Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Theor Popul Biol. Author manuscript; available in PMC 2016 June 23.

Published in final edited form as:

Theor Popul Biol. 2015 March ; 100C: 88–97. doi:10.1016/j.tpb.2014.12.004.

Covariation of gene frequencies in a stepping-stone lattice of populations[1]

Joseph Felsenstein[*]

Department of Genome Sciences and Department of Biology, University of Washington, Box 355065, Seattle, WA 98195-5065

Abstract

For a one- or two-dimensional lattice of finite length consisting of populations, each of which has the same population size, the classical stepping-stone model has been used to approximate the patterns of variation at neutral loci in geographic regions. In the pioneering papers by Maruyama (1970a, 1970b, 1971) the changes of gene frequency at a locus subject to neutral mutation between two alleles, migration, and random genetic drift were modeled by a vector autoregression model. Maruyama was able to use the spectrum of the migration matrix, but to do this he had to introduce approximations in which there was either extra mutation in the terminal populations, or extra migration from the subterminal population into the terminal population. In this paper a similar vector autoregression model is used, but it proves possible to obtain the eigenvalues and eigenvectors of the migration matrix without those approximations. Approximate formulas for the variances and covariances of gene frequencies in different populations are obtained, and checked by numerical iteration of the exact covariances of the vector autoregression model.

Keywords

Stepping-stone model; genetic drift; neutral mutation; migration; geographic differentiation

Stepping stone models of migration on rectangular lattices of populations have become of increasing interest as samples of many SNP loci have been collected in contiguous geographic areas, particularly in human populations. Stepping-stone models were pioneered independently by Malécot (1951) and by Kimura (1953; Kimura and Weiss, 1964). They considered lattices of infinite numbers of populations connected by migration in one and two dimensions, and derived expressions for the genetic variability expected in the populations in a balance between mutation and genetic drift of neutral alleles.

[1]This paper is dedicated to the memory of the late Takeo Maruyama (1936-1987)

This manuscript version is made available under the CC BY-NC-ND 4.0 license.

Mailing address: Joe Felsenstein, University of Washington, Department of Genome Sciences, Box 355065, Seattle, WA 98195-5065, phone number 206 543 0150, fax number 206 685 7301.

[*]joe@gs.washington.edu

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Lattices of finite size are of greater practical interest as models of real populations. Malécot (1948, 1950) pioneered them, using a model of a torus or circle of a finite number of populations. Maruyama (1970a, 1970b, 1971) was the first to consider stepping stone models for ordinary one- and two-dimensional linear or rectangular lattices of finite numbers of populations. He gave expressions for the variances and covariances of gene frequencies for arbitrary pairs of populations when there was a two-allele neutral mutation model in which the population had reached its equilibrium distribution. To do so he used an approximate model which had some lack of realism in the treatment of migration into the terminal populations of the lattice. He considered different variations in the way these terminal populations were modeled.

In the present paper I will treat a more exact model with a more realistic pattern of migration into the terminal populations. The results are similar, but not identical, to Maruyama's results.

Patterson et al. (2006) have pointed out the importance of using the eigenvectors and eigenvalues of the variation in gene frequencies in principal components analysis of data from different populations, an approach which goes back to Menozzi et al. (1978) and is reviewed by Cavalli-Sforza and Feldman (2003). An important issue is the interpretation of any significant patterns of geographic differentiation that are found. They do not necessarily indicate historical events such as waves of migration.

Novembré and Stephens (2008) have pointed out that the eigenvectors of a lattice model of migration are startlingly similar to principal components found for gene frequency patterns in geographic studies of genetic variation. When one of these principal components is seen, this makes it less obvious that it must arise from an historical invasion event, as the stepping stone models do not include historical invasions. They pointed out that migration matrices such as the ones Maruyama used fall into the class of Toeplitz matrices (Gray, 2006) for which the eigenvectors and eigenvalues can readily be computed. The more exact model which we use here has migration matrices that are not precisely Toeplitz matrices. It turns out that their eigenvalues have similarities with those of Maruyama's matrices, and their eigenvectors are readily found.

The immediate purpose of obtaining expressions for the covariances will be to approximate the joint distribution of gene frequencies in the populations by a multivariate normal distribution, for which the expectations and covariances are all that we need to determine the distribution. The expectations are easy to obtain; the covariances do still require one approximation but, when checked against an exact numerical solution of the equations, seem closer to the correct values than Maruyama's approximations are. The joint distribution of gene frequencies will not actually be normal, but for cases with small departures of gene frequencies from their expectations it will come close to being multivariate normal. Elsewhere I hope to discuss the development of an approximate maximum likelihood inference of the parameters of the model using this approximation. The formulas developed here may also be useful to others working with finite stepping-stone models who wish to have a closer approximation of the covariances of gene frequencies than has hitherto been available.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

1. The model

The finite stepping-stone model is a linear lattice of n_1 populations (if in one dimension) or a rectangular lattics of $n_1 \times n_2$ populations (if in two dimensions). Analogous models can be erected in higher numbers of dimensions. Each population has the same size, N individuals. The model has discrete, nonoverlapping generations. In each generation an infinite number of offspring are produced in each population. From these a random N are chosen to survive to be the adults of the next generation. The population sizes thus remain constant.

Each individual among the offspring is diploid, and has two parents. Migration enters the picture in the locations of the parents. For most of the populations in a one-dimensional lattice, there is a probability $1 - m$ that a particular individual drawn to be a parent comes from the same population. In a one-dimensional lattice there is a probability $\frac{1}{2}m$ that this parent comes from the population to the left, and a probability $\frac{1}{2}m$ that it comes from the population to the right. The exception is when the offspring population is the first or the last in the lattice. Then there is a probability $\frac{1}{2}m$ that it comes from the adjacent population, and the rest of the time, $1 - \frac{1}{2}m$ of the time, it comes from the same population.

If the lattice is a two-dimensional one, the same process is imagined to occur in both dimensions, completely independently. Thus for a population that is not on a boundary of the lattice, a parent for the (i, j) population may have come from that population, or from any of the 8 populations surrounding that one, as shown in Figure 1. It can have come from the four populations adjacent on diagonals with probability $\frac{1}{4}m^2$ each, and from the four populations adjacent in one direction with probability $\frac{1}{2}m(1 - m)$ each. With probability $(1 - m)^2$ it comes from population (i, j).

If the population is on one of the boundaries of the two-dimensional lattice, but not at a corner, its parents come from the two adjacent side populations with probability $\frac{1}{2}m(1 - \frac{1}{2}m)$ each, from the adjacent interior population with probability $\frac{1}{2}m(1 - m)$, from the two nearest diagonal populations with probability $\frac{1}{4}m^2$ each, and hence from the same population the remaining $(1 - \frac{1}{2}m)$ $(1 - m)$ of the time. When the population is in a corner of the lattice, a parent comes from the two nearby populations with probability $\frac{1}{2}m(1 - \frac{1}{2}m)$ each, from the one population one step away on the diagonal with probability $\frac{1}{4}m^2$, and from that population itself with the remaining probability $(1 - \frac{1}{2}m)^2$.

These seemingly complicated patterns are really just the consequence of having $\frac{1}{2}m$ come from each neighboring population in one dimension, with the two-dimensional pattern being independent movement in the two dimensions.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Figure 1 shows the migration rates in the present model in one- and two-dimensional lattices. This is slightly different from the migration pattern usually used in two-dimensional stepping-stone models. In most previous work, in two dimensions the migration can only come from one of the two populations immediately adjacent in one dimension, or immediately adjacent in the other dimension. Migrants could only come from populations $(i - 1, j)$, $(i + 1, j)$, $(1, j - 1)$, or $(i, j + 1)$. The present scheme, in which migration occurs or not in either dimension, independently, leads to greater mathematical tractability. It is worth noting that such a scheme is also implicit in Maruyama's papers.

The model follows the gene frequency of one allele at a locus, in the presence of migration, mutation, and genetic drift. The model of mutation has two alleles with mutation back and forth between them. If the mutation rate from A to a is μ, and the mutation rate from a to A is ν, the equilibrium gene frequency is $\bar{p} = \nu/(\mu + \nu)$. Mutation in such a one-locus model acts as if it were a form of migration. If we set an additional rate of migration into each population of $m_\infty = \mu + \nu$, and have these immigrant copies of the gene be drawn from a pool in which the gene frequency of A is $\bar{p}$, this will be indistinguishable from a model that has migration plus mutation between two alleles.

In the present model, parents are chosen according to the migration model, with the two parents of an individual independently drawn. Each parent contributes an allele to the offspring. Mutation occurs (or does not) for each copy of the gene. Each population thus has a pool of newborn offspring, whose genetic composition is characterized by the gene frequency of the A allele (which I will call p). Among these offspring, the model makes the presence or absence of the A allele at each copy from that pool independent of the other copies, so that we do not need to concern ourselves with the diploid genotype frequencies in the pool.

Genetic drift occurs by sampling N diploid individuals from the offspring pool, without replacement. As the presence of the A allele in each copy in the diploid individuals is independent, this has the same effect as drawing $2N$ times from a pool which has gene frequency p.

We can take the gene frequencies in the local populations and arrange them in a column vector, whose length is the number of populations. For the two-dimensional case this involves taking the gene frequencies in the rows of the array of populations, forming each into a column vector, and stacking them on top of each other, so that the first two entries in the vector are the gene frequencies for populations (1, 1) and (1, 2). Let the populations now be numbered in the order in which they appear in this vector.

For both one- and two-dimensional cases we will see below that we can use the migration pattern to create a migration matrix **M** whose elements m_{ij} are the probability that a copy of the gene found in the newborn offspring pool for population i came from a parent which was in population j, we can then write for population i the gene frequency in the next generation:

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$p_i' = (1 - m_\infty) \sum_{j=1}^{n} m_{ij} p_j + m_\infty \bar{p} + \varepsilon_i \quad (1)$$

where ε_i is the change due to genetic drift. The random variable p_i' is a binomial proportion after $2N$ trials with a probability of success equal to the sum of the first two terms on the right-hand side of this equation. The expectation of ε_i is thus zero.

This set of equations can be put into matrix form as

$$\mathbf{p}^{(t+1)} = (1 - m_\infty)\mathbf{M}\mathbf{p}^{(t)} + m_\infty \bar{p}\mathbf{1} + \varepsilon^{(t)}, \quad (2)$$

where the vector **1** is a column vector of 1's and $\mathbf{p}^{(t)}$ is the vector of gene frequencies in the populations in the adult stage of generation t. This equation is true for any pattern of recurrent migration, not just for stepping-stone lattices.

Note that, except for the error not being multivariate normally distributed, these equations are essentially a vector autoregression (VAR) model, beloved of econometricians. In this case they are (nearly) a VAR(1) model, as each generation depends only on the preceding one. In VAR models, the objective is to infer the regression coefficients from a series of observations. Here most of the regression structure is known, it is also known that the errors ε_i do not covary, and the objective is to predict the variances and covariances of the model at equilibrium.

This model is close to the models used by Maruyama (1970a, 1970b, 1971) but is not exactly the same. For the terminal population he considered two different models. One (his “absorbing boundary”) had migration rate $m/2$ into the terminal population from the subterminal population, and also an extra inflow of $m/2$ into from a pool at the equilibrium frequency. This allowed a fraction $1 - m$ of the terminal population to be nonmigrants, which made all the diagonal elements of the migration matrix equal. The second (his “reflecting boundary”) (Maruyama 1970a, 1970b) had migration at rate m from the subterminal population into the terminal population, rather than $m/2$. Again the migration matrix had all its diagonal elements equal. We will not adopt either of these measures, hoping to be able to cope with a migration matrix that does not have all of its diagonal values equal. It is also worth noting that Maruyama's variable **p** is not our **p**, but is the deviation of the gene frequency from its equilibrium value.

2. Transformation to uncorrelated variables

In the cases we consider, the matrix **M** will be symmetric. For all cases of symmetric migration among populations (not just for stepping-stone lattices of populations) it will be possible to write it in terms of its eigenvalues and eigenvectors as

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$\mathrm{M} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T \quad (3)$$

where the matrix **U** of eigenvectors is orthonormal, so that its transpose $\mathbf{U}^T$ is also its inverse. The diagonal elements of the matrix of eigenvalues $\mathbf{\Lambda}$ will be real numbers. Substituting this form of **M** into equation (2) we get

$$\mathbf{p}^{(t+1)} = (1 - m_\infty)\mathbf{U}\mathbf{\Lambda}\mathbf{U}^T\mathbf{p}^{(t)} + m_\infty \bar{p}\mathbf{1} + \varepsilon^{(t)}, \quad (4)$$

where **1** is the column vector whose entries are all 1. This equation can be premultiplied by $\mathbf{U}^T$ to get

$$\mathbf{U}^T\mathbf{p}^{(t+1)} = (1 - m_\infty)\mathbf{\Lambda}\mathbf{U}^T\mathbf{p}^{(t)} + m_\infty \mathbf{U}^T\bar{p}\mathbf{1} + \mathbf{U}^T\varepsilon^{(t)}, \quad (5)$$

since $\mathbf{U}^T$ is the inverse of **U** so that their product is the identity matrix and thus disappears from the matrix products.

We now consider a new vector of random variables

$$\mathbf{x}^{(t)} = \mathbf{U}^T\mathbf{p}^{(t)} \quad (6)$$

and write the vector $\mathbf{U}^T\bar{p}\mathbf{1}$ as **q**, which is a constant vector and not a random variable. We notice that since the expectations of the random variables $\varepsilon_i^{(t)}$ are zero, we can write the equation for $\mathbf{x}^{(t)}$ as

$$\mathbf{x}^{(t+1)} = (1 - m_\infty)\mathbf{\Lambda}\mathrm{x}^{(t)} + m_\infty\mathbf{q} + \boldsymbol{\eta}^{(t)} \quad (7)$$

The vectors of random variables $\boldsymbol{\eta}^{(t)}$ will have expectation zero.

The first eigenvalue λ_1 will be, for all the migration matrices we consider, 1, and its associated eigenvector will have all elements equal. The other eigenvectors will be orthogonal to this. It follows that **q** will have its first element nonzero, and all other elements zero.

3. Expectations

If we now take the expectations of the terms in equation (7), we can readily use equation (7), to show for each element of the vector **x** that in the long run that the process will approach a

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

stationary distribution. In that distribution, the expectations the elements of $x_i^{(t)}$ are each equal to the expectations of the corresponding elements of **q**. This is asymptotically true for t very large, when the effect of the initial frequencies of the $p_i^{(0)}$ have been lost. Going back to the elements of the vectors $\mathbf{p}^{(t)}$ in this limiting distribution, these all have equal expectations, Going back to the elements of the vectors $\mathbf{p}^{(t)}$ in this limiting distribution, these all have equal expectations, which are simply $\bar{p}$, the mutational equilibrium gene frequency. The equation for the expectations of the $\mathbf{x}^{(t)}$ in this limiting distribution, which will be called $\bar{\mathbf{x}}^{(t)}$, is then

$$\bar{\mathbf{x}}^{(t+1)} = (1 - m_\infty)\mathbf{\Lambda}\bar{\mathbf{x}}^{(t)} + m_\infty \mathbf{q}\,. \quad (8)$$

We can subtract this equation termwise from equation (7) and obtain an equation for the deviation of the $\mathbf{x}^{(t)}$ from their expectations. If we call these deviations $\mathbf{y}^{(t)}$, this is then

$$\mathbf{y}^{(t+1)} = (1 - m_\infty)\mathbf{\Lambda}\mathbf{y}^{(t)} + \boldsymbol{\eta}^{(t)} \quad (9)$$

It is easy to see that the expectations of the $\mathbf{y}^{(t)}$ are all zero. For this reason, the equation for the $\mathbf{y}^{(t)}$ is simpler than the equation for the $\mathbf{x}^{(t)}$, and we now use it to investigate the covariances.

4. Approximating the stochastic process

If we write, from equation (9), the expression for one of the elements of the vector $\mathbf{y}^{(t+1)}$, it is

$$y_i^{(t+1)} = (1 - m_\infty)\lambda_i y_i^{(t)} + \eta_i^{(t)} \quad (10)$$

If the elements of $\boldsymbol{\eta}^{(t)}$ were all independent of each other, it would be easy to show that the $y_i^{(t)}$ would be independent random variables. Actually, the $\eta_i^{(t)}$ are not, strictly speaking, independent. The vector $\boldsymbol{\eta}^{(t)}$ is a linear transformation of a set of independent binomial variates (each element of $\boldsymbol{\varepsilon}^{(t)}$ being the difference between a binomial frequency and its expectation). However the nonindependence is subtle and, for our purposes, unimportant. The objective of this paper is to derive the covariances of the gene frequencies, for use in a statistical inference when approximating the distribution of the gene frequencies as multivariate normal. It will be sufficient to be able to show that the $\boldsymbol{\eta}_i^{(t)}$ are approximately uncorrelated.

To do this another approximation will also be made. The $\varepsilon_i^{(t)}$ have variances that arise in the binomial sampling of the gene frequencies in going from the infinite number of newborn

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

individuals in a population to the finite number N of surviving adults. In any generation in which that gene frequency among the newborns is p_i, the binomial variance on sampling $2N$ copies of the gene is $p_i(1 - p_i)/(2N)$. At stationarity, the values of p_i across different generations have expectation $\bar{p}$ and a variance, which we can call σ_i^2. The expectation of the binomial variance across generations can be written in terms of this as

$$\mathbb{E}\left[\frac{1}{2N}p_i(1-p_i)\right] = \frac{1}{2N}(\mathbb{E}[p_i] - \mathbb{E}[p_i^2]) = \frac{1}{2N}(\bar{p} - \bar{p}^2 - \sigma_i^2) \quad (11)$$

At this point we do not know the values of the quantities σ_i^2. They depend on the very thing we want to compute, the variances and covariances of the gene frequencies p_i. There are likely to be differences between the variances σ_i^2, with terminal populations having a somewhat higher variance of gene frequency than interior populations.

In the limiting case when migration rates are large, the gene frequencies in all populations will be very similar. In that limit the variances σ_i^2 will be equal. The approximation made here will be to assume that we are near this limit, and that for the purposes of further calculation we can assume that all of the σ_i^2 are equal to the With that approximation, we can show that the random variables $\eta_i^{(t)}$ are not correlated. We have seen that they are the elements of the vector $\mathbf{U}^T\boldsymbol{\varepsilon}^{(t)}$. We know that the $\varepsilon_i^{(t)}$ have zero covariances, as they are independent random variables. If we use the approximation that the $\varepsilon_i^{(t)}$ have equal variances σ^2, then we can write the covariance matrix of the $\eta_i^{(t)}$ as

$$\mathbb{E}[\boldsymbol{\eta}\boldsymbol{\eta}^T] = \mathbb{E}[\mathbf{U}^T\varepsilon\varepsilon^T\mathbf{U}] = \mathbf{U}^T\mathbb{E}[\varepsilon\varepsilon^T]\mathbf{U} = \mathbf{U}^T(\frac{1}{2N}(\bar{p}(1-\bar{p}) - \sigma^2)\mathbf{I})\mathbf{U} = \frac{1}{2N}(\bar{p}(1-\bar{p}) - \sigma^2 \quad (12)$$
$$)\mathbf{U}^T\mathbf{U} = \frac{1}{2N}(\bar{p}(1-\bar{p}) - \sigma^2)\mathbf{I}$$

As this is diagonal, the covariances of different elements of $\boldsymbol{\eta}$ are thus, to close approximation, zero.

5. Covariances of gene frequencies

With the covariances of the $\eta_i^{(t)}$ now determined to (approximately) be zero, the random variables for the different i in equation (9) are now uncorrelated. They all have expectation zero. Its variance is easily determined by noting that the sum on the right-hand side is of two uncorrelated variables, whose covariance is therefore zero, so that:

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$\mathrm{Var}\left(y_i^{(t+1)}\right) = (1-m_\infty)^2\lambda_i^2\mathrm{Var}\left(y_i^{(t)}\right) + \mathrm{Var}\left(\eta_i^{(t)}\right) \quad (13)$$

Once this random process has reached stationarity, the variance of y_i in all generations is equal and, with our approximation of σ_i^2 by σ^2, we can then solve for the variance of y_i as

$$\mathrm{Var}(y_i) = \frac{\mathrm{Var}(\eta_i)}{1-(1-m_\infty)^2\lambda_i^2} = \frac{\bar{p}(1-\bar{p})-\sigma^2}{2N(1-(1-m_\infty)^2\lambda_i^2)}. \quad (14)$$

The remaining covariances are all zero.

The covariances of the x_i are the same as those of the y_i, since these vectors differ by a constant vector. To obtain the covariances of the gene frequencies p_i, we use the fact that the gene frequencies are a linear transformation of the y_i. From equation (6),

$$\mathbf{p} = \mathbf{U}(\mathbf{y}+\mathbf{q}), \quad (15)$$

so that since $\mathbf{q}$ is $\mathbf{U}^T\bar{p}\mathbf{1}$,

$$\mathbf{p} = \mathbf{U}\mathbf{y} + \bar{p}\mathbf{1} \quad (16)$$

so that

$$\mathbf{p} - \bar{p}\mathbf{1} = \mathbf{U}\mathbf{y}. \quad (17)$$

The covariances of $\mathbf{p}$ will then be

$$\mathbb{E}\left[(\mathbf{p}-\bar{p}\mathbf{1})(\mathbf{p}-\bar{p}\mathbf{1})^T\right] = \mathbb{E}[\mathbf{U}\mathbf{y}\mathbf{y}^T\mathbf{U}^T] = \mathbf{U}\mathbb{E}[\mathbf{y}\mathbf{y}^T]\mathbf{U}^T \quad (18)$$

From equation (14) we can now write the covariances of the p_i as

$$\mathrm{Cov}[\mathbf{p}] = \mathbf{U}\mathrm{diag}\left(\frac{\bar{p}(1-\bar{p})-\sigma^2}{2N(1-(1-m_\infty)^2\lambda_i^2)}\right)\mathbf{U}^T \quad (19)$$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Since the diagonal matrix in this expression can also be written in terms of the square of the diagonal matrix of eigenvalues, it turns out that this covariance matrix can also be written in terms of the inverse of a simple function of the migration matrix:

$$\mathrm{Cov}[\mathbf{p}] = (\bar{p}(1-\bar{p}) - \sigma^2)\left(2N\left(\mathbf{I} - (1-m_\infty)^2\mathbf{M}^2\right)\right)^{-1} \quad (20)$$

Of course, the diagonal elements of the covariance matrices are all assumed to be equal to σ^2. Although the above expressions compute σ^2 in terms of itself, we will find a way to untangle that.

We now know that the expectations of the p_i are all $\bar{p}$, and for the (approximated) process we have the covariances in terms of the eigenvalues and eigenvectors of **M**. Now all we need is to find those. The derivation since (1) has been general for any system of recurring migration whose migration rates among populations are symmetric, provided that the migration is strong enough to allow us to assume that the variances σ_i^2 are nearly equal. Now we need to use the particular pattern of migration on a finite lattice to obtain the eigenvalues and eigenvectors of the migration matrix.

6. The spectrum of the migration matrix

The matrix in one dimension is

$$\mathbf{M} = \begin{bmatrix} 1-\frac{1}{2}m & \frac{1}{2}m & 0 & \cdots & 0 & 0 & 0 \\ \frac{1}{2}m & 1-m & \frac{1}{2}m & \cdots & 0 & 0 & 0 \\ 0 & \frac{1}{2}m & 1-m & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1-m & \frac{1}{2}m & 0 \\ 0 & 0 & 0 & \cdots & \frac{1}{2}m & 1-m & \frac{1}{2}m \\ 0 & 0 & 0 & \cdots & 0 & \frac{1}{2}m & 1-\frac{1}{2}m \end{bmatrix} \quad (21)$$

For the one-dimensional case we will use n in place of n_1 for the number of populations, to reduce typographical stress.

Feller (1957, section XVI.3, pp. 389-390) treats a random walk with two reflecting boundaries, whose transition matrix is the above matrix for the case when $m = 1$. He uses a partial fractions method; as he notes, this is equivalent to a spectral decomposition of the

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

matrix. His quantity s_r is the reciprocal of the eigenvalue λ_{r+1}. Thus the eigenvalues for $m = 1$ are

$$\lambda_i = \cos\left(\frac{(i-1)\pi}{n}\right) \quad (22)$$

and if for $m = 1$ we call the matrix **R**, we have for general m

$$\mathbf{M} = (1-m)\mathbf{I} + m\mathbf{R} \quad (23)$$

so that the eigenvalues of **M** are the corresponding linear combinations which turn out to be

$$\lambda_i = 1 - m\left(1 - \cos\left(\frac{(i-1)\pi}{n}\right)\right) \quad (24)$$

The eigenvectors of **M** will be the same as those of **R**. In the Appendix it is shown that if these are scaled so as to be orthonormal the right eigenvectors can be written as

$$(\mathrm{U}^T)_{ij} = \sqrt{\frac{\Delta_i}{n}}\cos\left(\frac{(i-1)(2j-1)\pi}{2n}\right), \; i = 1, 2, \ldots, n, \; j = 1, 2, \ldots, n \quad (25)$$

where the quantity Δ_i is 1 if $i = 1$ and 2 otherwise. Maruyama found a similar quantity necessary in the expressions for his eigenvectors.

By comparison, Maruyama's eigenvalues were

$$\lambda_i = 1 - m\left(1 - \cos\left(\frac{i\pi}{n+1}\right)\right). \quad (26)$$

His eigenvectors could also be written in terms of cosines, but were different from the expressions for our matrix.

That there is a connection to Maruyama's matrix is made clearer if we consider a “shift matrix” **S** which has the diagonal above the main diagonal filled with 1s, and the rest of the elements zero:

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$\mathbf{S} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & 1 \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 \end{bmatrix} \quad (27)$$

One can write Maruyama's matrix as $(1-m)\mathbf{I} + \frac{1}{2}m\mathbf{S} + \frac{1}{2}m\mathbf{S}^T$. Our matrix $\mathbf{M}$ can also be written in terms of $\mathbf{S}$, as

$$\mathbf{M} = \mathbf{I} + \frac{1}{2}m\mathbf{S} + \frac{1}{2}m\mathbf{S}^T - \frac{1}{2}m(\mathbf{S}\mathbf{S}^T + \mathbf{S}^T\mathbf{S}). \quad (28)$$

7. Solving for the covariances

Using equation (19) together with equations (26) and (25) we can now write the covariance of populations in the one-dimensional case as

$$\mathrm{Cov}[p_i, p_j] = \frac{1}{2N}(\bar{p}(1-\bar{p}) - \sigma^2) \sum_{k=1}^{n} \frac{\Delta_k}{n} \cos\left(\frac{(k-1)(2i-1)\pi}{2n}\right) \cos\left(\frac{(k-1)(2j-1)\pi}{2n}\right) \quad (29)$$

$$\times \frac{1}{1-(1-m_\infty)^2\left(1-m\left[1-\cos\left(\frac{(k-1)\pi}{n}\right)\right]\right)^2}$$

For the two-dimensional case we note that the migration matrix $\mathbf{M}$ is the result of independent migration in the two dimensions; it turns out that

$$\mathbf{M} = \mathbf{M}^{(1)} \otimes \mathbf{M}^{(2)}, \quad (30)$$

where the superscripts indicate the two dimensions and $\otimes$ is the Kronecker product of matrices. It is well-known for Kronecker products that if $\mathbf{\Lambda}^{(1)}$ and $\mathbf{\Lambda}^{(2)}$ are the diagonal matrices of the eigenvalues of (respectively) $\mathbf{M}^{(1)}$ and $\mathbf{M}^{(2)}$, then the diagonal matrix of eigenvalues of $\mathbf{M}$ is the Kronecker product $\mathbf{\Lambda}^{(1)} \otimes \mathbf{\Lambda}^{(2)}$, which is the diagonal matrix whose diagonal elements are all $n_1 \times n_2$ products of one of the eigenvalues of $\mathbf{M}^{(1)}$ and one of the eigenvalues of $\mathbf{M}^{(2)}$. Similarly, the $n_1\, n_2 \times n_1\, n_2$ matrix $\mathbf{U}$ of eigenvectors of $\mathbf{M}$ is the Kronecker product of the $n_1 \times n_1$ matrix $\mathbf{U}^{(1)}$ of eigenvectors of $\mathbf{M}^{(1)}$ and the $n_2 \times n_2$ matrix $\mathbf{U}^{(2)}$ of eigenvectors of $\mathbf{M}^{(2)}$.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

If for the two-dimensional case we denote the gene frequency of the population at position (i, j) as p_{ij}, the result is that

$$\mathrm{Cov}[p_{ij}, p_{k\ell}] = \frac{1}{2N}(\bar{p}(1-\bar{p}) - \sigma^2) \sum_{g=1}^{n} \sum_{h=1}^{n} \frac{\Delta_g}{n} \frac{\Delta_h}{n} \cos\left(\frac{(g-1)(2i-1)\pi}{2n}\right) \cos \quad (31)$$

$$\left(\frac{(h-1)(2j-1)\pi}{2n}\right) \times \cos\left(\frac{(g-1)(2k-1)\pi}{2n}\right) \cos\left(\frac{(h-1)(2\ell-1)\pi}{2n}\right)$$

$$\times \frac{1}{1-(1-m_\infty)^2\left(1-m\left[1-\cos\left(\frac{(g-1)\pi}{n}\right)\right]\right)^2\left(1-m\left[1-\cos\left(\frac{(h-1)\pi}{n}\right)\right]\right)^2}$$

Higher numbers of dimensions can be accommodated in an exactly analogous way. The generalization of the two-dimensional case to having different migration rates m_1 and m_2 in the two dimensions is straightforward.

8. Solving for σ^2

There is still the vexing matter of the variance σ^2. One method of inferring it is to start with $\sigma^2 = 0$ and use equations (29) or (31) to compute all the Cov[i,i]. Then from these new estimates of the σ_i^2, make a new estimate of σ^2 by averaging these. One might want to continue this until the value of σ^2 converges. It actually does not converge if $4Nm < 1$. However if in each iteration we instead make a weighted average of two quantities, one the mean of the new σ_i^2 and the other the previous value of σ^2, with their weights $4Nm$ and 1, this seems always to converge rapidly. Let us call this the F1 approximation.

An alternative approach that seems reasonable is, for a given pair of populations, i and j, to

- Compute $\sigma^2 = \mathrm{Cov}[p_i, p_j]$ and do this iteratively until it converges (as before using the weighted average with weights $4Nm$ and 1). Use this for the computation of $\mathrm{Cov}[p_i, p_i]$.
- Do the same with $\mathrm{Cov}[p_j, p_j]$. Use this for the computation of $\mathrm{Cov}[p_j, p_j]$.
- For the computation of $\mathrm{Cov}[p_i, p_j]$ use the average of these two values of σ^2.

This will be referred to as the F2 approximation.

We will see that with $4Nm$ moderately large, it will make little difference which of these two methods of inferring σ^2 we use. The F2 method is more tedious computationally, requiring as it does multiple iterative estimations of σ^2.

9. Alternative equations

The covariances between populations can be arranged in a square matrix

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$\mathbf{C} = \left[\mathbb{E}[(p_i - \bar{p})(p_j - \bar{p})]\right]. \quad (32)$$

Using equation (2) we can show that at equilibrium **C** will satisfy

$$\mathbf{C} = (1 - m_\infty)^2 \mathbf{M}\mathbf{C}\mathbf{M}^T + \mathbf{Q}, \quad (33)$$

where **Q** is the covariance matrix of ε, which will be diagonal. This is a linear equation in the c_{ij}, though a big one: for example, for a 30 × 30 lattice the matrix **C** will be 900 × 900. This equation is general to any pattern of recurring migration.

It is worth noting that the linear equations in the c_{ij} can be rewritten in a more conventional form if we convert the covariance matrices **C** and **Q** into vectors by stacking their columns into stacks s(**C**) and s(**Q**) so that

$$s(\mathbf{C}) = (c_{11}, c_{21}, c_{31}, \ldots, c_{n1}, c_{12}, c_{22}, \ldots, c_{2n}, \ldots, c_{1n}, c_{2n}, \ldots, c_{nn})^T \quad (34)$$

and similary for s(**Q**). The equations then can be written using a Kronecker product:

$$s(\mathbf{C}) = (1 - m_\infty)^2 (\mathbf{M} \otimes \mathbf{M}) s(\mathbf{C}) + s(\mathbf{Q}). \quad (35)$$

Moving all the terms involving **C** to the left side

$$(\mathbf{I} - (1 - m_\infty)^2 \mathbf{M} \otimes \mathbf{M}) s(\mathbf{C}) = s(\mathbf{Q}) \quad (36)$$

whereby

$$s(\mathbf{C}) = (\mathbf{I} - (1 - m_\infty)^2 \mathbf{M} \otimes \mathbf{M})^{-1} s(\mathbf{Q}) \quad (37)$$

which is a slightly more general version of equation (20). The matrix inversion will always be possible if $m_\infty > 0$.

This is a very large set of equations – for a 30 × 30 lattice of populations there will be 810,000 equations in as many unknowns. It is unlikely to be a practical numerical way of solving for the c_{ij}, but use of a transformation analogous to equation (6),

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$\mathbf{x} = (\mathbf{U}^T \otimes \mathbf{U}^T)\mathrm{p}, \quad (38)$$

allows us to solve for the covariances of **p** in a way exactly analogous to our derivation above. The approximation of the σ_i^2 is the same in this set of equations as in the previous forms.

10. Longer-range migration

The migration matrix allows only a single step of migration in each dimension. One straightforward way to model multiple-step migration would be to allow in each dimension a Poisson-distributed number of steps of migration, with expectation m. Thus, in a one-dimensional lattice a fraction

$$\frac{e^{-m}m^k}{k!} \quad (39)$$

of the individuals have undergone k steps of migration, where each step is according to the matrix

$$\mathbf{R} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & \cdots & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & \cdots & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & \frac{1}{2} \\ 0 & 0 & 0 & \cdots & \frac{1}{2} & \frac{1}{2} \end{bmatrix}. \quad (40)$$

As noted above in the discussion of computing the eigenvalues and eigenvectors of **M**, the matrix **R** is the special case of the one-step migration matrix with $m = 1$. The resulting migration matrix when there is a Poisson-distributed number of migration steps is

$$\mathbf{M} = \sum_{k=0}^{\infty} \frac{1}{k!} e^{-m} m^k \mathbf{R}^k \quad (41)$$

This is true for any matrix **R** of one-step movements. The eigenvectors of a matrix polynomial like this are well-known to be the eigenvectors of **R**, which are the eigenvectors

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

of the one-step migration matrix, given in equation (25). The eigenvalues of **M** in the case of a lattice can be written in terms of the eigenvalues in equation (26) – the ith one is

$$\sum_{k=0}^{\infty} \frac{1}{k!} e^{-m} m^k \lambda_i^k = e^{-m} \sum_{k=0}^{\infty} \frac{1}{k!} (m\lambda_i)^k = e^{m(\lambda_i - 1)} \quad (42)$$

If we replace λ_i in this expression with the ith eigenvalue of the migration matrix that has $m = 1$, we get from equation (26) for the ith eigenvalue of the multistep migration matrix

$$\lambda_i = \exp\left(m\left[\cos\left(\frac{(i-1)\pi}{n}\right) - 1\right]\right) \quad (43)$$

For the gene frequency covariances, this can be substituted into equations (29) or (31) in place of the eigenvalue expressions there. The eigenvalues in equation (26) and those in equation (43) become asymptotically the same as m becomes small, as one would hope they would.

For two dimensions, tractability requires that the movement in the two dimensions be an independent random walk in each dimension. The number of steps in the x direction is a random Poisson variable, and so is the number of steps in the y direction, and these are independent. This will also be true if the total number of steps is a Poisson variate, and each step is independently chosen to be in the x direction or in the y direction, by independent Bernoulli variables (coin tosses). The eigenvalues are then all possible products of two quantities, each computed by equation (43).

11. Diffusion limit

It is well-known that if we consider a series of cases with increasing population sizes N, and decreasing values of the strengths of the deterministic evolutionary forces (here m_∞ and m), such that the products Nm_∞ and Nm are constant, and if we also observe the gene frequencies on a time scale in which one unit of time is N generations, that the stochastic process of gene frequency change approaches a diffusion process. This is of considerable interest, as the approximation is usually very close.

Taking these limits, terms in m_∞^2 and m^2, and all of their higher powers drop out of the equations. The result for the covariances is that equation (29) approaches the limit

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$\mathrm{Cov}[p_i, p_j] = (\bar{p}(1-\bar{p}) - \sigma^2) \sum_{k=1}^{n} \frac{\Delta_k}{n} \cos\left(\frac{(k-1)(2i-1)\pi}{2n}\right) \cos\left(\frac{(k-1)(2j-1)\pi}{2n}\right) \quad (44)$$

$$\times \frac{1}{4Nm_\infty + 4Nm\left(1 - \cos\left(\frac{(k-1)\pi}{n}\right)\right)}$$

In the case of multiple-step migration, as we take N larger, m becomes small, and the occurrence of more than one step of migration becomes vanishingly rare, so it is easy to show that the diffusion limit is this same expression. The analogous expressions for the two-dimensional lattice are easily obtained and involve two terms in $4Nm$.

12. Fleming and Su's approximation

Fleming and Su (1974) have derived another approximation to the finite stepping-stone models, by approximating the space as continuous. We will not give their formulas here, but below we will compare them to our approximations. I have showed elsewhere (Felsenstein, 1975) that models of finite populations of organisms migrating in continuous geographical spaces encounter some difficulties in maintaining a Poisson random field distribution. Fleming and Su's equations cannot be seen as exact for a organisms randomly distributed in a continuous geographical space. We must instead regard them as an approximation to the stepping-stone model.

To do this I have taken their interval $(-\ell, \ell)$ and divided it into n equal intervals, with the n populations of the stepping-stone model being regarded as located at the midpoints of those intervals. This is to some extent an arbitrary choice.

13. Numerical comparisons

The approximation involved in using a common σ^2 for all of the variances is expected to be a good one if $4Nm \gg 1$, but it is worth doing numerical checks. Table 1 shows numerical comparisons. The exact covariances for the model are calculated using equation (33), with the variances in the diagonal matrix **Q** obtained from equation (11). This equation was iterated many times until it converged.

These values, exact under our model, were compared with the approximations in equation (29). They were also compared with Maruyama's approximations in his 1970a paper, Table 4, and also with with the continuous approximation of Fleming and Su (1974). Table 1 shows the results for $n = 10$, $m_\infty = 0.001$, $m = 0.1$, $N = 25$, and $\bar{p} = 0.2$, the case that Maruyama considered. The exact value is denoted by E, our approximations by F1 and F2, Maruyama's approximation by M, and Fleming and Su's approximation by FS. Owing to the symmetry of the case, populations 6 through 10 have the same variances as populations 5 through 1 so they are not shown.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

14. Numerical comparisons

The approximation involved in using a common σ^2 for all of the variances is expected to be a good one if $4Nm \gg 1$, but it is worth doing numerical checks. Table 1 shows numerical comparisons. The exact covariances for the model are calculated using equation (33), with the variances in the diagonal matrix **Q** obtained from equation (11). This equation was iterated many times until it converged.

These values, exact under our model, were compared with the approximations in equation (29). They were also compared with Maruyama's approximations in his 1970a paper, Table 4, and also with with the continuous approximation of Fleming and Su (1974). Table 1 shows the results for $n = 10$, $m_\infty = 0.001$, $m = 0.1$, $N = 25$, and $\bar{p} = 0.2$, the case that Maruyama considered. For Fleming and Su's approximation, which uses a continuous space on the interval (−1, 1), the locations corresponding to the n populations were taken to be $1 - \frac{1}{n}, 1 - \frac{3}{n}, 1 - \frac{5}{n}, \ldots, -\frac{1}{n}, \frac{1}{n}, \ldots, 1 - \frac{1}{n}$. These are n points equally spaced, each being the center of an interval which is $\frac{1}{n}$ of the total interval, which is of length 2. The exact value is denoted by E, our approximations by F1 and F2, Maruyama's approximation by M, and Fleming and Su's approximation by FS. Owing to the symmetry of the case, populations 6 through 10 have the same variances as populations 5 through 1 so they are not shown.

While generally similar, then two approximations differ noticeably, especially for the terminal population. In Maruyama's approximation, there is an extra injection of mutation into the terminal populations. This damps its departure from the mutational equilibrium which is caused by genetic drift. Both approximations are considerably lower than the true variance, the expected variance in the F1 approximation developed in this paper being 18% low in the terminal population and 26% low in the central population. The F2 approximation is slightly worse in the populations near the ends, and slightly better in the central region, but perhaps not enough to make it worthwhile in view of its higher computational burden. The Fleming-Su approximation is worse. It does agree with the other values by showing higher variance in the terminal populations.

Table 1 has $4Nm = 10$ and $4Nm_\infty = 0.1$. When $4Nm = 100$, as in Table 2, the F1 approximation is much better, being 0.56% high in the terminal populations, and 0.3% low in the center populations. The F2 approximation is again lower than the F1 approximation in the end region and higher than it in the central region; in most populations it is farther from the exact value than is the F1 approximation.

In longer one-dimensional stepping stone models, as in Table 3, we can see that the exact variances in the terminal populations are higher than in the center, but that this rapidly declines and becomes relatively constant as we move toward the center of the lattice. The F1 approximation tracks this quite closely, being about 3.6% too high in the terminal populations and 0.5% too low in the center. The F2 approximation is in general closer to the exact value, and through most of the central region is nearly exact. The Fleming-Su approximation does even less well than before.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

We can also compare the correlations between the gene frequencies of pairs of populations. Table 4 shows these for the same case as Table 1 above, where $N = 25$. For this case Maruyama also calculated the correlations from his reflecting boundary approximation, in Table 4 of his 1970b paper. Both of our approximations are very close, and his quite a bit farther away. The (1,9) element of this table has the greatest departure from the exact value for both our and Maruyama's approximation, and his is 8.24% high while ours is 0.67% low. (There is a typographical error in the Theoretical number in the (9,10) element of his table). The F2 approximation is slightly better than the F1 approximation for pairs of populations near the ends of the region, and slightly better for all other pairs.

Table 5 shows the correlations for the case when $N = 250$. It can be shown that the F1 approximation will not depend on N, so our predictions are the same as in Table 4. In fact, the expressions for predicted correlations using our approximation also do not depend on $\bar{p}$ or on m_∞ either – they depend only on the geometry of the populations and on m. Compared to these predictions, the exact values are a bit more different in Table 5, being up to 5% low (in the case of the (10,1) element), but more typically being about 1% low. The F2 approximations do depend on N, but for the correlations they are so close to the values of the F1 approximation that there seems little point going to the extra effort of computing them.

15. Usefulness

The present solution is approximate, but less so that previous efforts. It is intended for use in a likelihood-based (or Bayesian-based) inference of migration and/or mutation parameters, scaled as a fraction of the local effective population size. However the approximation is specific to rectangular lattices which have reached equilibrium between mutation, migration, and genetic drift. Both aspects, the rectangularity and the equilibrium, may be questioned for natural populations, the equilibrium being a particularly severe challenge for human populations. In continents such as Europe, we need assurance that there has been enough time since historical population movements such as those associated with the spread of agriculture. Given perhaps only 7000 years since the establishment of agriculture there, that is a stringent requirement.

The availability of approximations based on processes in a continuum of finite length, instead of the stepping stone lattice, presents another challenge. We have compared our formulas for the variances with Fleming and Su's (1974) approximation, and for these cases the continuous approximation was a worse approximation. Barton and Wilson (1995) have developed approximations for coalescent times for an approximate model of a continuum of finite length. The papers by Barton et al. (2002, 2010) make similar approximations that allow for extinction and recolonization, which have not been discussed here. These papers use, respectively, a model of a two-dimensional infinite continuum, and a model of a finite torus. Wilkins and Wakeley (2002) and Wilkins (2004) use a different approximate model to develop other approximations to coalescence times in a one- or two-dimensional finite continuum. Neither of these approximations has been developed into an approximation for variances and covariances of local gene frequencies, although it does not seem difficult to do so. Barton and Wilson (1995) approximate these covariances by using Malécot's (1951) formulae for the one- and two-dimensional infinite continuum approximation.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

If the stepping-stone model is regarded as the truth, and these continuum models are intended as approximations to it, the present formulas may help evaluate the accuracy of the continuum approximations. But to the extent that both the stepping-stone models and the continuum models are regarded as approximations to the more subtle structure of real populations, we would have to model those, perhaps in individual-based simulations, to evaluate whether either approach is viable.

Acknowledgments

Early work on this project was supported by NIH grant R01 GM071639-01A1/G126IE. Later work was partially supported by NSF grant DEB 0742517, Mary Kuhner and Kevin McCracken, principal investigators. It was also supported by "life support" interim salary support from the Department of Genome Sciences of the University of Washington, by NSF Cooperative Agreement No. DBI-0939454 (BEACON), and by NSF grant DEB 1019583, J. Felsenstein and F. L. Bookstein co-PIs. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

I wish to thank Nick Barton, William G. Hill, Monty Slatkin, and Eric Rynes for helpful comments on an earlier version of the manuscript, and the reviewers for this journal for their very helpful remarks.

Appendix

Feller (1957, section XVI.3, p. 390) gives the elements of the ith right eigenvector of matrix $\mathbf{M}$ when $m = 1$ as

$$u_{ij} = \sin\left(\frac{(i-1)j\pi}{n}\right) - \sin\left(\frac{(i-1)(j-1)\pi}{n}\right), \quad (45)$$

where they are determined up to an arbitrary multiplicative constant.

Note the trigonometric relationship

$$\sin(\theta+\phi) + \sin(\theta-\phi) = 2\cos(\theta)\sin(\phi) \quad (46)$$

which may readily be verified using the addition law for sines.

Taking

$$\theta = \frac{1}{2}\left(\frac{(i-1)j\pi}{n} + \frac{(i-1)(j-1)}{n}\right) = \frac{(i-1)(2j-1)\pi}{2n} \quad (47)$$

and

$$\phi = \frac{1}{2}\left(\frac{(i-1)j\pi}{n} - \frac{(i-1)(j-1)\pi}{n}\right) = \frac{(i-1)\pi}{2n} \quad (48)$$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

we find that the elements of the eigenvector can be written as

$$u_{ij} = \cos\left(\frac{(i-1)(2j-1)\pi}{2n}\right)\sin\left(\frac{(i-1)\pi}{2n}\right) \quad (49)$$

The sine factor does not depend on j, so it is a scaling of the ith eigenvector. These eigenvectors each need to be rescaled so that the eigenvectors are orthonormal. If they are written as

$$u_{ij} = \sqrt{\frac{\Delta_i}{n}}\cos\left(\frac{(i-1)(2j-1)\pi}{2n}\right) \quad (50)$$

where Δ_i is 1 for $i = 1$ and 2 otherwise, it can be shown that this is the appropriate scaling.

References

Barton NH, Wilson I. Genealogies and geography. Philos Trans R Soc Ser B, Biol Sci. 1995; 349:49–59.

Barton NH, Depaulis F, Etheridge AM. Neutral evolution in spatially continuous populations. Theor Popul Biol. 2002; 61:31–48. [PubMed: 11895381]

Barton NH, Etheridge AM, Véber A. A new model for evolution in a spatial continuum. Electron J Probab. 2010; 15:162–216.

Cavalli-Sforza LL, Feldman MW. The application of molecular genetic approaches to the study of human evolution. Nat Genet. 2003; 33(Supplement):266–275. [PubMed: 12610536]

Felsenstein J. A pain in the torus: some difficulties with models of isolation by distance. Am Nat. 1975; 109:359–368.

Fleming WH, Su CH. Some one-dimensional migration models in population genetics theory. Theor Popul Biol. 1974; 5:431–449. [PubMed: 4460257]

Gray RM. Toeplitz and circulant matrices: a review. Found Trends Commun Inf Theory. 2006; 2(3): 155–239.

Kimura M. “Stepping-stone” model of population. Annu Rep Natl Inst Genet, Jpn. 1953; 3:62–63.

Kimura M, Weiss GH. The stepping stone model of population structure and the decrease of genetic correlation with distance. Genet. 1964; 49:561–576.

MalécotG. Les mathématiques de l'hereditéMasson; Paris: 1948

Malécot G. Quelques schémas probablistes sur la variabilité des populations naturelles. Ann Univ Lyon, Sci, sect A. 1950; 13:37–60.

Malécot G. Un traitement stochastique des problèmes linéaires (mutation, linkage, migration) en génétique de populations. Ann Univ Lyon, Sci, sect A. 1951; 14:79–117.

Maruyama T. Stepping stone models of finite length. Adv Appl Probab. 1970a; 2(2):229–258.

Maruyama T. Analysis of population structure. I. One-dimensional stepping-stone models of finite length. Ann Hum Genet. 1970b; 34:201–214. [PubMed: 5493850]

Maruyama T. Analysis of population structure. II. Two-dimensional stepping stone models of finite length and other geographically structured populations. Ann Hum Genet. 1971; 35:179–196. [PubMed: 5159533]

Menozzi P, Piazza A, Cavalli-Sforza L. Synthetic maps of human gene frequencies in Europeans. Sci. 1978; 201:786–792.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Novembré J, Stephens M. Interpreting principal component analyses of spatial variation. Nat Genet. 2008; 40(5):646–649. [PubMed: 18425127]

Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. 2006; 2(12):e190. [PubMed: 17194218]

Wilkins JF, Wakeley J. The coalescent in a continuous, finite, linear population. Genet. 2002; 161:873–888.

Wilkins JF. A separation-of-timescales approach to the coalescent in a continuous population. Genet. 2004; 168:2227–2244.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

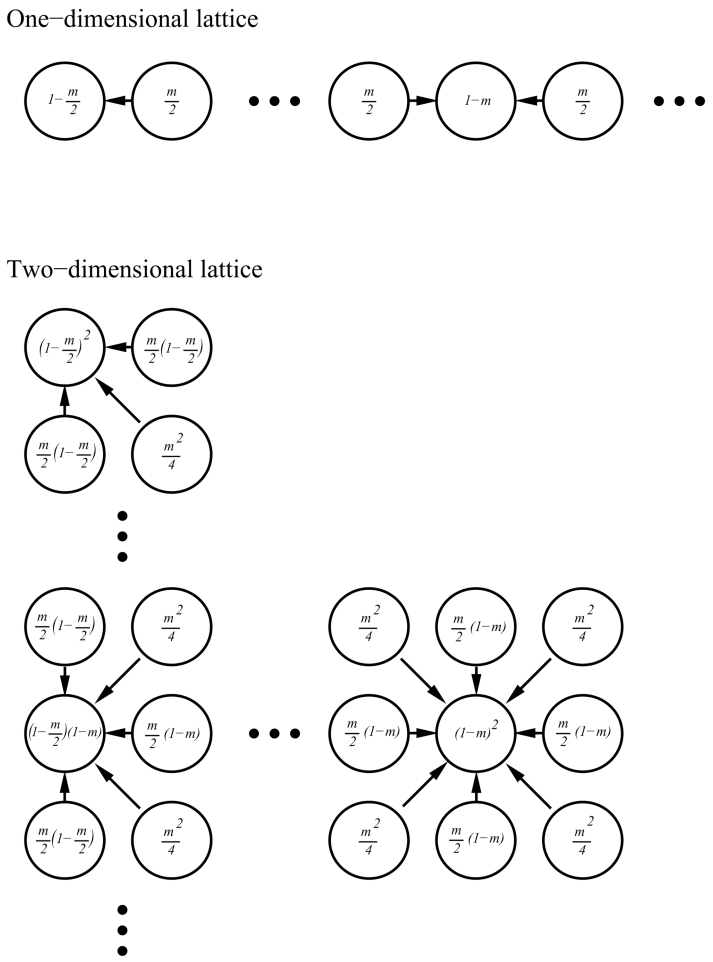

Figure 1.
Diagram of the migration pattern in a one- and two-dimensional stepping stone model. In each case the recipient population is shown together will all populations contributing migrants to it, with the fraction of the recipient population coming from each of these populations shown. This is shown for an interior population and a terminal population for the one-dimensial stepping stone model, and for an interior population, a side population, and a corner population for a two-dimensional stepping stone model.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 1

Comparison of exact solution (E) for the variances of population gene frequencies in a 10-stone stepping-stone model for $n = 10$, $m_\infty = 0.001$, $m = 0.1$, $N = 25$, and $\bar{p} = 0.2$ with the approximations of this paper (F1 and F2), Maruyama's reflecting-boundary approximation (M), and Fleming and Su's (1974) approximation. Approximation F1 uses the value of σ^2 averaged over all populations. Approximation F2 uses a value of σ^2 calculated for that population. As the migration pattern is symmetrical, only the values for populations 1-5 are shown.

population	E	F1	F2	M	FS
1	0.1281	0.1046	0.0961	0.0982	0.0811
2	0.1161	0.0946	0.0922	0.0928	0.0720
3	0.1121	0.0877	0.0892	0.0879	0.0665
4	0.1107	0.0834	0.0872	0.0857	0.0634
5	0.1103	0.0814	0.0862	0.0846	0.0619

Theor Popul Biol. Author manuscript; available in PMC 2016 June 23.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2

Comparison of exact solution (E) for the variances of population gene frequencies in a 10-stone stepping-stone model for $n = 10$, $m_\infty = 0.001$, $m = 0.1$, $N = 250$, and $\bar{p} = 0.2$ with the approximation of this paper (F). Approximation F1 uses the value of σ^2 averaged over all populations. Approximation F2 uses a value of σ^2 calculated for that population. FS is the Fleming-Su approximation. As the migration pattern is symmetric, only the values for populations 1-5 are shown.

population	E	F1	F2	FS
1	0.02115	0.02127	0.02089	0.01396
2	0.01918	0.01924	0.01913	0.01207
3	0.01784	0.01784	0.01790	0.01087
4	0.01700	0.01697	0.01712	0.01016
5	0.01660	0.01655	0.01674	0.00984

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 3

Comparison of exact solution (E) for the variances of population gene frequencies in a 100-stone stepping-stone model for $n = 100$, $m_\infty = 0.001$, $m = 0.1$, $N = 250$, and $\bar{p} = 0.2$ with the approximations of this paper (F1 and F2). Approximation F1 uses the value of σ^2 averaged over all populations. Approximation F2 uses a value of σ^2 calculated for that population. As the migration pattern is symmetric, only values for the first half of the lattice are shown.

population	E	F1	F2
1	0.01900	0.01968	0.01870
2	0.01691	0.01744	0.01679
3	0.01534	0.01575	0.01532
4	0.01416	0.01447	0.01419
5	0.01328	0.01351	0.01333
10	0.01126	0.01139	0.01129
20	0.01066	0.01061	0.01066
30	0.01062	0.01057	0.01062
40	0.01062	0.01057	0.01062
50	0.01062	0.01057	0.01062

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 4

Exact solution and two approximations for the correlations of population gene frequencies for $n = 10$, $m_\infty = 0.001$, $m = 0.1$, $N = 25$, and $\bar{p} = 0.2$. Within each nondiagonal cell, the uppermost value is the exact solution, the next two values our F1 and F2 approximations, and the lower value is Maruyama's reflecting-boundary approximation, from Table 4 of his 1970b paper.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	1.0									
2	0.9292	1.0								
	0.9286									
	0.9290									
	0.9410									
3	0.8607	0.9223	1.0			Exact				
	0.8600	0.9223				F1				
	0.8611	0.9225				F2				
	0.8795	0.9264				M				
4	0.7931	0.8499	0.9171	1.0						
	0.7924	0.8499	0.9174							
	0.7940	0.8504	0.9175							
	0.8175	0.8586	0.9209							
5	0.7287	0.7808	0.8424	0.9139	1.0					
	0.7279	0.7807	0.8428	0.9144						
	0.7297	0.7814	0.8430	0.9144						
	0.7541	0.7928	0.8496	0.9177						
6	0.6690	0.7167	0.7731	0.8386	0.9128	1.0				
	0.6679	0.7164	0.7734	0.8392	0.9134					
	0.6696	0.7170	0.7736	0.8392	0.9134					
	0.6971	0.7315	0.7834	0.8455	0.9166					
7	0.6152	0.6589	0.7106	0.7705	0.8386	0.9139	1.0			
	0.6136	0.6581	0.7105	0.7709	0.8392	0.9144				
	0.6149	0.6585	0.7106	0.7709	0.8392	0.9144				
	0.6427	0.6756	0.7233	0.7804	0.8455	0.9177				
8	0.5676	0.6080	0.6555	0.7106	0.7731	0.8424	0.9171	1.0		

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

	1	2	3	4	5	6	7	8	9	10
	0.5655	0.6065	0.6548	0.7105	0.7734	0.8428	0.9174			
	0.5662	0.6066	0.6548	0.7106	0.7736	0.8430	0.9175			
	0.5974	0.6268	0.6709	0.7233	0.7834	0.8496	0.9209			
9	0.5265	0.5639	0.6080	0.6589	0.7167	0.7808	0.8499	0.9223	1.0	
	0.5238	0.5618	0.6065	0.6581	0.7164	0.7807	0.8499	0.9223		
	0.5240	0.5618	0.6066	0.6585	0.7170	0.7814	0.8504	0.9225		
	0.5571	0.5855	0.6268	0.6756	0.7315	0.7928	0.8586	0.9264		
10	0.4916	0.5265	0.5676	0.6152	0.6690	0.7287	0.7931	0.8607	0.9292	1.0
	0.4883	0.5238	0.5655	0.6136	0.6679	0.7279	0.7924	0.8600	0.9286	
	0.4883	0.5240	0.5662	0.6149	0.6696	0.7297	0.7940	0.8611	0.9290	
	0.5321	0.5571	0.5974	0.6427	0.6971	0.7541	0.8175	0.8795	0.9410	

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 5

Comparison of exact solution (E) for the correlations of population gene frequencies in a 10-stone stepping-stone model for $n = 10$, $m_\infty = 0.001$, $m = 0.1$, $N = 250$, and $\bar{p} = 0.2$ with the approximation of this paper. The 10 × 10 lower triangle contains three numbers, the uppermost one being the exact correlation and the lower two our approximations F1 and F2. Approximation F1 is calculated with one value of σ^2 for all populations; approximation F2 is calculated using variances from approximations of σ^2 local to the two populations involved, and an average of these values is used to compute the covariance between the two populations.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	1.0									
2	0.9332	1.0								
	0.9286									
	0.9287									
3	0.8657	0.9226	1.0			Exact				
	0.8600	0.9223				F1				
	0.8600	0.9223				F2				
4	0.7985	0.8501	0.9151	1.0						
	0.7924	0.8499	0.9174							
	0.7925	0.8499	0.9174							
5	0.7350	0.7816	0.8398	0.9106	1.0					
	0.7279	0.7807	0.8428	0.9144						
	0.7279	0.7807	0.8428	0.9144						
6	0.6770	0.7193	0.7715	0.8346	0.9092	1.0				
	0.6679	0.7164	0.7734	0.8392	0.9134					
	0.6680	0.7164	0.7734	0.8392	0.9134					
7	0.6259	0.6645	0.7115	0.7681	0.8346	0.9106	1.0			
	0.6136	0.6581	0.7104	0.7709	0.8392	0.9144				
	0.6136	0.6581	0.7105	0.7709	0.8392	0.9144				
8	0.5820	0.6174	0.6603	0.7115	0.7715	0.8398	0.9151	1.0		
	0.5655	0.6065	0.6548	0.7104	0.7734	0.8428	0.9174			
	0.5655	0.6065	0.6548	0.7105	0.7734	0.8428	0.9174			
9	0.5449	0.5778	0.6174	0.6645	0.7193	0.7816	0.8501	0.9226	1.0	
	0.5238	0.5618	0.6065	0.6581	0.7164	0.7807	0.8499	0.9223		

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

	1	2	3	4	5	6	7	8	9	10
	0.5238	0.5618	0.6065	0.6581	0.7164	0.7807	0.8499	0.9223		
10	0.5140	0.5449	0.5820	0.6259	0.6770	0.7350	0.7985	0.8657	0.9332	1.0
	0.4883	0.5238	0.5655	0.6136	0.6679	0.7279	0.7924	0.8600	0.9286	
	0.4883	0.5238	0.5655	0.6136	0.6680	0.7279	0.7925	0.8600	0.9287	