki and a referee for helpful suggestions. Much of this research was performed whilst the author was at the School of Mathematical Sciences, Queen Mary and Westfield College, University of London, as a Research Assistant of the Belgian National Fund for Scientific Research. Parts of this work were funded by the UK EPSRC (grant no. GR/G 11101) and by a Drapers' Company/QMW research studentship. The author is a Royal Society Dorothy Hodgkin Fellow.

References

Balding, D. J., Donnelly, P., Nichols, R. A.: Comment on "DNA fingerprinting: a review of the controversy" by K. Roeder. Statist. Sci. **9**, 248–251 (1994)

Balding, D. J., Nichols, R. A.: DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands. Forensic Sci. Inter. **64**, 125–140 (1994)

Balding, D. J., Nichols, R. A.: A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. Genetica **96**, 3–12 (1995)

Barton, N. H., Wilson, I.: Genealogies and geography. Phil. Trans. R. Soc. Lond. B **349**, 49–59 (1995a)

Barton, N. H., Wilson, I.: Genealogies and geography. In: P. H. Harvey et al.: New uses for new phylogenies (pp. 23–56) Oxford: Oxford University Press 1995b

Cannings, C.: The latent roots of certain Markov chains arising in genetics: A new approach. I. Haploid models. Adv. Appl. Prob. **6**, 260–290 (1974)

Chakraborty, R., Danker-Hopfe, H.: Analysis of population structure: A comparative analysis of different estimators of Wright's fixation indices. In: C. R. Rao and R. Chakraborty: Handbook of Statistics (Vol. 8, Chap 7: pp. 203–254) Amsterdam: Elsevier, North-Holland 1991

Cockerham, C. C., Weir, B. S.: Estimation of gene flow from F-statistics. Evolution **47**, 855–863 (1993)

Crow, J. F., Aoki, K.: Group selection for a polygenic behavioral trait: Estimating the degree of population subdivision. Proc. Nat. Acad. Sci. USA **81**, 6073–6077 (1984)

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

Feller, W.: An Introduction to Probability Theory and Its Applications, Vol. II. New York: Wiley 1966

Feller, W.: An Introduction to Probability Theory and Its Applications, Vol. I (3rd edition) New York: Wiley 1968

Foreman, L. A., Smith, A. F. M., Evett, I. W.: Bayesian analysis of deoxyribonucleic acid profiling data in forensic identification applications. J. R. Statist. Soc. A **160**, 429–459 (1997)

Griffiths, R. C.: The number of heterozygous loci between two randomly chosen completely linked sequences of loci in two subdivided population models. J. Math. Biol. **12**, 251–261 (1981)

Herbots, H. M.: Stochastic models in population genetics: genealogy and genetic differentiation in structured populations. Ph.D. thesis, University of London 1994

Herbots, H. M.: The structured coalescent. In: P. Donnelly and S. Tavaré: Progress in population genetics and human evolution (IMA Volumes in Mathematics and its Applications, vol. 87, pp. 231–255) New York: Springer-Verlag 1997

Hey, J.: A multi-dimensional coalescent process applied to multi-allelic selection models and migration models. Theoret. Popul. Biol. **39**, 30–48 (1991)

Hudson, R. R.: Gene genealogies and the coalescent process. In: D. J. Futuyma and J. Antonovics: Oxford Surveys in Evolutionary Biology (vol. 7, pp. 1–44) Oxford: Oxford University Press 1990

Kingman, J. F. C.: On the genealogy of large populations. Adv. Appl. Prob. **19A**, 27–43 (1982a)

Kingman, J. F. C.: The coalescent. Stoch. Proc. Appl. **13**, 235–248 (1982b)

Kingman, J. F. C.: Exchangeability and the evolution of large populations. In: G. Koch and F. Spizzichino: Exchangeability in probability and statistics (pp. 97–112) Amsterdam: North-Holland 1982c

Latter, B. D. H.: The island model of population differentiation: a general solution. Genetics **73**, 147–157 (1973)

Li, W.-H.: Distribution of nucleotide differences between two randomly chosen cistrons in a subdivided population: the finite island model. Theoret. Popul. Biol. **10**, 303–308 (1976)

Malécot, G.: Les Mathématiques de l'Hérédité. Paris: Masson 1948

Malécot, G.: Quelques schémas probabilistes sur la variabilité des populations naturelles. Ann. Univ. Lyon, Sciences, Sect. A **13**, 37 (1950)

Malécot, G.: The Mathematics of Heredity. San Francisco: Freeman 1969

Malécot, G.: Heterozygosity and relationship in regularly subdivided populations. Theoret. Popul. Biol. **8**, 212–241 (1975)

Maruyama, T.: Effective number of alleles in a subdivided population. Theoret. Popul. Biol. **1**, 273–306 (1970)

Morton, N. E.: Genetic structure of forensic populations. Proc. Nat. Acad. Sci. USA **89**, 2556–2560 (1992)

Nagylaki, T.: The strong-migration limit in geographically structured populations. J. Math. Biol. **9**, 101–114 (1980)

Nagylaki, T.: The robustness of neutral models of geographical variation. Theoret. Popul. Biol. **24**, 268–294 (1983)

Nagylaki, T.: The influence of spatial inhomogeneities on neutral models of geographical variation. I. Formulation. Theoret. Popul. Biol. **33**, 291–310 (1988)

Nagylaki, T.: Fixation indices in subdivided populations. Genetics, **148**, 1325–1332 (1998a)
Nagylaki, T.: The expected number of heterozygous sites in a subdivided population. Genetics, **149**, 1599–1604 (1998b)
Nagylaki, T., Barcilon, V.: The influence of spatial inhomogeneities on neutral models of geographical variation. II. The semi-infinite linear habitat. Theoret. Popul. Biol. **33**, 311–343 (1988)
Nagylaki, T., Keenan, P. T., Dupont, T. F.: The influence of spatial inhomogeneities on neutral models of geographical variation. III. Migration across a geographical barrier. Theoret. Popul. Biol. **43**, 217–249 (1993)
Nei, M.: Analysis of gene diversity in subdivided populations. Proc. Nat. Acad. Sci. USA **70**, 3321–3323 (1973)
Nei, M.: Molecular Population Genetics and Evolution. New York: American Elsevier, North Holland 1975
Nichols, R. A., Balding, D. J.: Effects of population structure on DNA fingerprint analysis in forensic science. Heredity **66**, 297–302 (1991)
Notohara, M.: The coalescent and the genealogical process in geographically structured population. J. Math. Biol. **29**, 59–75 (1990)
Sawyer, S.: Results for the stepping stone model for migration in population genetics. Annals of Probability **4**, 699–728 (1976)
Slatkin, M.: Gene flow in natural populations. Ann. Rev. Ecol. Syst. **16**, 393–430 (1985)
Slatkin, M.: The average number of sites separating DNA sequences drawn from a subdivided population. Theoret. Popul. Biol. **32**, 42–49 (1987)
Slatkin, M.: Inbreeding coefficients and coalescence times. Genet. Res., Camb. **58**, 167–175 (1991)
Slatkin, M.: Isolation by distance in equilibrium and non-equilibrium populations. Evolution **47**, 264–279 (1993)
Slatkin, M., Barton, N. H.: A comparison of three indirect methods for estimating average levels of gene flow. Evolution **43**, 1349–1368 (1989)
Strobeck, C.: Average number of nucleotide differences in a sample from a single subpopulation: a test for population subdivision. Genetics **117**, 149–153 (1987)
Takahata, N.: Gene identity and genetic differentiation of populations in the finite island model. Genetics **104**, 497–512 (1983)
Takahata, N.: The coalescent in two partially isolated diffusion populations. Genet. Res., Camb. **52**, 213–222 (1988)
Teugels, J. L.: Inleiding tot de discrete wiskunde, het kansrekenen en de statistiek. Deel 1: Discrete Wiskunde. Leuven: Acco 1986
Weir, B. S.: The effects of inbreeding on forensic calculations. Annu. Rev. Genet. **28**, 597–621 (1994)
Wright, S.: Evolution in Mendelian populations. Genetics **16**, 97–159 (1931)
Wright, S.: The genetical structure of populations. Annals of Eugenics **15**, 323–354 (1951)