doi: 10.1111/j.1365-294X.2004.02393.x

Seeing ghosts: the effect of unsampled populations on migration rates estimated for sampled populations

MONTGOMERY SLATKIN

Department of Integrative Biology, University of California, Berkeley, CA 94720–3140, USA

Abstract

In 2004, the term 'ghost population' was introduced to summarize the effect of unsampled subpopulations that exchange migrants with other subpopulations that have been sampled. Estimated long-term migration rates among populations sampled will be affected by ghost populations. Although it would be convenient to be able to define an apparent migration matrix among sampled populations that incorporate the exchange of migrants with ghost populations, no such matrix can be defined in a way that predicts all features of the coalescent process for the true migration matrix. This paper shows that if the underlying migration matrix is symmetric, it is possible to define an apparent migration matrix among sampled subpopulations that predicts the same within-population and between-population homozygosities among sampled populations as is predicted by the true migration matrix. Application of this method shows that there is no simple relationship between true and apparent migration rates, nor is there a way to place an upper bound on the effect of ghost populations. In general, ghost populations can create the appearance of migration between subpopulations that do not actually exchange migrants. Comparison with published results from the application of the program, MIGRATE, shows that the apparent migration rates inferred with that program in a three-subpopulation model differ from those based on pairwise homozygosities. The apparent migration matrix determined by the method described in this paper probably represents the upper bound on the effect of ghost populations.

Keywords: F_{ST}, gene flow, population structure

Received 12 July 2004; revision received 29 September 2004; accepted 29 September 2004

Introduction

In a species distributed over a wide geographical area, it is almost never possible to sample all parts of the range, and hence it is almost never possible to have a complete picture of the geographical pattern of genetic variation. Instead, samples are taken from some locations in the hope that those samples are typical and permit inference about the species as a whole. The problem that will be addressed in this paper is the extent to which this hope can be realized when the goal is to estimate migration rates among locations sampled.

A variety of indirect methods are available to estimate migration rates from genetic data (Neigel 1997). All methods rely on population genetic models whose assumptions may not be satisfied by real populations. Methods based on Wright's F_{ST} (Slatkin 1993; Rousset 1997) and methods that rely on an underlying multipopulation coalescent process (Beerli & Felsenstein 1999; Bahlo & Griffiths 2000; Beerli & Felsenstein 2001) assume that a long-term genetic equilibrium has been attained. Even when these assumptions are satisfied, inference about migration may depend on which populations are sampled. Beerli (2004) introduced the term 'ghost population' for a population that represents the collective effect of unsampled populations on estimates of migration rates among populations sampled. In the Beerli & Felsenstein (2001) program MIGRATE, it is possible to allow for a ghost population when estimating migration rates among populations sampled. Here I will use the term more generally to mean any population that is not sampled but is connected by migration to populations that are sampled. The goal of this paper is to quantify the effect of ghost populations on migration rates inferred among populations that are sampled. It will be clear that, even under the highly idealized conditions considered in this paper, the effect of ghost populations is far from simple.

Correspondence: Montgomery Slatkin, Fax: 510-643-6264; E-mail: slatkin@socrates.berkeley.edu

Multipopulation coalescent model

Takahata (1988), Notohara (1990), Wilkinson-Herbots (1998) and others have developed the coalescent theory applicable to a set of local, randomly mating populations among which there is gene flow. For convenience, local populations will be referred to as demes and the collection of demes together will be called the population. The species is monoecious and the population comprises d demes with the ith deme containing N_i diploid individuals. There is random mating within each deme. The migration pattern among the demes is described by the backwards migration matrix **m**. The ijth element of **m** for $i \neq j$, m_{ij}, is the probability that a randomly chosen copy of a locus in deme i was derived from a copy in deme j in the previous generation. For later convenience, the diagonal elements of **m** will be denoted by $1 - m_{ii}$. Because every copy has to be in some deme in the previous generation.

$$m_{ii} = \sum_{j \neq i} m_{ij}$$

We assume that genetic material is obtained from individuals in a geographically dispersed species and that genotypes at numerous effectively unlinked genetic loci are determined. Differences in allelic states are determined by whatever means is appropriate for the kind of loci surveyed, e.g. SNPs, microsatellites, allozymes or RFLPs (restriction fragment length polymorphism). The coalescent (without recombination) is a Markov process that models the genetic state of the sample. The important idea in the coalescent is that the ancestry of the sample is analysed separately from the mutation process that determines genetic state (Hudson 1990). For convenience of discussion, I will refer to each copy of a locus in the sample as a gene. The term gene seems slightly preferable to allele because the latter term suggests identity of genetic state. Assume that $c(t)$ is a d-vector that describes the ancestry of the genes at a locus at time t in the past. The elements of $c(t)$, $c_i(t)$, are the numbers of ancestral lineages present in deme i at t, and the $c_i(0)$ are the numbers of genes sampled from each deme. Under the assumption that the N_i are large and m_{ij} are all small, the ancestry of the sample can, to a good approximation, be described by a continuous-time Markov chain for which the nonzero transition probabilities are as follows. The probability that $c_i(t)$ decreases by one and the others are unchanged in a time interval of length dt is,

$$\Pr(c_i \to c_i - 1) = \frac{c_i(c_i - 1)}{4N_i} dt$$

which represents the effect of a single coalescent event. The probability that $c_i(t)$ decreases by one and $c_j(t)$ increases by one (representing migration from deme j to deme i, $j \neq i$) is

$$\Pr(c_i \to c_i - 1, c_j \to c_j + 1) = c_i m_{ij} dt$$

These transition probabilities completely specify the coalescent process and allow us to solve in principle for the joint probability distribution,

$$p(c_1(t) \ldots, c_d(t))$$

given the initial conditions. In practice, an analytic solution cannot be found for samples larger than two copies, except in very simple cases (Takahata & Slatkin 1990). The process can be efficiently simulated, however, and it forms the basis for likelihood and Bayesian methods for estimating migration rates (Beerli & Felsenstein 1999; Bahlo & Griffiths 2000; Beerli & Felsenstein 2001; Nielsen & Wakeley 2001).

For understanding how ghost populations affect estimates of elements of the migration matrix, assume that samples are taken only from the first n of the d demes. That is, $c_i(0) = 0$ for $i > n$. The genetic state of the sample is determined by the gene genealogy of the sample and the locations of mutations on the branches (Hudson 1990). Mutations occur independently of geographical location so the likelihood of the data depends on the migration matrix through the distribution of gene genealogies generated by the coalescent process. To incorporate the effect of gene flow with ghost populations into an apparent migration matrix for the n deme sampled, it would have to be possible to define a matrix with elements $\tilde{m}_{ij}$ in such a way that gene genealogies generated by the coalescent process for the n demes sampled have the same probability distribution as gene genealogies from the true coalescent process for the d demes in the population. It is not possible, however, to define an apparent migration matrix which has this property because there are more states of the Markov chain for the coalescent process for the full population than for the subset of demes sampled. Although samples are taken only from n demes, ancestors of the copies sampled may be in any of the d demes, so the complete set of configurations must be accounted for, and that set of configurations requires the $d \times d$ migration matrix, not an $n \times n$ submatrix.

This conclusion does not mean that in particular cases, an apparent migration matrix cannot be found that approximately incorporates ghost populations, but there is no underlying principle that ensures such a matrix exists or that its elements are independent of both the sample size from each deme and the method used to estimate apparent migration rates. In other words, there is no underlying set of parameters ($\tilde{m}_{ij}$) to which estimates obtained from different methods would be expected to converge in the limit of large sample sizes.

Pairwise homozygosities

Relationship between migration matrix and pairwise coalescence times

There is a close relationship between expected coalescence times of pairs of genes and the homozygosity within and between demes (Slatkin 1991; Slatkin 1993). Let H_{ij} be the probability of identity in state of two genes, one drawn from deme i and the other from deme j, and let $P_{ij}(t)$ be the probability distribution of the coalescence times of those two genes. For the infinite alleles model of mutation (meaning that each mutant is new),

$$H_{ij} = \int_0^{\infty} P_{ij}(t)e^{-2\mu t}dt \tag{1}$$

where μ is the mutation rate (Hudson 1990). Equation 1 can be understood intuitively as meaning that two genes are identical in state if they coalesce before there is a mutation on either lineage. If the mutation rate is small, meaning that $1/\mu$ is much larger than any of the coalescence times in the model, the right hand side of Equation 1 can be expanded in a Taylor series to obtain,

$$H_{ij} \approx 1 - 2\mu\int_0^{\infty} tP_{ij}(t)dt = 1 - 2\mu\bar{t}_{ij} \tag{2}$$

to order μ (Slatkin 1991), where $\bar{t}_{ij}$ is the average coalescence time between genes drawn from demes i and j. Note that $\bar{t}_{ij} = \bar{t}_{ji}$ because it does not matter which deme is regarded as i or j. Although the H_{ij} can estimate F_{ST} for each pair of populations (Slatkin 1993), the analysis here is easier if the H_{ij} themselves are used.

There is a further simplification if the migration matrix is symmetric, $m_{ij} = m_{ji}$, and if it is aperiodic and irreducible. Under these restrictions,

$$\bar{t}_{ii} = 4N_T$$

independently of the migration matrix, where,

$$N_T = \sum_{i=1}^{d} N_i$$

is the total number of individuals in the population (Strobeck 1987; Hey 1991). A further simplification is obtained by noting that the coalescence of two genes sampled from different demes occurs in two phases (Slatkin 1991). The first phase is between the time the sample is taken and the first time in the past the ancestral lineages are in the same deme, and the second phase is before the ancestral lineages are first in the same deme. The expected length of the second phase is $4N_T$ independently of the migration matrix, and the expected length of the first phase, denoted by u_{ij}, depends on the migration matrix. Therefore,

$$\bar{t}_{ij} = 4N_T + u_{ij}. \tag{3}$$

The u_{ij} can be found by applying the standard theory of Markov chains, summarized by Ewens (2004). The first entry of the two ancestral lineages into the same deme is defined to be absorption, and the expected times to absorption can be found by solving the system of $d(d-1)/2$ linear equations,

$$(m_{ii} + m_{jj})u_{ij} - \sum_{k\neq i,j}^{d} (m_{ik}u_{kj} + m_{jk}u_{ik}) = 1 \tag{4}$$

for $n \geq i > j \geq 1$, where the sum is over all $k = 1, \ldots, d$, except for $k = i$ and $k = j$. Equation 4 is an approximation that assumes all the off-diagonal elements of the migration matrix are small, an assumption made in the full coalescent model as well.

This set of $d(d-1)/2$ equations is a linear system for $d(d-1)/2$ unknowns, the u_{ij}; the general theory of Markov processes ensures that there is always a unique solution with $u_{ij} > 0$. Therefore, given the backwards migration matrix, we have a straightforward way to predict the u_{ij} and hence the H_{ij} for all i and j.

Equation 4 can be also regarded as a set of linear equations for m_{ij} in terms of the u_{ij}. As such, they provide a way to estimate the m_{ij} from estimates of the H_{ij}. Equations 2 and 3 imply,

$$\frac{H_{ij} - H_0}{2\mu} = \hat{u}_{ij} \tag{5}$$

where $H_0 = H_{ii}$ is the within-deme homozygosity (which is the same for each deme under the assumptions of this model), μ is assumed to be known, and the carat (^) indicates estimated values of u_{ij}. By substituting the $\hat{u}_{ij}$ for u_{ij} in Equation 4, estimates of m_{ij} are obtained. The total population size, N_T, can also be estimated:

$$\hat{N}_T = \frac{1 - H_0}{4\mu} \tag{6}$$

This analysis does not mean that the H_{ij} provides the best way to estimate N_T and the m_{ij} from real data sets. They do not. Likelihood methods of kind developed by Bahlo & Griffiths (2000), Beerli & Felsenstein (2001) and Nielsen & Wakeley (2001) use all the data, not only within and between-population homozygosities, and therefore should provide better estimates. But if a very large amount of data were available and the model's assumptions were exactly satisfied, estimates based on maximum likelihood and the H_{ij} should converge to the true values of N_T and m_{ij}.

Because Equation 4 is a linear system of equations for the $d(d-1)/2$-values of m_{ij} as functions of u_{ij}, a unique solution will always exist provided that the determinant of the coefficient matrix is nonzero. But the theory of linear equations does not ensure that all $m_{ij} \geq 0$, which is necessary for biological reality. In fact, it is easy to choose values of u_{ij} for which Equation 4 implies that one or more of the m_{ij} are negative. For example, if $d = 3$, $u_{21} = u_{31} = 1000$ and $u_{32} = 2000$, then Equation 4 implies that $m_{21} = m_{31} = 0.001$ and $m_{32} = -0.00025$. A set of u_{ij} for which the solutions to Equation 4 are non-negative will be called feasible. A set for which one or more m_{ij} is negative will be called infeasible.

Apparent migration matrix

The preceding theory provides a way to define an apparent migration matrix among a subset of demes sampled. Assume that d is the true number of demes and m_{ij} $(i \neq j)$ are the true off-diagonal elements of the migration matrix. Then the solution to Equation 4 for u_{ij} provides set the 'true' absorption times that determine the expected values of the H_{ij}. Now assume that the number of demes sampled, n, is smaller than d. Because these n demes are part of the larger set of demes, the u_{ij} and hence the expected values of H_{ij} are determined by the true $d \times d$ migration matrix. However, they are in general not the solution to the corresponding system of $n(n-1)/2$ equations for the demes sampled. To be specific, assume that the first n of d demes are sampled, and that u_{ij} are the solutions to Equation 4 for the true migration matrix m_{ij}. The elements of the apparent migration matrix, the $\tilde{m}_{ij}$, among the demes sampled are the solutions to

$$(\tilde{m}_{ii} + \tilde{m}_{jj})u_{ij} - \sum_{k \neq i,j}^{n} (\tilde{m}_{ik}u_{jk} + \tilde{m}_{jk}u_{ik}) = 1 \qquad (7)$$

for $n \geq i > j \geq 1$. The apparent migration rates depend on the true migration rates and population sizes through the solutions to the two systems of Equations 4 and 7. There appears to be no simple relationship between m_{ij} and $\tilde{m}_{ij}$ and no intuitive way to characterize the effect of ghost populations.

We have already seen that a set of values of u_{ij} may be infeasible. It is possible, then, that the u_{ij} associated with the true migration matrix may be infeasible for the apparent migration matrix. That situation, however, appears to be very unlikely. I generated symmetric $d \times d$ migration matrices with off-diagonal elements having values randomly chosen within specified limits, computed u_{ij} from Equation 4 and then tested the feasibility of u_{ij} for every subset representing from $n = 3$ to $n = d - 1$ demes sampled. In running thousands of sets of replicates, no infeasible subsets of u_{ij} were found. It is possible to find a counter example, however, and one is presented in succeeding discussions. It appears to represent an unusual situation that arises only if migration is very restricted, as it is in a one-dimensional stepping-stone model. In most cases, an apparent migration matrix with non-negative elements can be computed for each subset of populations.

The within-population homozygosity is also affected by the presence of ghost populations. As Equation 6 shows, the total population size, N_T, rather than the number of individuals in the demes sampled is estimated by H_0. If we were attempting to use this method to estimate the average deme size, N, then our estimate would be too large by a factor d/n, because N would be estimated by N_T/n but N_T would actually be dN.

Examples

Three-deme models

To illustrate the preceding results, I consider several examples. The simplest nontrivial case is with $d = 3$ and $n = 2$. If $m_{31} = 0$, the three populations can be thought of as being in a line with deme 2 in the middle. In this case, Equation 4 reduces to a 3×3 system of equations, that can be written in matrix form as

$$\begin{pmatrix} 2m_{21} + m_{32} & -m_{32} & 0 \\ -m_{32} & m_{32} + m_{21} & -m_{21} \\ 0 & -m_{21} & 2m_{32} + m_{21} \end{pmatrix} \begin{pmatrix} u_{21} \\ u_{31} \\ u_{32} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \qquad (8)$$

and that has the solution

$$u_{21} = \frac{5m_{21} + 4m_{32}}{6m_{21}(m_{21} + m_{32})}$$

$$u_{31} = \frac{4m_{21}^2 + 7m_{21}m_{32} + 4m_{32}^2}{6m_{21}m_{32}(m_{21} + m_{32})} \qquad (9)$$

$$u_{32} = \frac{5m_{32} + 4m_{21}}{6m_{32}(m_{21} + m_{32})}$$

If only two of the three demes are sampled, then Equation 7 implies

$$\tilde{m}_{ij} = \frac{1}{2u_{ij}} \qquad (10)$$

for whichever two demes are sampled. For example, if $m_{21} = m$ and $m_{32} = wm$, then if $w = 5$,

$\tilde{m}_{21} = 36m/50 = 0.72m$, $\tilde{m}_{31} = 90m/139 \approx 0.65m$, and
$\tilde{m}_{32} = 90m/29 \approx 3.1m$

A second three-deme model corresponds to a case examined by Beerli (2004). He considered several models in which the demes sampled exchanged migrants with one or more unsampled demes. In his scenarios D and E, there is symmetric migration between the two demes sampled and a third deme denoted 'world'. In my notation, both scenarios can be parameterized as $m_{21} = m$ and $m_{31} = m_{32} = wm$, where demes 1 and 2 are the ones sampled and deme 3 is the world. Beerli's scenario D corresponds to $w = 1$ and his scenario E corresponds to $w = 10$. Specializing the general model to this case implies,

$u_{21} = 5/[2m(3 + 2w)]$

and hence,

$$\tilde{m}_{21} = \frac{m(3 + 2w)}{5} \qquad (11)$$

Therefore, $\tilde{m}_{21} = m$ for $w = 1$ and $\tilde{m}_{21} = 4.6m$ for $w = 10$. Beerli's (2004) Fig. 4, parts B1 and B2, shows that for scenario D, MIGRATE infers the correct migration rate between demes 1 and 2, which is consistent with the result here. For scenario E, MIGRATE infers that the migration rate is roughly 2.5 times the correct rate, which is slightly more than half of the result predicted here.

The reason for the difference between the apparent migration rates computed by the two methods is that MIGRATE also estimates the deme size scaled by the mutation rate, $\Theta = 4N\mu$. Beerli's (2004), Fig. 3, parts B1 and B2 shows that estimates of Θ are roughly correct for scenario D and about 50% too large for scenario E. The present method behaves differently. As discussed above, H_0 would estimate the apparent deme size to be $3N/2$. The effect of ghost populations on estimates of the apparent Θ and m_{21} obtained from MIGRATE depends on the extent of immigration from the ghost population. With $w = 1$, there is no effect on the estimates of Θ or m_{21}. With strong gene flow with the ghost population, $w = 10$, the results are comparable to those here: Θ is increased by 50% and m_{21} is increased by about a factor of 2.5, which is close to $(2/3)4.6 = 3.07$, the value obtained with the method developed here when the effect of the ghost population on the estimate of deme size is taken account of.

This example illustrates the fact that different methods for estimating apparent migration rates give different answers. This difference results from the fact that MIGRATE uses the full coalescent process, in which the rate of coalescent events within populations is proportional to $c_i(c_i - 1)$. With a large sample from each deme, coalescent events are occurring so frequently that immigration from the ghost population has little effect unless the immigration rate is large. In Beerli's scenario D $4Nm_{31} = 4Nm_{32} = 1$, so the within-deme coalescent process dominates and correct estimates are obtained. In scenario E, $4Nm_{31} = 4Nm_{32} = 10$, so immigration from the ghost population dominates and the ghost population affects the estimate of both N and m_{21} roughly to the same extent as predicted by the theory presented here. This kind of intuitive argument can help explain results from MIGRATE and similar programs but cannot substitute for a thorough analysis of a particular data set.

Stepping-stone models

The linear stepping stone model is one used in many theoretical studies as an extreme case in restricted migration. Assume that the d demes are arranged in a circle and that migration is only between adjacent demes at rate $m/2$ per generation. The pairwise coalescence times depend on k, the number of steps separating two demes,

$$u_k = \frac{(d-k)k}{2m} \tag{12}$$

(Slatkin 1991). If d is an integer that is a multiple of n and the demes sampled are evenly spaced (i.e. $d/n - 1$ demes are unsampled between each pair of demes sampled), we can find the apparent migration rate without using the general formulation. Between adjacent demes sampled, Equation 12 implies

$$u = \frac{\left(d - \frac{d}{n}\right)\frac{d}{n}}{2m} = \frac{(n-1)\left(\frac{d}{n}\right)^2}{2m} \tag{13}$$

Therefore, for adjacent sampled demes, Equation 13 reduces to Equation 12 if n replaces d and $\tilde{m} = mn^2/d^2$ replaces m. The apparent migration rate is reduced by the square of fraction of demes sampled.

Because of the symmetry of the model and the choice of demes sampled, the sampling scheme does induce apparent migration among sampled demes that are not apparently adjacent. That can be shown by demonstrating that Equation 7 implies $\tilde{m}_{ij} = 0$ unless $|i - j| \leq 1$. In other words, the geometric structure of migration is preserved, although its apparent magnitude is altered. That is not true in general, however. Consider the same model but assume that the demes sampled are not evenly spaced. For example, suppose that $m = 0.01$, $d = 20$, $n = 4$, and the four demes sampled are 1, 3, 5 and 17. The off-diagonal elements of the apparent migration matrix are

$$\tilde{m} = \begin{pmatrix} 0.000659 & -0.000048 & 0.000247 \\ & 0.000558 & -6.56 \times 10^{-7} \\ & & 0.000036 \end{pmatrix} \tag{14}$$

The negative values are not the result of rounding error or of using such a large value of m that the linear approximation on which Equation 7 is based is not valid. Reducing m by one or two orders of magnitude reduces the elements of $\tilde{m}$ proportionally. This result shows that a feasible set of u_{ij} can be infeasible for a subset of demes sampled. That appears to be the case for all choices of demes sampled from a linear stepping-stone model with $n > 2$ unless they are evenly spaced, although I could not find a general proof.

A two-dimensional stepping stone model is analysed in the same way. For simplicity, I will consider only a model on a $d_1 \times d_2$ 'torus' of populations and assume that d_1 and d_2 are even. Given that two genes are sampled from demes i steps apart in one direction and j steps apart in the other, the expected time until the ancestors of those genes are in the same population is,

$$u_{ij} = \sum_{l=0}^{d_2/2} \sum_{k=0}^{d_1/2} \frac{f_{kl}[1 - \cos(2\pi ik/d_1)\cos(2\pi jl/d_2)]}{\delta(k)\delta(l)(1 - f_{kl})} \tag{15}$$

where $\delta(k) = 1$ if $k = 0$ and $1/2$ if $k > 0$, $f_{00} = 0$ and,

$$f_{kl} = [1 - m(1 - \cos(2\pi k/d_1))]^2[1 - m(1 - \cos(2\pi l/d_2))]^2$$

otherwise. These expressions correct typographical errors in Equation 8 of Slatkin (1993). The calculations done in that paper were based on the correct equations.

Equation 15 can be used to find the average absorption times for a set of demes sampled and then Equation 7 can be used to find the elements of $\tilde{m}$. To illustrate, assume $d_1 = d_2 = 50$ and that five adjacent demes in a line are sampled. The symmetry of the model ensures that it does not matter which five. If $m = 0.01$, then the off-diagonal elements of the apparent migration matrix are,

$$\tilde{m} = 10^{-6}\begin{pmatrix} 13.7 & 3.2 & 1.0 & 1.9 \\ & 15.3 & 3.5 & 1.0 \\ & & 15.3 & 3.2 \\ & & & 13.7 \end{pmatrix}$$

Two features of this result are worth noting. First, the two-dimensional stepping-stone model creates the appearance of migration between nonadjacent populations. Second, the apparent level of migration between adjacent populations is more than two orders magnitude smaller than the actual level, $m = 2.5 \times 10^{-3}$ as compared to 1.37×10^{-5} and 1.53×10^{-5}. In other words, much lower levels of migration in a two-dimensional stepping-stone model result in the same degree of differentiation as in a one-dimensional model.

Discussion and conclusions

The problem of describing the effect of ghost populations on estimates of migration rates among sampled populations appears not to have a general solution in the sense that there is no definition of an apparent migration matrix among sampled populations that will predict all properties of genetic samples. The underlying coalescent process for the true population necessarily requires a larger configuration space than a coalescent model for only the subpopulations sampled. Consequently, it seems impossible in general to quantify the effect of ghost populations on estimated migration rates among populations sampled. Even for a particular population, different sample sizes and different methods of inference can lead to different estimates.

For a more restricted problem, that of symmetric migration and estimates of migration rates based on homozygosities within and between populations, it is almost always possible to find an apparent migration matrix that summarizes the effect of ghost populations. Cases in which an apparent migration matrix cannot be found appear to be exceptional and require that the true pattern of migration be quite restricted, as in the linear stepping-stone model.

The results in this paper are intended to clarify the theoretical relationship between the true pattern of migration among local populations and the apparent pattern among populations sampled. Even if the model's assumptions are satisfied, it is obviously not possible to work backwards and infer properties of the true migration matrix from the apparent migration matrix; even the true number of local populations (d) is unknown. Instead, the theory can be used to explore the relationship among models and to guide interpretation of results obtained from estimates of pairwise migration rates. The apparent migration matrix obtained using pairwise homozygosities probably represents an upper bound of the effect of immigration from unsampled populations because pairwise homozygosities give the most weight to migration relative to coalescence. As seen in the comparison with Beerli's (2004) results, apparent migration rates obtained using Beerli & Felsenstein's (2001) MIGRATE are less than predicted by considering pairwise homozygosities.

Acknowledgements

This research was supported in part by a grant from NIH (GM40282) and by the Miller Institute of Basic Research in Science. I thank P. Peerli and F. Rousset for helpful comments on an earlier version of this paper.

References

Bahlo M, Griffiths RC (2000) Inference from gene trees in a subdivided population. *Theoretical Population Biology*, **57**, 79–95.

Beerli P (2004) Effect of unsampled populations on the estimation of population sizes and migration rates between sampled populations. *Molecular Ecology*, **13**, 827–836.

Beerli P, Felsenstein J (1999) Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics*, **152**, 763–773.

Beerli P, Felsenstein J (2001) Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proceedings of the National Academy of Science of the USA*, **98**, 4563–4568.

Ewens WJ (2004) *Mathematical population genetics. I. Theoretical Introduction*. Springer, New York.

Hey J (1991) A multi-dimensional coelescent process applied to multi-allelic selection models and migration models. *Theoretical Population Biology*, **19**, 30–42.

Hudson RR (1990) Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology*, **7**, 1–44.

Neigel JE (1997) A comparison of alternative strategies for estimating gene flow from genetic markers. In: *Annual Review of Ecology and Systematics* (ed. Fautin DG), Vol. 28, pp. 105–128. Annual Reviews Inc, Palo Alto, California, USA.

Nielsen R, Wakeley J (2001) Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics*, **158**, 885–896.

Notohara M (1990) The coalescent and the genealogical process in geographically structured population. *Journal of Mathematical Biology*, **29**, 59–75.

Rousset F (1997) Genetic differentiation and estimation of gene flow from *F*-statistics under isolation by distance. *Genetics*, **145**, 1219–1228.

Slatkin M (1991) Inbreeding coefficients and coalescence times. *Genetical Research*, **58**, 167–176.

Slatkin M (1993) Isolation by distance in equilibrium and non-equilibrium populations. *Evolution*, **47**, 264–279.

Strobeck C (1987) Average number of nucleotide differences in a sample from a single subpopulation: a test for population subdivision. *Genetics*, **117**, 149–154.

Takahata N (1988) The coalescent in two partially isolated diffusion populations. *Genetical Research*, **52**, 213–222.

Takahata N, Slatkin M (1990) Genealogy of neutral genes in two partially isolated populations. *Theoretical Population Biology*, **38**, 331–350.

Wilkinson-Herbots HM (1998) Genealogy and subpopulation differentiation under various models of population structure. *Journal of Mathematical Biology*, **37**, 535–585.

The author is a Professor in the Department of Integrative Biology, University of California at Berkeley.