Detecting Isolation by Distance Using Phylogenies of Genes

Montgomery Slatkin* and Wayne P. Maddison*†

*Department of Zoology, University of California, Berkeley, California 94720, and †Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721

Manuscript received December 28, 1989
Accepted for publication May 30, 1990

ABSTRACT

We introduce a method for analyzing phylogenies of genes sampled from a geographically structured population. A parsimony method can be used to compute $s$, the minimum number of migration events between pairs of populations sampled, and the value of $s$ can be used to estimate the effective migration rate $M$, the value of $Nm$ in an island model with local populations of size $N$ and a migration rate $m$ that would yield the same value of $s$. Extensive simulations show that there is a simple relationship between $M$ and the geographic distance between pairs of samples in one- and two-dimensional models of isolation by distance. Both stepping-stone and lattice models were simulated. If two demes $k$ steps apart are sampled, then $\hat{s}$, the average value of $s$, is a function only of $k/(Nm)$ in a one-dimensional model and is a function only of $k/(Nm)^2$ in a two-dimensional model. Furthermore, $\log(\hat{M})$ is approximately a linear function of $\log(k)$. In a one-dimensional model, the regression coefficient is approximately $-1$ and in a two-dimensional model the regression coefficient is approximately $-0.5$. Using data from several locations, the regression of $\log(\hat{M})$ on $\log$(distance) may indicate whether there is isolation by distance in a population at equilibrium and may allow an estimate of the effective migration rate between adjacent sampling locations. Alternative methods for analyzing DNA sequence data from a geographically structured population are discussed. An application of our method to the data of R. L. Cann, M. Stoneking and A. C. Wilson on human mitochondrial DNA is presented.

DNA sequence data will provide detailed information about the genetic state of natural populations but at present there are few methods available for relating sequences of genes from different members of the same species to population genetic processes. In this paper, we will describe a method for using relationships among genes and the geographic locations from which those genes are sampled to learn about dispersal in geographically structured populations. The results presented in this paper are extensions of those in a previous paper (SLATKIN and MADISON 1989) in which we developed a method for estimating the average level of gene flow from knowledge of phylogenies of genes sampled from an island model of population structure. In this paper, we will show that the analysis of phylogenies of genes combined with geographic distances separating genes sampled can provide additional information about the pattern of gene flow in a geographically structured population.

CLADISTIC ANALYSIS OF GENE PHYLOGENIES

We assume that a sample of genes, i.e. nonrecombining segments of DNA, is taken from two or more geographic locations. We assume that there is no recombination among genes so our approach is applicable to mitochondrial DNA in plants and animals and to plastid DNA in plants. Because there is no recombination, the phylogeny of the genes sampled represents the history of their past coalescent events. We assume that there is sufficient variation among the genes sampled that their phylogeny can be inferred by one of the standard methods (reviewed by FELSENSTEIN 1988). We can then use a parsimony method for computing the minimum number of dispersal events, which we denote by $s$, that is consistent with the phylogeny of the genes and their geographic locations. This approach may not indicate the actual history of dispersal in the population but indicates the minimum amount of dispersal that must have occurred.

In our previous paper (SLATKIN and MADISON 1989) we assumed that all migration events are equivalent, in effect assuming an island model of population structure in which every location is equally accessible from every other location. We showed that the distribution of $s$ depends on the product $Nm$, where $N$ is the size of each deme in the island model and $m$ is the immigration rate. We found that for $Nm \geq 1$, the distribution of $s$, $p(s)$, is approximately normal with the mean, $\hat{s}$, dependent on $Nm$. Figure 1 shows the dependence of $\hat{s}$ on $Nm$ when samples are taken from two demes. Because $p(s)$ is approximately normal, the value of $s$ computed in a particular data set can be used to estimate $Nm$, and the variance of the distri-

Genetics 126: 249–260 (September, 1990)
bution provides confidence limits on that estimate.

The goal of our earlier paper was to provide an estimate of the average level of gene flow, \( Nm \), in an island model that had reached an equilibrium under gene flow and genetic drift. In this paper we will consider models of isolation by distance (Wright 1943), including both one- and two-dimensional stepping-stone models and models of a continuously distributed population, and show how \( s \) computed for pairs of populations depends on the geographic distance between locations that are sampled. Values of \( s \) computed for all pairs of locations sampled can indicate both the overall level of gene flow and the extent of gene flow between pairs of populations.

**SIMULATION MODEL**

The simulation model used here is similar to that used by Slatkin and Maddison (1989), with the main difference being that a variety of different population structures can be assumed. We assume that genes are sampled from two different geographic locations and that samples of equal size, \( n \), are drawn from each location. In all of our simulations, each replicate begins with the sample of genes and their geographic locations. The model then proceeds backwards in time. We assume that the genes sampled are selectively equivalent and use the coalescent theory of Kingman (1982a, b). This approach allows us to simulate only the direct ancestors of the \( 2n \) genes sampled. In each generation, there are two possible changes in the configuration of the genes: two (or more) genes may be descended from a common ancestor in the previous generation (a "coalescent event") or a gene (or more than one gene) may be descended from a gene in another location (a "dispersal event"). We considered two models of population structure, a stepping-stone model of a population divided into discrete demes and a lattice model that approximates a continuously distributed population.

**Stepping-stone model:** In the stepping-stone model, we used the theory of coalescents in a subdivided population to compute the probabilities of coalescent and dispersal events in each generation. We modeled a population comprising \( d \) demes each containing \( N \) haploid individuals. Following Takahata's (1988) simulation approach, we assumed that \( N \) is large enough and the migration rates are all small enough that only a single coalescent or dispersal event can occur in each generation. Takahata calls this the "diffusion limit" of the process. In a deme with \( i \) genes that are ancestors of the genes sampled, the probability that there is a coalescent event is approximately \( i(i-1)/2N \) and the probability that one of these genes is descended from an immigrant is approximately \( \nu m \), where \( m \) is the immigration rate to that deme.

In the stepping-stone model, we assumed that immigrants to each deme come only from adjacent demes. We considered both a one-dimensional model in which each deme, except the two end demes, had two neighbors and a two-dimensional model in which each internal deme had four neighbors. Demes were equally likely to receive immigrants from each adjacent deme. Hence in the one-dimensional model, the internal demes could receive their immigrants only from the two adjacent demes and the two end demes received their immigrants only from the demes next to them. In the two-dimensional model, demes at the edges could receive their immigrants only from each of the three adjacent demes and demes at the corners could receive immigrants from each of the two adjacent demes. Dispersal along diagonals was not permitted. We assumed all demes had the same immigration rate, \( m \). If an immigration event occurred, the immigrant was equally likely to have come from any of the demes that were possible sources of the immigrants.

**Lattice model:** The lattice model assumes that one haploid individual is located at each vertex of a two-dimensional lattice. The coordinates of each gene sampled were specified. Then in each generation each gene present that was ancestral to one of the genes sampled "chose" its ancestor by moving a number of lattice points that was determined by a random draw from the distribution of dispersal distances. We used discretized versions of an exponential and a gaussian distribution of dispersal distances with variance \( \sigma^2 \). We allowed for both reflecting boundaries, representing a flat surface, and periodic boundaries, representing a "torus." When two genes chose the same ancestor, a coalescent event occurred and the number of ancestral genes was reduced by one. We always assumed a number of lattice points far greater than the
Detecting Isolation by Distance

A

\[ \frac{k}{M} = 5 \]

\[ \frac{k}{M} = 3 \]

\[ \frac{k}{M} = 1 \]

\[ s \]

\[ k \]

0

2

4

6

8

10

0

2

4

6

8

10

\( k/M^2 = 1 \)

\( k/M^2 = 0.333 \)

\( k/M^2 = 0.03 \)

B

\[ \bar{s} \]

\[ k \]

0

2

4

6

8

10

0

2

4

6

8

10

\( \log(k) \)

\[ 10 \]

\[ 8 \]

\[ 6 \]

\[ 4 \]

\[ 2 \]

\[ 0 \]

\[ 0.0 \]

\[ 0.5 \]

\[ 1.0 \]

\[ 1.5 \]

FIGURE 2.—Simulation results from a stepping-stone model of population structure. In all cases, \( N = 10,000 \) and samples of 16 genes were drawn from two demes \( k \) steps apart. Each point represents the average of 100 replicates. In (A), a 50 \( \times \) 1 linear array of demes was assumed and the value of \( M = Nm \) was adjusted in each set of simulations so that the ratio \( k/M \) was constant. The demes sampled were symmetrically placed on either side of the center. The two curves for \( k/M = 1 \) show the results for independent sets of 100 replicates. In (B), a 20 \( \times \) 20 square array of demes was assumed and the value of \( M \) was adjusted so that the ratio \( k/M^2 \) was constant. The demes sampled were close to the center of one axis and symmetrically placed on either side of the center of the other axis.

number of genes sampled so the probability of more than two genes choosing the same ancestor was low. If however three genes did choose the same ancestor, the first two coalesced and the third simply occupied the same location.

To obtain results from the lattice model that were comparable to those from the stepping-stone model, we assumed that two quadrants were sampled. Each quadrant was regarded as a single sampling location and \( n \) genes were chosen from each quadrant randomly without replacement. These quadrants represented samples at distinct locations that are typically taken from populations that are regarded as being continuously distributed. The distance between two quadrats is the distance between the lattice points at their centers.

In both the stepping-stone and lattice models, we started each replicate simulation by assuming the genes sampled were in their initial locations and then ran the process backwards until a single ancestral gene was left. During each replicate we kept track of the phylogeny of the genes sampled. We then applied the parsimony method to compute \( s \) for that replicate using the algorithm described in our previous paper (SLATKIN and MADDISON 1989).

SIMULATION RESULTS

Stepping-stone model: In the stepping-stone model, we found that the distribution of \( s, p(s) \) depends on \( M = Nm \), the number of dimensions, and \( k \), the number of steps separating the two demes sampled. The first result of interest is that \( \bar{s} \) depends only on the ratio \( k/M \) in the one-dimensional stepping-stone model and on \( k/M^2 \) in the two-dimensional model, as shown in Figure 2.

The second result of interest is the dependence of \( \bar{s} \) on \( k \) for a fixed value of \( M \). As expected, \( \bar{s} \) decreases with increasing \( k \). The decrease with \( k \) differs between the one- and two-dimensional models, as shown in Figure 3. The value of \( \bar{s} \) itself is not of as much interest as the resulting estimate of \( Nm \) that is obtained using our results for an island model. For each of the sets of replicates, we used the value of \( \bar{s} \) we obtained and the results from an island model with the same sample sizes (Figure 1 above and Table 1; SLATKIN and MADDISON (1989)) to find the value of \( Nm \) for an island model.
model that would give the same value of \( \bar{s} \). This allows us to describe the results from stepping-stone models in terms of an equivalent \( Nm \) for the island model, \( \bar{M} \), as shown in Figure 4. Figure 4 shows that \( \log(\bar{M}) \) is approximately a linear function of \( \log(k) \). The slope is approximately \(-1\) in a one-dimensional model and \(-0.5\) in a two-dimensional model. The intercept depends on the actual value of \( \bar{M} \) and is close to \( \log(\bar{M}) \).

**Lattice model:** The results from the lattice model are very similar to those from the stepping-stone model. We found that there was no detectable difference between the results for the gaussian and exponential dispersal distributions so all the results we discuss were obtained using the exponential distribution. We also found that there was no detectable difference between the model with reflecting and periodic boundaries as long as the longer dimension of the lattice was greater than \( k \) by more than a factor of four. Figure 5 shows that \( \bar{s} \) depends on \( k/\sigma^2 \) in a compressed lattice (20 \( \times \) 1000) and on \( k/\sigma^4 \) in a square lattice (300 \( \times \) 300). These results are analogous to those shown in Figure 3.

As in the stepping-stone model, \( \log(\bar{M}) \) is approximately linear in \( \log(k) \) with a slope of approximately \(-1\) in the compressed lattice and \(-0.5\) in the square lattice as shown in Figure 6. The intercepts of these curves are approximately \( \log(\pi \sigma^2) \), which is the logarithm of Wright's (1943) neighborhood size for a haploid population.

**EXAMPLES**

**Simulated data:** To illustrate how our approach can be used, we will examine the results from single replicates. A single replicate represents a data set that might be obtained from a study in mtDNA. In the first two examples we assumed that nine locations were sampled. The question is whether we could use these samples to gain any insight into the structure of the populations from which these samples were taken. From the results shown in Figure 4, we would expect \( \log(\bar{M}) \) to decrease linearly with the logarithm of the geographic distance separating the samples. Furthermore, we would expect the intercept of the regression line (the value of \( a \)) to indicate the value of \( Nm \) or \( \pi \sigma^2 \), and the slope of the regression line (the value of \( b \)) to indicate whether gene flow was primarily in one dimension (\( b \approx -1.0 \)) or in two dimensions (\( b \approx -0.5 \)).

The analysis of a typical data set with samples from \( r \) locations proceeds as follows. First, a phylogeny of all genes in the sample is inferred, using a method such as parsimony or minimum evolution. Then values of \( s \), denoted by \( s_{ij} \) \((i, j = 1 \ldots r, i \neq j)\), between all

---

**TABLE 1**

Values of \( s_{ij} \) (above the diagonal) and \( \bar{M}_{ij} \) (below the diagonal) for single replicates of a stepping-stone model

<table>
<thead>
<tr>
<th>Sample</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Part A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>---</td>
<td>7</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>4.95</td>
<td>---</td>
<td>6</td>
<td>6</td>
<td>7</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>7.93</td>
<td>3.24</td>
<td>---</td>
<td>7</td>
<td>8</td>
<td>8</td>
<td>6</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>7.93</td>
<td>3.24</td>
<td>4.95</td>
<td>---</td>
<td>8</td>
<td>5</td>
<td>7</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td>7.93</td>
<td>4.95</td>
<td>7.93</td>
<td>4.95</td>
<td>---</td>
<td>7</td>
<td>7</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>6</td>
<td>4.95</td>
<td>1.99</td>
<td>7.93</td>
<td>1.99</td>
<td>4.95</td>
<td>---</td>
<td>6</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>4.95</td>
<td>1.99</td>
<td>3.24</td>
<td>4.95</td>
<td>4.95</td>
<td>3.24</td>
<td>---</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>4.95</td>
<td>1.99</td>
<td>7.93</td>
<td>3.24</td>
<td>7.93</td>
<td>3.24</td>
<td>4.95</td>
<td>---</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>4.95</td>
<td>1.99</td>
<td>4.95</td>
<td>3.24</td>
<td>7.93</td>
<td>4.95</td>
<td>4.95</td>
<td>7.93</td>
<td>---</td>
</tr>
</tbody>
</table>

Part B

<table>
<thead>
<tr>
<th>Sample</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>---</td>
<td>6</td>
<td>7</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>3.24</td>
<td>---</td>
<td>7</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>4.95</td>
<td>4.95</td>
<td>---</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>1.30</td>
<td>1.30</td>
<td>1.99</td>
<td>---</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>0.75</td>
<td>0.75</td>
<td>0.75</td>
<td>1.99</td>
<td>---</td>
<td>6</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>0.75</td>
<td>0.75</td>
<td>0.75</td>
<td>0.75</td>
<td>3.24</td>
<td>---</td>
<td>6</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>7</td>
<td>0.35</td>
<td>0.35</td>
<td>0.35</td>
<td>0.35</td>
<td>1.30</td>
<td>1.30</td>
<td>---</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>8</td>
<td>0.35</td>
<td>0.35</td>
<td>0.35</td>
<td>0.35</td>
<td>1.30</td>
<td>0.75</td>
<td>4.95</td>
<td>---</td>
<td>7</td>
</tr>
<tr>
<td>9</td>
<td>0.35</td>
<td>0.35</td>
<td>0.35</td>
<td>0.35</td>
<td>1.30</td>
<td>0.75</td>
<td>3.24</td>
<td>4.95</td>
<td>---</td>
</tr>
</tbody>
</table>

The values of \( \bar{M} \) were computed by interpolating values in Table 1 of SLATKIN and MADISON (1989) for the appropriate sample size. In both parts, \( N = 10,000, m = 0.001 \) (\( Nm = 10 \)), and samples of 16 genes were taken from 9 demes. In the simulations, the demes were labeled with a pair of indices \((i, j)\). In Part A, the demes were in a 21 \( \times \) 21 array. The upper left hand deme was \((0,0)\); the lower right hand deme was \((20,20)\). The demes sampled were in a square array with 3 demes separating the nearest demes sampled. The identities of the demes sampled are as follows: 1 (6, 6); 2 (10, 6); 3 (14, 6); 4 (6, 10); 5 (10, 10); 6 (14, 10), 7 (6, 14); 8 (10, 14); 9 (14, 14). In Part B, the demes were in a 50 \( \times \) 1 linear array with the demes numbered in order beginning with 0. The demes sampled were in evenly spaced with three demes separating adjacent demes samples. The identities of the demes are as follows: 1 (7); 2 (11); 3 (15); 4 (19); 5 (23); 6 (27); 7 (31); 8 (35); 9 (39).
pairs of sampling locations are found by focusing on

genes sampled from each pair of locations and work-
ing backward down the phylogeny. Finally, values of
\( \hat{M} \) are obtained from \( s_j \) by using our results for an
island model (SLATKIN and MADISON 1989). If the
sample sizes are the same in each location, then a
single table can be used for converting the integer
values of \( s \) into values of \( \hat{M} \). It is possible to compare
the resulting values of \( \hat{M} \) with exceptions under dif-
ferent models of isolation by distance that would
represent different possible pathways of dispersal.

Table 1, part A, shows values of \( s_j \) and \( \hat{M} \) obtained

from a single replicate when samples are taken from a
3 \times 3 square array of locations in a 21 \times 21 stepping-
stone model. In Figure 7A, we plotted the values of
\( \hat{M} \) against two different sets of assumed geographic
distances between the demes sampled. First we as-
umed the "correct" distances counting one unit of
distance between the nearest demes sampled, which
were in fact four steps apart. We will call this assump-
tion about the distances separating the sampling lo-
cations the "3 \times 3 distance matrix," for which there
were only five distance classes (1, \( \sqrt{2} \), 2, \( \sqrt{5} \), and \( \sqrt{8} \)).
Here, we are using Euclidean distance between pairs
of locations sampled. As an alternative, we considered
the possibility that the samples were in a linear array.
We chose a particular ordering (1, 2, 3, 6, 5, 4, 7, 8, 9) for the
samples that would represent a possible pathway of dispersal that was restricted by a ge-
ographic barrier. We will call this assumption about
distances separating sampling locations the "9 \times 1
distance matrix," for which there are 8 distance classes
(1, \ldots, 8). For each of the two distance matrices, we
computed the averages of \( \hat{M} \) for all pairs of locations
in the same distance class. The results are shown in
Figure 7A.

The slope of the regression line for the 3 \times 3
distance matrix is \( b = -0.45 \) and intercept is \( a = 0.78 \).
The value of \( b \) is consistent with two-dimensional gene
flow and not one dimensional gene flow. The value of \( a \) indicates that \( Nm \approx 10^{0.78} = 6.02 \), which is less
than the actual rate of gene flow between adjacent
demes (\( Nm = 10 \)) but does represent the gene flow
between demes separated by 4 steps. In contrast, the
9 \times 1 distance matrix fits much less well. The results
are not convincingly linear and the slope of the regression line, \( b = -0.14 \), is not in agreement with a model of isolation by distance.

To determine the statistical power of our method, we carried out the above comparison of the \( 3 \times 3 \) and the \( 9 \times 1 \) distance matrices in 100 replicates from the same program. Each replicate was analyzed as in the above example and yielded two values of \( b \), one for each distance matrix. Figure 8A shows the distributions of these values. We conclude that the results presented in Figure 7A are somewhat fortuitous. It does seem likely that the \( 9 \times 1 \) distance matrix would be rejected if the data are actually from a two-dimensional population, but the extent of variation in values of \( b \) for the \( 3 \times 3 \) distance matrix is quite large for these sample sizes. The problem is that there are too few distance classes to provide much power to detect isolation by distance.

We considered another example in part B, Table 1,
Detecting Isolation by Distance

Figure 8.—Regression coefficients for replicate simulations of different models of gene flow and different hypotheses about distances between sampling locations. In all parts 16 individuals from each of 9 locations were sampled: the $3 \times 3$ distance matrix indicates the hypothesis that populations sampled are in a square array as in Table 1 and the $9 \times 1$ distance matrix indicates the hypothesis that populations sampled are in a linear array as in Table 1. (A) Results from a $21 \times 21$ stepping-stone model of population structure with samples from 9 locations in a square with a spacing of 4 between the nearest locations sampled. These simulations are the same as those for Table 1, part A. (B) Results for a $51 \times 1$ stepping-stone model. Samples were taken from nine demes separated by 4 steps each. These simulations are the same as those in Table 1, part B. (C) Results for a $21 \times 21$ stepping stone model. Samples were taken from 9 demes in a line separated by one step. (D) Results for an island model with 100 demes and samples from 9 demes chosen arbitrarily.

Table 2
Results from analyzing the data of CANN, STONEKING and WILSON (1988) as described in the text

<table>
<thead>
<tr>
<th></th>
<th>African</th>
<th>European</th>
<th>Asian</th>
<th>New Guinean</th>
<th>Australian</th>
</tr>
</thead>
<tbody>
<tr>
<td>African (20)</td>
<td>—</td>
<td>9</td>
<td>7</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>European (45)</td>
<td>4.3/2790</td>
<td>—</td>
<td>14</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>Asian (34)</td>
<td>2.7/7540</td>
<td>7.9/4750</td>
<td>—</td>
<td>5</td>
<td>9</td>
</tr>
<tr>
<td>New Guinea (17)</td>
<td>1.9/10590</td>
<td>1.2/7800</td>
<td>1.4/3050</td>
<td>—</td>
<td>4</td>
</tr>
<tr>
<td>Australian (20)</td>
<td>2.3/14050</td>
<td>3.2/11280</td>
<td>4.3/6510</td>
<td>1.2/3460</td>
<td>—</td>
</tr>
</tbody>
</table>

Pairwise values of $s$ are above the diagonal and values of both $M$ and geographic distance in miles are below the diagonal. Sample sizes are in parentheses.

which presents values of $s$, and $\bar{M}$ for nine samples from a one-dimensional population. Now the $9 \times 1$ distance matrix is correct and the $3 \times 3$ distance matrix is not. Plotting the results in the same way as in Figure 7A, we can see that the $9 \times 1$ distance matrix provides a better fit to the data than does the
3 × 3 distance matrix. Using the 9 × 1 distance matrix, the regression coefficients are \( a = 0.56 \) and \( b = -1.21 \) in rough agreement with expectations. The estimate of \( Nm \) using \( a \) is \( 10^{0.56} = 3.63 \) which is approximately correct for the spacing of samples we have assumed. For this example, we also carried out 1000 replicate simulations to find the distribution of \( b \) values under the two hypotheses about gene flow. The results are shown in Figure 8B. In this case, the results of the 9 × 1 distance matrix are clustered around \(-1\), making values of \( b \) as large as \(-0.5\) significantly different. In Figure 8B, we did not plot the results from the 3 × 3 model because they were too variable and would not fit on the same scale.

We also considered one other set of replicate simulations to illustrate that more information about isolation by distance is obtained if there are more distance classes. We simulated a stepping-stone model with 441 demes in a square array, as in Table 1, part A, but sampled 16 individuals from nine demes in a line. Figure 8C shows the distribution of \( b \) values we obtained. There is less variation than we found in Figure 8A, which represents different samples from the same population.

In any application of our method, we can consider the alternative that there is no isolation by distance, either because gene flow occurs between locations independently of distance, as in the island model, or because the population is not at a genetic equilibrium. It is tempting to use a significance test for regression coefficients to determine whether the values of \( b \) obtained differ significantly from zero. That cannot be done, however, because the different pairs of values of \( \hat{M} \) and \( \hat{k} \) are not independent. Instead, the only way we see to determine significance of the regression coefficients for a particular set of samples is to carry out a simulation study for samples of the same size from an island model. Figure 8D shows the results of one such set of 1000 simulations of an island model for the sample sizes used in Table 1 (16 individuals from each of 9 locations). Each value replicate was subject to the same analysis as is shown in Table 1 for both the 3 × 3 distance matrix and the 9 × 1 distance matrix. In 95% of the outcomes, the \( b \) values were between \(-0.7 \) and 0.7 for the 3 × 3 distance matrix and \(-0.4 \) and 0.4 for the 9 × 1 distance matrix.

We conclude from these examples that detecting isolation by distance is possible but that samples from a relatively large number of locations are needed to have much accuracy. It is preferable that samples span a wide range of geographic distances. With 9 samples of 16 individuals each, it is likely that a two-dimensional pattern of isolation by distance could be detected and almost certain that a one-dimensional pattern could be detected. In the absence of an analytic theory, it is necessary to carry out a simulation study tailored to a particular data set to determine exactly what conclusions can be drawn.

Human mitochondrial DNA: The data of CANN, STONEKING and WILSON (1987) provide an opportunity to apply our method. CANN, STONEKING and WILSON (1987) examined mtDNA from 146 humans from five racial groups, African, European, Asian, New Guinean, and Australian. CANN, STONEKING and WILSON constructed a cladogram of all 146 individuals using a parsimony criterion. We used their cladogram and their assignment of racial identities to compute pairwise \( s \) values between races. The results are shown in Table 2. These \( s \) values show that the data show evidence of geographic restriction. They are all too small to be consistent with sampling from a panmictic population (W. P. MADISON and M. SLATKIN, unpublished data).

The analysis of this data set illustrates a problem that is common in the analysis of mtDNA data. In published data sets at least, there is usually insufficient resolving power of the biochemical methods used to allow the construction of a completely bifurcating (or dichotomous) tree of all individuals sampled. Instead, there are multifurcations (or polytomies) that cannot be resolved without additional information. To find each value of \( s \), a particular resolution of the multifurcations must be assumed. In our previous paper (SLATKIN and MADISON 1989), we suggested that the resolution should be chosen to minimize the value of \( s \) because that resolution will result in the smallest number of migration events consistent with the phylogeny. That is what we did in computing the values of \( s \) in Table 2. It is worth noting, however, that computing \( s \) in this way between all pairs of sampling locations (in this example, races) may lead to different resolutions of the same multifurcation when different pairs of locations are considered. Other ways of resolving multifurcations are possible and they will in-

![Figure 9](image-url)
of locations including the Middle East. We assumed (from west Africa and assigned them to Lagos, Nigeria be between the demes sampled if demes black Americans, we assumed those individuals came they all came from Beirut (34°N, 35°E). European individuals came from a variety of areas in east and southeast Asia including the Philippines. We assumed they all came from Hong Kong (22°N, 114°E). We assumed that all individuals from New Guinea came from a location roughly in the center of the eastern part of the island (8°N, 147°E) and that all Australian individuals came from a location roughly in the center of the northern part of Australia (24°N, 134°E). We are aware that the locations assigned to these samples are approximate and somewhat arbitrary. We present them for illustrative purposes only. We estimated the distances between these locations by finding the minimum (great circle) distance between nearest neighbors and assumed that the dispersal between other pairs followed the pathways between nearest neighbors. That choice of distance seemed preferable to using the great circle distance which would give a smaller distance between Africa and Australia than between Africa and New Guinea. Distances between nearest neighbors were obtained using the “Map” utility distributed by Apple Computer for use on a Macintosh. The pairwise geographic distances computed in this way are also shown in Table 2.

To determine whether there is a decrease of $\mu$ with distance, we estimated geographic distances between these races. Because the African samples were from black Americans, we assumed those individuals came from west Africa and assigned them to Lagos, Nigeria (6°N, 3°E). European individuals came from a variety of locations including the Middle East. We assumed they all came from Beirut (34°N, 35°E). Asian individuals came from a variety of areas in east and

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>1.414</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>2</td>
<td>1.414</td>
<td></td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>3</td>
<td>13</td>
<td>13</td>
<td></td>
<td>1.414</td>
</tr>
<tr>
<td>4</td>
<td>13</td>
<td>13</td>
<td>1.414</td>
<td></td>
</tr>
</tbody>
</table>

Matrix A contains the approximate distances between pairs of demes sampled and matrix B contains the approximate distances that would be between the demes sampled if demes 2 and 3 were reversed. See the legend of Figure 8 for details of the simulations.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>1.414</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>2</td>
<td>1.414</td>
<td></td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>3</td>
<td>13</td>
<td>13</td>
<td></td>
<td>1.414</td>
</tr>
<tr>
<td>4</td>
<td>13</td>
<td>13</td>
<td>1.414</td>
<td></td>
</tr>
</tbody>
</table>

Values of $P_j$ for the generalized cladistic method described in the text

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>1.414</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>2</td>
<td>1.414</td>
<td></td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>3</td>
<td>13</td>
<td>13</td>
<td></td>
<td>1.414</td>
</tr>
<tr>
<td>4</td>
<td>13</td>
<td>13</td>
<td>1.414</td>
<td></td>
</tr>
</tbody>
</table>

Average divergence times of pairs of genes sampled from demes $i$ and $j$

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>454 (354)</td>
<td>457 (351)</td>
<td>658 (300)</td>
<td>602 (338)</td>
</tr>
<tr>
<td>2</td>
<td>497 (348)</td>
<td>627 (312)</td>
<td>589 (335)</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>353 (293)</td>
<td>417 (308)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>419 (372)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

These values are in terms of $N = 10,000$ generations and were obtained from the same simulations that produced the values plotted in Figure 8. The averages were taken over all pairs of different genes sampled from pairs of demes sampled. The numbers in parentheses are standard deviations computed from the mean squared deviations of the divergence times.
TABLE 5
Average coalescent times of genes in samples $i$ and $j$ in the simulations described in Table 1

<table>
<thead>
<tr>
<th>$i$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Part A (samples in a square)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>196</td>
<td>205</td>
<td>231</td>
<td>237</td>
<td>313</td>
<td>248</td>
<td>225</td>
<td>233</td>
<td>226</td>
</tr>
<tr>
<td>2</td>
<td>208</td>
<td>229</td>
<td>237</td>
<td>298</td>
<td>251</td>
<td>230</td>
<td>233</td>
<td>229</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>250</td>
<td>245</td>
<td>305</td>
<td>260</td>
<td>241</td>
<td>245</td>
<td>245</td>
<td>242</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>245</td>
<td>301</td>
<td>277</td>
<td>237</td>
<td>236</td>
<td>237</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td>321</td>
<td>310</td>
<td>299</td>
<td>302</td>
<td>300</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td>271</td>
<td>261</td>
<td>274</td>
<td>265</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>231</td>
<td>227</td>
<td>222</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>235</td>
<td>224</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>233</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Part B (samples in a line) |    |    |    |    |    |    |    |    |    |
| 1   | 42 | 117| 91 | 146| 165| 164| 167| 167| 167|
| 2   | 115| 107| 119| 124| 124| 127| 125| 126|    |
| 3   | 101| 139| 153| 149| 157| 157| 157|    |    |
| 4   | 76 | 60 | 71 | 57 | 53 | 53 |    |    |    |
| 5   | 38 | 48 | 29 | 26 | 26 |    |    |    |    |
| 6   | 58 | 41 | 41 | 41 | 41 |    |    |    |    |
| 7   | 21 | 18 | 18 | 18 | 18 |    |    |    |    |
| 8   |    | 13 | 13 | 13 | 13 |    |    |    |    |
| 9   |    |    |    |    |    |    |    |    |    |

Times are measured in units of $1/N$ ($N = 10,000$).

only 0.1% of the variation in the data. There are of course many approximations that went into the analysis of these data but if there were an effect of isolation by distance it still might be detectable. Instead we are led to the conclusion that the similarity of human mtDNAs in different regions is not maintained by ongoing gene flow among local populations at approximate equilibrium which is of course consistent with our knowledge about the relatively recent dispersal of human races.

DISCUSSION

Our results show that if the assumptions of our model are satisfied, it is possible to detect restricted dispersal in a geographically structured population. The key assumptions are that the genes sampled are selectively equivalent and that there is no recombination. In our previous paper (SLATKIN and MADDISON 1989), we suggest that some kinds of selection might reduce the effective population sizes of demes, leading to an overestimate of the effectiveness of selection. We do not yet know the effect of recombination but are currently investigating that problem. Relatively large numbers of samples are needed to carry out the analysis we suggest. Ten or more randomly chosen individuals per location are necessary to provide sufficient within-location variation. Fewer than ten individuals per location would suggest that errors in the reconstruction of the phylogeny of genes would be likely to have large effects on the results. The number of locations that must be sampled depends on the hypothesis to be tested. There is more resolving power with more locations.

The method we have described above makes use of only part of the information in a sample of genes from different geographic locations. We have considered other ways to look at data that use somewhat different information and we have found that other statistics are less informative than pairwise values of $s$ for understanding population structure. We will discuss briefly two alternatives.

**Generalized cladistic methods:** When considering samples of genes from only two locations at a time, $s$ is the minimum number of migration events between those two locations. It is possible to consider samples from more than two locations together and compute the minimum number of migration events in the entire phylogeny of genes. We did this in our previous paper (SLATKIN and MADDISON 1989) under the assumption that migration events between all pairs of populations were equally difficult, which is equivalent to assuming an island model of migration. The goal in that paper was estimating the average level of gene flow among demes. It is possible to generalize that method by assuming values for a matrix that quantifies the relative difficulties of different dispersal events. We will call such a matrix $P_{ij}$ which is defined for $i \neq j$. In general, $P_{ij}$ does not need to be symmetric. The entries of $P_{ij}$ indicate how much weight is given to dispersal events between locations $i$ and $j$. The island model is represented by the values $P_{ij} = 1$. A linear stepping-stone model is represented by $P_{ij} = |i - j|$ if the locations are numbered in linear order. A two-
dimensional stepping-stone model could be represented by the $P_{ij}$ being either the linear distances between two locations or the Manhattan distances. More general patterns of gene flow could be represented by values of $P_{ij}$ that indicate distances between locations measured along pathways of possible gene flow.

Given a sample of genes and their phylogeny and given the matrix $P_{ij}$ it is possible to compute the minimum amount of migration that has occurred. The minimum amount of migration is computed using a parsimony criterion as described by Sankoff and Rousseau (1975) and Maddison and Maddison (1989). For each choice of $P_{ij}$ the result is a single number, which we will still call $s$. We could generalize the method of Slatkin and Maddison (1989) by assuming the $P_{ij}$ matrix is known and using it to calculate $s$ in order to estimate $N_m$. We will take a different approach here because our concern is with distinguishing among different patterns of gene flow, i.e., different $P_{ij}$ matrices. Different choices of $P_{ij}$ represent different hypotheses about gene flow among the locations sampled, and we would expect that choices of $P_{ij}$ that better represented the actual pathways would result in smaller values of $s$, if the $P_{ij}$ are standardized in some way. For example, if locations $i$ and $j$ are actually close together while locations $p$ and $q$ are distant, then an incorrect $P_{ij}$ matrix that weighs $i - j$ migration events heavily and $p - q$ migration events lightly would be expected to yield a high $s$ value, while a matrix using the correct weighting would give a smaller $s$ value. The question is whether this method for comparing different hypotheses about gene flow has any statistical power.

To illustrate this approach, assume we have samples from four locations. Table 3 shows the $P_{ij}$ matrices used to test these two hypotheses. They are standardized by making the sums of the elements the same. We used our simulation program to generate the histories of samples of 16 genes from each of the four populations and then computed the value of $s$ for both $P_{ij}$ matrices. By doing this for a large number of replicates, we could generate the distribution of $s$ under each hypothesis about gene flow. Graphs of the distributions are shown for each of two cases in Figure 10. These results are typical of what we found in other cases. The distributions overlap sufficiently that it would be difficult for a single value of $s$ computed from a mitochondrial data set, for example, to allow us to distinguish these two hypotheses with confidence. With large sample sizes from many more sampling locations, greater resolution is possible but the pairwise method we described above always seemed preferable.

**Branch lengths:** So far we have considered only the topology of the phylogeny of genes and not the branch lengths. The number of nucleotide differences between a pair of genes will indicate the time since they are descended from a common ancestral gene if it is assumed that mutation rates are constant. Strobeck (1987) and Slatkin (1987) suggested that the number of nucleotide differences between genes sampled from the same and from different locations could be used to estimate migration rates. In terms of phylogenies of genes, their method is equivalent to comparing branch lengths within and between sampling locations. In discussing branch lengths directly, we are ignoring the problem of estimating them from nucleotide differences.

In our simulations, we computed the average and the standard deviation of branch lengths separating genes sampled from the same and from different populations. The result for each set of replicates was a matrix of average branch lengths that could be compared with the known patterns and rates of gene flow.

There are three features of our results that make branch lengths poorly suited for inferring levels and patterns of gene flow. (i) In a subdivided population, branch lengths are usually very long. Strobeck (1987) and Slatkin (1987) showed that, in general, the average divergence time of a pair of genes sampled from the same deme is proportional to the total size of the population, not the deme size. The average divergence time of genes drawn from different demes is even longer. Some typical results for the stepping-stone model are shown in Table 4. (ii) The coefficients of variation of divergence times are large, typically greater than one half, as shown in Table 4. (iii) Variation in branch lengths between different pairs of demes is highly correlated, as illustrated in Table 5, which presents the average branch lengths within and between sampling locations in the simulations done for Table 1. The reason is that all pairs of genes that are on different sides of a deep branch will have the same divergence times. Therefore, even if there were no error in estimating divergence times from nucleotide differences, branch lengths do not seem well suited to estimating levels of gene flow. Any estimator that depended on branch lengths would be subject to too much intrinsic variability to be useful.

Average branch lengths do depend somewhat on distances between samples, as shown in Tables 4 and 5, so differences in branch lengths might be checked for consistency with results obtained using pairwise $s_{ij}$ values. Our conclusion about branch lengths should not be interpreted as meaning they cannot be useful, only that they appear not to be useful to estimate average levels of gene flow.

**CONCLUSIONS**

We have shown that our measure of genetic similarity between pairs of populations, $s$, leads to very
simple predictions in both the stepping-stone and lattice models of isolation by distance. If $s$ is used to compute $\hat{M}$, the equivalent value of $Nm$ in an island model, then $\hat{M}$ decreases approximately linearly with distance in a one-dimensional model and approximately with the square root of distance in a two-dimensional model. Pairwise values of $\hat{M}$ are in effect measures of genetic distance between populations that indicate the amount of gene flow between them.

Our results suggest that it may be possible to distinguish ongoing gene flow from historical association of populations as an explanation for observed genetic similarities. If geographic distance can be assumed to indicate the potential for gene flow between locations, then plotting $\log(\hat{M})$ against the logarithm of distance should yield a significant regression if gene flow is the cause of genetic similarity. On the other hand, if genetic similarity is caused by the fact that the populations observed are recently derived from an ancestral population, then the regression of $\log(\hat{M})$ against the logarithm of distance would not be expected to have such a simple form. For example, if all the populations in an area were the product of a rapid range expansion in the recent past, then values of $\hat{M}$ for all pairs of populations would be approximately the same. Patterns of historical association might easily mimic patterns of gene flow, but there at least is potential for distinguishing the two possibilities.

This research was supported in part by a U.S. National Institutes of Health research grant GM40282 to M.S., by the Miller Foundation for Basic Research and by a Natural Sciences and Engineering Research Council of Canada post-doctoral fellowship to W.P.M. We thank N. H. Barton, S. A. Frank, N. Takahata and two referees for helpful comments on an earlier version of this paper.

LITERATURE CITED


