The Coalescent in a Continuous, Finite, Linear Population

Jon F. Wilkins*,[1] and John Wakeley†

**Program in Biophysics and †Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138*

Manuscript received June 10, 2001
Accepted for publication March 4, 2002

ABSTRACT

In this article we present a model for analyzing patterns of genetic diversity in a continuous, finite, linear habitat with restricted gene flow. The distribution of coalescent times and locations is derived for a pair of sequences sampled from arbitrary locations along the habitat. The results for mean time to coalescence are compared to simulated data. As expected, mean time to common ancestry increases with the distance separating the two sequences. Additionally, this mean time is greater near the center of the habitat than near the ends. In the distant past, lineages that have not undergone coalescence are more likely to have been at opposite ends of the population range, whereas coalescent events in the distant past are biased toward the center. All of these effects are more pronounced when gene flow is more limited. The pattern of pairwise nucleotide differences predicted by the model is compared to data collected from sardine populations. The sardine data are used to illustrate how demographic parameters can be estimated using the model.

MIGRATION often plays an important role in shaping patterns of genetic diversity. Under conditions of restricted gene flow, the geographical and genetic structures of a population tend to become correlated. In the most basic terms, we expect individuals in close geographical proximity to be genetically more similar than geographically distant individuals. This differentiation will arise even in the absence of local adaptation, due to locally occurring genetic drift. As a result, it is possible to use existing patterns of neutral genetic variation to make inferences about the geographic structure of populations.

The best-studied models of geographic structure are the island (Wright 1931; Maruyama 1970a) and stepping-stone (Kimura and Weiss 1964) models. Both types of model assume a population composed of a number of subpopulations, or demes, connected to each other through migration. Each deme is assumed to be panmictic. In the island model, there is no explicit geography, in that each migration event occurs via a common migrant pool. Stepping-stone models, in contrast, permit migration only between neighboring demes. In the one-dimensional model, the demes are arrayed in a line, and each deme exchanges migrants only with the two adjacent demes. The analogous two-dimensional model assumes a grid of demes, with each deme exchanging migrants with some number of neighbors (*e.g.*, four).

Coalescent theory differs from classical population genetics in its focus on the time to the most recent common ancestor of two or more sequences, rather than on the properties of the population as a whole. This focus on the genealogical structure of a sample provides a framework in which properties of populations can be estimated. Coalescent theory applied to geographically structured populations with discrete demes has been formalized as the "structured coalescent" (see, *e.g.*, Wilkinson-Herbots 1998).

The coalescent model developed in this article assumes a population distributed uniformly along a finite, one-dimensional habitat. Gene flow is restricted, so locations of parents and offspring are correlated. A diffusion approximation is used to characterize the locations of ancestors of sampled sequences. Applied to pairs of sequences, this approach fully specifies the probability density for the times and locations of their most recent common ancestors and also provides summary statistics such as the mean time to coalescence.

The model is analogous to the one-dimensional stepping-stone model, but with some important differences that illustrate the motivation for this work. Although there has been a lot of work on finite stepping-stone models (discussed below), most analyses of the stepping-stone model rely on nonrealistic treatments of habitat boundaries. The best-studied models fall into two categories. Models in the first category assume that the ends of the array are joined together (the circular stepping-stone model, or the toroidal model in two dimensions; *e.g.*, Maruyama 1970b; Nagylaki 1974a, 1977; Strobeck 1987; Slatkin 1991). This assumption secures mathematical tractability by making all demes identical and migration isotropic (Strobeck 1987), but is directly applicable to few systems in nature (*e.g.*, a population inhabiting the entire coast of an island). Models in the second category assume a linear habitat of infinite

[1] *Corresponding author:* Harvard University, Biological Laboratories-2102, 16 Divinity Ave., Cambridge, MA 02138.
E-mail: jfwilkin@fas.harvard.edu

length (*e.g.*, Weiss and Kimura 1965; Nagylaki 1974b, 1976; Sawyer 1976, 1977). While these models provide useful insights regarding the short-term behavior of populations in a one-dimensional habitat, they predict infinite divergence between individuals (Sawyer 1976; Griffiths 1981; Wilkinson-Herbots 1998).

The model studied here assumes a finite linear habitat (*e.g.*, along a stretch of coastline). The analysis indicates that the expected pattern of genetic diversity does, in fact, depend on location in the habitat, suggesting that application of a circular model to a finite linear population is problematic. At the very least, this misapplication entails discarding information encoded in the variation in genetic diversity along the population range.

Models of isolation by distance in a continuous population date back to Wright (1943), who defined the effective neighborhood population size as the reciprocal of the probability of self-fertilization. That is, the neighborhood size is approximately the number of individuals within the single-generation dispersal range. Wright's work shows that the correlation between adjacent individuals and the differentiation between neighborhoods both increase as the neighborhood size becomes small compared with the total population size. However, much of the theoretical work since has focused on populations subdivided into discrete demes. While a discrete model may be appropriate for many organisms, others may be distributed more or less continuously across a particular range, but nevertheless be geographically structured due to limited gene flow. The model presented here assumes a continuous population, but can be applied in modified form to the discrete-demes model.

Work from within the classical population genetics paradigm provides some insight to the properties of finite linear models similar to the one considered here. Finite one-dimensional stepping-stone models have been analyzed by Maruyama (1970c), Fleming and Su (1974), and Malécot (1975). These analyses derive expectations for classical measures such as the covariance in gene frequencies across demes. Nagylaki and Barcilon (1988) have considered probabilities of identity in a semiinfinite linear habitat. Maruyama (1971) has also derived probabilities of identity for a continuous population on a torus and the rate of decrease in genetic variability in a finite two-dimensional population (Maruyama 1970d, 1972). Hey (1991) has compared the mean coalescence time for a pair of sequences sampled from opposite ends of a finite linear stepping stone with that of a pair sampled at random from the entire population. The result that coalescence times are longer near the center of the habitat range is consistent with findings of Herbots (1994, pp. 66 and 145–146), who found a similar pattern in linear stepping-stone models with three and five demes.

While all of these results are intimately linked to the distribution of coalescence times (Slatkin 1991), many classical population-genetic analyses do not make use of all of the information in DNA sequence data, making the coalescent approach presented here preferable. Furthermore, all of these analyses rely on approximations that assume a large local population size, an assumption that is relaxed in the present analysis. However, it is noteworthy that the results derived here are consistent with, and in some cases anticipated by, much of this previous work.

A model similar to the one presented here was proposed by Barton and Wilson (1995, 1996), who applied a coalescent approach to a continuous population in two dimensions, deriving recursion equations that describe the coalescent process for a pair of sequences. These equations agree closely with simulated distributions of coalescence times. However, the method becomes cumbersome for long coalescence times and does not readily lead to summary statistics such as the moments of the probability distribution. While limited to pairs of genes in a one-dimensional habitat, the model presented here is easily applied to both long and short coalescence times and yields summary statistics that can be used to make inferences regarding demographic history from genetic data. Simulation results confirm that the diffusion approximation used in this model provides an accurate characterization of the entire coalescent process under a broad range of parameter values.

THE MODEL

The model assumes a uniformly distributed population of N haploid individuals in a linear habitat, but can be applied to a population of $N/2$ diploid individuals without modification. Distance is scaled such that any location along the habitat is indexed by a number between 0 and 1 (with $\frac{1}{2}$ being the midpoint of the habitat). Absolute density-dependent population regulation is assumed. Each individual occupies a space of width $1/N$, from which all other individuals are excluded. The structure of the population is a one-dimensional lattice, as in the voter model (Holley and Liggett 1975), a contact-process model used in many ecological applications (Durrett and Levin 1994). The distribution of coalescence times is found using a continuous approximation. Another way to think of the model is as a stepping-stone model consisting of N demes, each of size 1.

Generations are nonoverlapping. Each individual produces a very large number of gametes, which are dispersed according to a normal distribution centered at the location of the individual and with variance $2\sigma_m^2$. Thus in each generation an effectively infinite number of gametes arrives at each location. One of these gametes is selected at random to become the adult at that location in the next generation. The distribution of the origins of those gametes is the correlation of the normal with a delta function at the individual's location.

Thus, the location of a parent is normally distributed around the location of its offspring, with variance $2\sigma_m^2$ (a grandparent is normally distributed around the same location with variance $4\sigma_m^2$, and so on).

The boundaries of the habitat are reflecting, so a gamete that would otherwise land outside the habitat range is reflected back an equal distance within it. Each individual thus has the same expected number of offspring regardless of its location. This means that migration is conservative, so migration alone is sufficient to maintain the relative population densities at all locations in the habitat (Nagylaki 1980). Nonreflecting boundaries would correspond to the case where those gametes dispersing outside the habitat range are lost. In such a system, individuals near the edges of the habitat would have a reduced effective fecundity relative to those nearer to the center.

As Felsenstein (1975) pointed out, most continuous-space models in population genetics assume a uniform population density that would not actually be maintained by the proposed reproductive scheme. A normal distribution of gametes without severe density regulation generates a population that is clumped together at certain locations and sparsely populated at others. With its absolute density regulation at all locations, the model of reproduction proposed here will immediately generate and maintain a population that is uniformly distributed across its habitat range.

Applying a coalescent approach to the analysis of this model involves tracking the location of the ancestors of a particular sequence back in time. The location of a single sampled lineage can be approximated using a diffusion process with diffusion constant σ_m^2. The precise location of a lineage is known only for the generation in which sampling occurred, so its location in the past is represented here as a probability distribution. Considering only a single sequence sampled from a particular location z^0 and its ancestors, one can imagine how this probability distribution broadens going back in time. In the distant past, the distribution becomes completely flat, when the ancestral sequence is equally likely to have been anywhere in the range. The exact distribution of ancestral locations can be derived for such a system using a Fourier series.

The analysis presented here derives the distribution of the time to coalescence for a pair of sequences drawn from locations z_1^0 and z_2^0 in the habitat. Each lineage is subject to a diffusion process backward in time, with the probability of coalescence related to the overlap of the two probability distributions (Figure 1). The solution employs a two-dimensional Fourier method, but is more complicated than the case of a single sequence because the ancestral location distributions $z_1(t)$ and $z_2(t)$, conditional on not yet having coalesced, are not independent of each other.

If we consider a single lineage, disregarding habitat boundaries for the moment, it is equally likely to have come from a location to its left or its right in the previous generation. It follows that when we consider two lineages at some distance from each other, they are equally likely to have been closer together or farther apart in the previous generation. However, if the two shared a common ancestor in the previous generation, they must have been closer together. Thus, conditional on not coalescing in the previous generation, the two lineages are slightly more likely to have been farther apart than closer together. In contrast to the single-sequence case, as time in the past becomes very large, the ancestral lineages are not equally likely to be found anywhere in the habitat range. If the two lineages are still distinct, they are likely to have been more geographically separated than a uniform distribution dictates.

The analysis involves a transformation of variables to create two new parameters that do diffuse independently. The first parameter encodes the distance between the two sequences, and the second their average position:

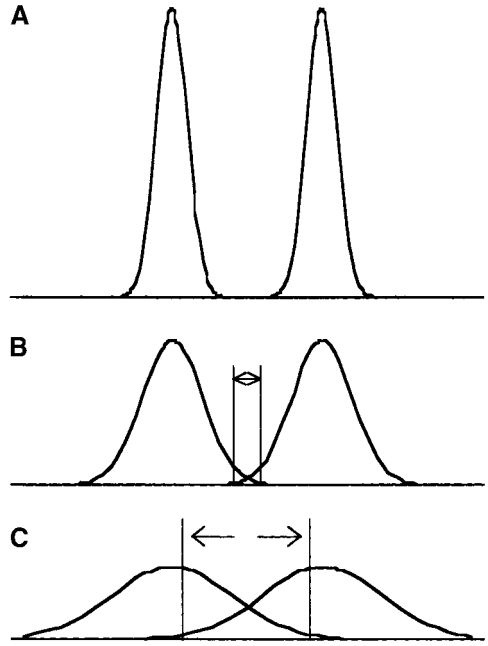

Figure 1.—The diffusion process for two lineages. The graphs illustrate the process of lineage diffusion backward in time. (A) The probability distribution for the locations of the two lineages at a time in the recent past. For sequences sampled from different locations, there is little overlap in the two distributions and therefore little possibility of a coalescent event. Going farther back in the past (B and C), the overlap between the two distributions initially increases, and the probability of the two sequences sharing a common ancestor increases. More recent coalescent events are most likely to occur near the center of the space separating the two samples (B). In the more distant past, the overlap between the two decreases and broadens, and the range over which coalescent events are likely to have occurred becomes less well defined (C).

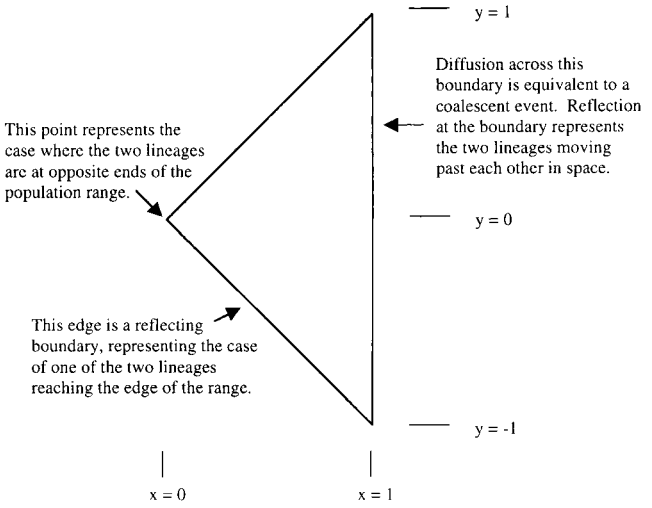

FIGURE 2.—The state space of the system in the transformed coordinates. Diffusion of the two lineages in the one-dimensional habitat is represented as the diffusion of a single point in a two-dimensional state space. The x coordinate encodes the distance between the two lineages, with $x = 0$ corresponding to the maximal separation of the two and $x = 1$ corresponding to the two lineages being at the same location. The y coordinate encodes the average position of the two lineages, with $y = -1$ corresponding to both lineages being at the end of the range where $z = 0$ and $y = 1$ corresponding to $z = 1$. The point $x = 0$, $y = 0$ is at the center of the square that is generated by the three reflections.

$$x = 1 - |z_1 - z_2| \tag{1}$$

$$y = z_1 + z_2 - 1. \tag{2}$$

The "1" terms in Equations 1 and 2 are included for mathematical convenience, and other coordinate systems would yield the same results, so long as diffusion is isotropic. The distribution of ancestral locations t generations in the past can now be represented as a two-dimensional probability surface in x and y:

$$U(x, y, t). \tag{3}$$

This probability distribution is nonzero within a right triangle ranging from -1 to 1 in y and from $|y|$ to 1 in x (Figure 2).

The genealogical history is now modeled as a single diffusion process in this triangular state space. The diffusion constant is $2\sigma_m^2$, twice that for the single-particle diffusion process discussed above. This factor of two arises from the fact that the new parameters are the sum and difference of two σ_m^2 single-particle diffusion processes.

Diffusion is subject to two different boundary conditions in the triangular state space. The two short sides of the triangle are reflecting boundaries. Reflection at these lines corresponds to reflection of a lineage off of a habitat boundary. The long side of the triangle is a more complex, partially reflecting, partially absorbing boundary. Positions along this line represent states where the two ancestral lineages are very close together in space. Reflection is equivalent to the two lineages moving past each other. Absorption is equivalent to a coalescent event.

The diffusion process in the transformed state space is isotropic, but not separable. Although diffusion in one dimension is independent of diffusion in the other dimension, it is not independent of location, due to the fact that the state space is triangular. The diagonal reflecting boundaries can be eliminated by placing a mirror image of the state space opposite the boundary. Reflection at the boundary is now represented as movement into the mirror-image state space. Three such reflections transform the state space into a square ranging from -1 to 1 in both x and y. All four edges are coalescent boundaries, and the symmetric shape makes the x and y diffusion processes separable as

$$U(x, y, t) = U_x(x, t)\, U_y(y, t), \tag{4}$$

where each distribution satisfies the diffusion equation

$$\frac{\partial U_x}{\partial t} = 2\sigma_m^2 \frac{\partial^2 U_x}{\partial x^2} \tag{5}$$

$$\frac{\partial U_y}{\partial t} = 2\sigma_m^2 \frac{\partial^2 U_y}{\partial y^2}. \tag{6}$$

Each pair of locations (z_1, z_2) in the real habitat corresponds to four locations in this square space, one in each of the four images of the triangular state space (see Equation A40).

The boundary conditions at the edges of the square depend on both the population density and the dispersal (migration) rate. If the population density and dispersal rate are very high, then when the two lineages come close together, they are likely to pass by each other rather than share a common ancestor, because there are a large number of individuals within their dispersal range. This corresponds to a more reflecting boundary in the two-dimensional state space. On the other hand, if the population density and dispersal rate are low, neighborhood size is small, and two lineages that are close together in space will be more likely to share a common ancestor, corresponding to a more absorbing boundary. Mathematically speaking, the flux rate of the probability distribution across the boundary is equal to the probability that the two lineages coalesce in the previous generation.

Because the model assumes perfect density-dependent population regulation, there is exactly one haploid lineage in each span of width $1/N$. A coalescent event occurs when the two ancestral lineages are found within the same $1/N$ span in a given generation. Note that giving the lineages a finite physical width addresses the concern raised by SAWYER (1976) that common ancestry in a continuous model requires two lineages to have a

physical separation of zero, which leads to pathological behaviors when applied to models of more than one dimension. In our continuous approximation of the population, we assume that a coalescent event occurs whenever the two lineages are separated by a distance of $<1/(2N)$, that is, when $1 - 1/(2N) < x < 1$. This approximates the probability that both lineages are found within the same fixed span of width $1/N$. Applying this criterion at the boundaries, the result is an infinite series of sine and cosine terms. The full solution is derived in the APPENDIX, and the main results are reproduced here in the text. For example, the joint probability density for the locations of the two lineages is given by

$$U_x(x, t) = \sum_{i=1}^{\infty} \frac{\alpha_i f_i(x_0)\cos(\alpha_i x)}{\alpha_i + \sin(\alpha_i)\cos(\alpha_i)} e^{-2\sigma_m^2 \alpha_i^2 (t-1/2)} + \sum_{j=1}^{\infty} \frac{\alpha_j^* f_j^*(x_0)\sin(\alpha_j^* x)}{\alpha_j^* - \sin(\alpha_j^*)\cos(\alpha_j^*)} e^{-2\sigma_m^2 \alpha_j^{*2} (t-1/2)} \quad (7)$$

$$U_y(y, t) = \sum_{i=1}^{\infty} \frac{\alpha_i f_i(y_0)\cos(\alpha_i y)}{\alpha_i + \sin(\alpha_i)\cos(\alpha_i)} e^{-2\sigma_m^2 \alpha_i^2 (t-1/2)} + \sum_{j=1}^{\infty} \frac{\alpha_j^* f_j^*(y_0)\sin(\alpha_j^* y)}{\alpha_j^* - \sin(\alpha_j^*)\cos(\alpha_j^*)} e^{-2\sigma_m^2 \alpha_j^{*2} (t-1/2)}. \quad (8)$$

These series can be truncated for purposes of making calculations without significant loss of accuracy. The α_i and α_j^* terms are determined by the boundary conditions (see Equations A13), and the f_i and f_j^* terms are determined by the two sampling locations (Equations A21). Discussion of Fourier-series solutions of the diffusion equation can be found in most texts (*e.g.*, CHURCHILL and BROWN 1987). The treatment of initial conditions used here is standard, and the boundary conditions are incorporated using the Sturm-Liouville method.

At time t, the probability of the two lineages not yet having coalesced is equal to the volume under the probability surface defined by $U(x, y, t)$ within the square $-1 < x < 1$, $-1 < y < 1$. Intuitively, this is because coalescence is represented by the diffusion of the probability density out of the square. Thus, the instantaneous rate of coalescence at time t is given by the rate at which the probability volume within the square is decreasing at that time. From this relationship it is possible to derive the expectation of the time to coalescence for two sequences sampled from locations corresponding to x_0 and y_0:

$$E[T] = \frac{1}{2} + \sum_i \sum_j \frac{2}{\sigma_m^2(\alpha_i^2 + \alpha_j^2)} \frac{\sin(\alpha_i) f_i(x_0)}{\alpha_i + \sin(\alpha_i)\cos(\alpha_i)} \frac{\sin(\alpha_j) f_j(y_0)}{\alpha_j + \sin(\alpha_j)\cos(\alpha_j)}. \quad (9)$$

It is possible using this approach to derive a number of other analytic results, including the full probability distribution for the time to coalescence (Equations A25 and A33). Each of the moments of the distribution has a simple form analogous to that for the expectation (Equations A26–A31 and A34). It is also possible to write down the exact distribution of the locations of two lineages that have not yet coalesced (A40), as well as the distribution of locations of the coalescent events (A44–A46).

RESULTS

The results derived in this article have been compared to simulation data to assess the accuracy of the diffusion approximation in this context. This section contains analytic and simulation results for a number of values over a range of parameters. These results provide both reassurance regarding the accuracy of the equations and insight into the behavior of the coalescent process in a continuous, linear habitat.

Monte Carlo simulations were performed backward in time using two different migration/coalescence processes. In the first process, the locations of the two lineages were kept as floating point numbers. Migration each generation was performed by drawing a random number from a normal distribution. If the new location for the lineage lay outside the habitat range, the new location was selected by reflection at the habitat boundary. A coalescent event occurs if, after translocation and reflection, the two lineages lie within a distance $1/(2N)$ of each other. Times and locations of coalescent events were averaged over a large number of sample runs.

The second process used a discrete lattice model. Each of the two lineages was assigned an integer location between 1 and N. Each generation two pseudorandom integers were drawn from independent Poisson distributions for each lineage, which were then translocated by the difference of the two Poissons. This produces a discrete distribution that approximates the shape of a normal distribution. Reflections were performed as in the first process. If, after migration and reflection, the two lineages were at the same location (had the same integer location value), a coalescent event was considered to have occurred. Again, the times and locations of the coalescent events were averaged over a large number of sample runs.

The relative value of the mean time to coalescence is determined largely by the product $N\sigma_m^2$, which is analogous to Nm in discrete-deme models of geographical structure. Presented here are three sets of data representing three different values of $N\sigma_m^2$. In all three cases $N = 200$. The values of σ_m^2 are 0.005, 0.0005, and 0.00005 ($N\sigma_m^2 = 1.0$, 0.1, and 0.01). For each value of σ_m^2, the mean time to coalescence was determined for a pair of sequences for a number of sampling locations. Figure 3 presents mean time to coalescence determined analytically from Equation 9 for the parameters used in Table 2 ($N\sigma_m^2 = 0.1$). Mean times to coalescence in Tables 1–3 are determined by three methods. The first two values are simulation results (by the Poisson and normal meth-

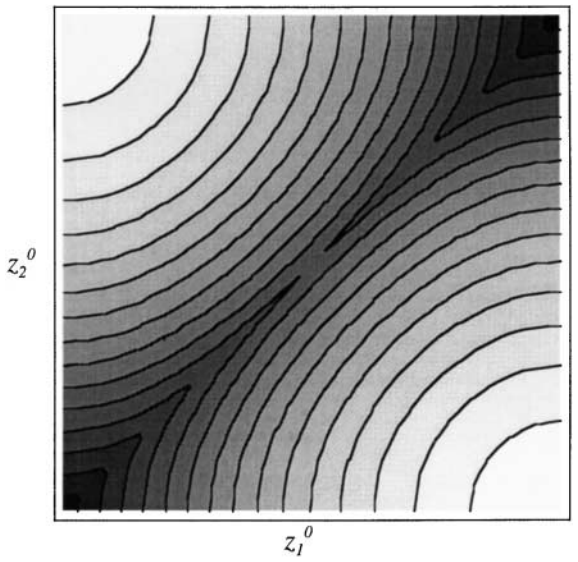

FIGURE 3.—This surface represents the mean coalescence times for pairs of sequences drawn from various locations in the habitat range. The data here are equivalent to those presented in Table 2 ($N = 100$, $\sigma_m^2 = 0.001$). The x- and y-axes represent the two sampling locations. Darker areas correspond to shorter mean coalescence times. The lower-left and upper-right corners represent the case where both samples are taken from one end of the range. The other two corners (with the longest coalescence times) represent the case where the two samples are taken from opposite ends of the range. Mean coalescence times are longer for samples taken from the center of the array than for samples taken from near the ends.

ods described above) from 1 million replicates. The third value is the analytically determined value from Equation 9.

Inspection of the tables reveals that the analytically derived mean time to coalescence is in good agreement with simulated data, with the results of the three methods differing typically by no more than 0.5%. Tables 1–3 and Figure 3 also immediately reveal two features of this model. First is the intuitively pleasing result that the mean time to coalescence increases with the physical distance between the two sampled sequences. The rate of increase with distance is dependent on the migration rate, with lower migration rates corresponding to higher rates of genetic divergence with distance.

A second, less intuitive, result from these data is the dependence of time to coalescence on the location of the two sampled sequences (in contrast to their separation). The pattern is most easily seen by considering pairs of adjacent sequences sampled from various locations along the habitat. A pair of adjacent sequences sampled from the center of the population range has a longer mean time to coalescence than pairs sampled closer to the ends, an effect that is more pronounced at lower migration rates. This result is anticipated by classical population genetics results in which the probability of identity by descent is higher for demes near a reflecting boundary (MARUYAMA 1970c; NAGYLAKI and BARCILON 1988) and by the work of HERBOTS (1994).

The model also provides results regarding the locations of the lineages and common ancestors (Equations A40–A46). Figure 4 shows the probability surface for the time and location of coalescence for three different pairs of sequences (from Equation A44). Recent coalescent events are likely to be found in the region between the two sampling sites. In the more distant past, the probability distribution depends only on the migration rate and not the sampling locations. This distant-past distribution is biased toward the center of the range. Another result that can be derived from the model is the distribution of the lineage locations conditional on their still being separate at a time t in the past (Equation A40). In the distant past, this distribution is skewed toward the edges. Intuitively, if the two lineages are separated by a very deep genealogical branch, it most likely results from their having spent a lot of time at opposite ends of the range. A number of other results are also derived and presented in the APPENDIX, including the strong-migration limit (NAGYLAKI 1980, 2000) and application of these results to a linear array of demes.

APPLICATION TO DATA

The expected time to coalescence can be used to estimate demographic parameters. In this section published sequence data are used to fit the model and estimate the effective population size and the genetic dispersal rate. Our purpose here is not to determine specific parameter values for a particular organism. In fact, the population considered below is likely to violate one or more assumptions of the model. Our goal is simply to illustrate the fact that patterns of genetic diversity such as the one predicted by the model may be found in nature. We also want to emphasize the fact that the finite linear model makes different predictions under neutrality than either an island or a circular model and that these differences may alter our interpretation of sequence data.

We have applied the model developed here to sequence data collected from the mitochondrial control region in the five different regional forms of sardines (*Sardinops*) in the Indian and Pacific oceans (BOWEN and GRANT 1997). Sardines are characterized by an antitropical distribution and are restricted to five temperate upwelling zones off the coasts of Japan, California, Chile, Australia, and South Africa. Temperate waters extend continuously from South Africa through Australia to Chile and from Japan to North America. The two temperate zones are separated by warmer tropical waters in the Pacific. However, BOWEN and GRANT (1997) point out that this tropical zone is fairly narrow in the eastern Pacific, along the west coast of Mexico, suggesting that genetic contact between the California and Chile sardine populations is or has been possible. In

TABLE 1

Comparison of derived and simulated mean coalescence times ($N = 100$; $\sigma_m^2 = 0.0001$)

	z_2^0										
$N\sigma_m^2 = 0.01$	0.0075	0.1025	0.2025	0.3025	0.4025	0.5025	0.5975	0.6975	0.7975	0.8975	0.9975
z_1^0											
0.0025	36.60	281.68	669.21	1113.47	1570.23	2001.93	2371.48	2668.08	2932.13	3078.44	3133.22
	37.47	279.72	667.07	1110.76	1569.15	1995.55	2364.42	2684.84	2926.59	3079.56	3129.56
	37.41	280.99	669.28	1113.92	1568.95	2001.45	2368.60	2690.41	2931.44	3081.25	3133.56
0.0975		141.70	549.87	1014.72	1487.27	1928.50	2305.56	2632.17	2878.58	3032.74	
		141.72	548.70	1014.94	1484.19	1929.06	2306.08	2630.55	2883.32	3033.48	
		142.65	549.32	1014.88	1485.21	1929.46	2305.30	2634.11	2880.09	3032.88	
0.1975			206.42	712.98	1226.90	1705.66	2112.37	2460.49	2727.98		
			204.91	709.88	1224.66	1706.43	2110.57	2460.18	2720.40		
			206.29	712.63	1228.65	1708.52	2111.00	2461.33	2722.62		
0.2975				245.53	809.11	1344.81	1794.98	2178.62			
				244.31	809.61	1343.76	1788.08	2173.63			
				244.74	808.97	1346.20	1792.28	2178.23			
0.3975					264.38	852.93	1361.66				
					263.59	851.07	1363.00				
					265.72	854.03	1359.55				
0.4975						273.02					
						271.91					
						272.41					

Tables 1–3 present a comparison of analytically derived mean coalescence times with values determined by simulation. Row and column headings indicate the sampling locations for the two sequences, in terms of distance along the population range. The leftmost nonempty cell in each row represents the case of two adjacent sequences at various points along the habitat. Mean coalescence times are longer for pairs of sequences that are more distant, as expected, but also for pairs of sequences drawn from closer to the center of the range.

the western Pacific, the tropical zone is much broader, making genetic exchange between Japan and Australia unlikely.

On the basis of these observations, it may be reasonable to apply the finite, linear model to this system, treating the populations as linearly arrayed from Japan to California to Chile to Australia to South Africa, with genetic contact possible only between adjacent populations. The total length of this range is ~25,000 miles, with the five sampling sites occurring at ~5000-mile intervals. These five sites yield 15 pairwise comparisons, which were fit to the model by finding the parameter values that minimize the sum of the squares of the differences between the predicted and observed values. The statistical properties of these estimators are not investigated here. However, in this case, the model does appear to fit the data reasonably well, reproducing the same pattern of genetic diversity, and it provides a framework that highlights certain features of the data.

Figure 5 shows the observed and expected average number of pairwise nucleotide differences (proportional to the expected coalescence time) in the data set for parameter values minimizing the sum of the squares of the errors. The observed data manifest the two key features of the model: increasing genetic differentiation with distance and greater genetic diversity near the center of the habitat. The product $N\sigma_m^2$ was estimated to be 0.017, and $\theta (= 2N\mu)$ was estimated to be 1.64. Assuming a per generation mutation rate of $\sim 7.5 \times 10^{-6}$ for the mitochondrial control region (on the basis of a divergence rate of 15% per million years between lineages and a mean generation time of 2 years, Butler *et al.* 1996), this gives a total mitochondrial effective population size of 2×10^5. This value is likely much smaller than the census size, possibly resulting from a higher variance of reproductive success than that assumed by the model and also consistent with suggestions of population fluctuations in the species (Soutar and Isaacs 1974). This produces an estimate of $\sigma_m^2 = 8.5 \times 10^{-8}$, which corresponds to a mean intergenerational migration distance of ~8 miles and a (mitochondrial) neighborhood size of ~200.

Inferences can also be drawn from differences in observed and expected values. In the fourth column of Figure 5 (labeled "Calif."), the observed interpopulation values are all lower than the corresponding expected values. In the fifth column ("Japan"), on the other hand, the expected values are lower than the observed. This pattern suggests that the California and Chile populations are more closely connected genetically than might

TABLE 2

Comparison of derived and simulated mean coalescence times ($N = 100$; $\sigma_m^2 = 0.001$)

	z_2^0										
$N\sigma_m^2 = 0.1$	0.0075	0.1025	0.2025	0.3025	0.4025	0.5025	0.5975	0.6975	0.7975	0.8975	0.9975
z_1^0											
0.0025	119.82	163.96	223.05	280.07	333.74	381.82	421.42	454.09	478.93	494.39	499.97
	120.64	163.78	223.23	280.18	333.85	382.85	420.77	454.56	479.53	494.62	499.57
	120.25	164.58	223.64	281.02	334.57	382.57	421.76	455.20	479.78	494.87	500.11
0.0975		162.54	213.61	272.87	326.64	375.96	415.52	449.19	473.93	489.99	
		162.57	214.46	272.85	327.37	375.94	415.54	449.84	473.32	489.23	
		163.01	214.99	273.36	327.72	376.38	416.08	449.93	474.80	490.07	
0.1975			198.57	249.50	306.66	357.14	398.07	433.25	458.77		
			198.55	248.96	306.15	356.94	397.92	432.57	459.22		
			198.55	249.78	306.62	357.31	398.56	433.68	459.45		
0.2975				221.89	270.87	325.54	369.21	406.74			
				221.88	270.71	324.91	368.85	406.40			
				222.20	271.71	325.73	369.57	406.81			
0.3975					235.49	281.81	328.99				
					235.82	281.91	329.18				
					235.77	282.31	329.72				
0.4975						240.45					
						240.48					
						240.20					

See Table 1 legend for details.

be predicted from distance alone, whereas the Japan and California populations are genetically more distant than expected. This observation supports Bowen and Grant's (1997) argument that the tropical barrier in the eastern Pacific is, or has recently been, traversible. In fact, the data suggest that the tropical water in the

TABLE 3

Comparison of derived and simulated mean coalescence times ($N = 100$; $\sigma_m^2 = 0.01$)

	z_2^0										
$N\sigma_m^2 = 1.0$	0.0075	0.1025	0.2025	0.3025	0.4025	0.5025	0.5975	0.6975	0.7975	0.8975	0.9975
z_1^0											
0.0025	188.44	190.30	197.90	204.65	210.72	216.22	220.51	223.88	226.56	228.25	228.44
	188.45	191.10	198.18	205.18	211.41	216.87	220.68	224.16	227.15	227.98	228.81
	188.43	191.53	198.19	205.20	211.47	216.82	221.04	224.53	227.05	228.57	229.09
0.0975		192.85	197.43	204.55	210.39	215.47	219.72	223.90	226.49	227.45	
		193.30	197.77	204.27	210.65	216.18	220.30	223.77	226.13	227.52	
		193.15	198.24	204.74	210.96	216.32	220.54	224.05	226.57	228.09	
0.1975			199.03	203.39	208.78	214.40	218.49	222.23	224.11		
			199.82	203.53	209.24	214.71	218.58	222.28	225.39		
			199.53	203.79	209.44	214.77	219.01	222.54	225.08		
0.2975				203.85	207.00	211.61	215.78	219.63			
				203.90	207.25	211.90	216.37	220.01			
				204.11	207.47	212.23	216.46	220.02			
0.3975					206.53	208.47	212.25				
					206.95	209.15	212.57				
					206.84	209.30	212.97				
0.4975						207.26					
						207.88					
						207.74					

See Table 1 legend for details.

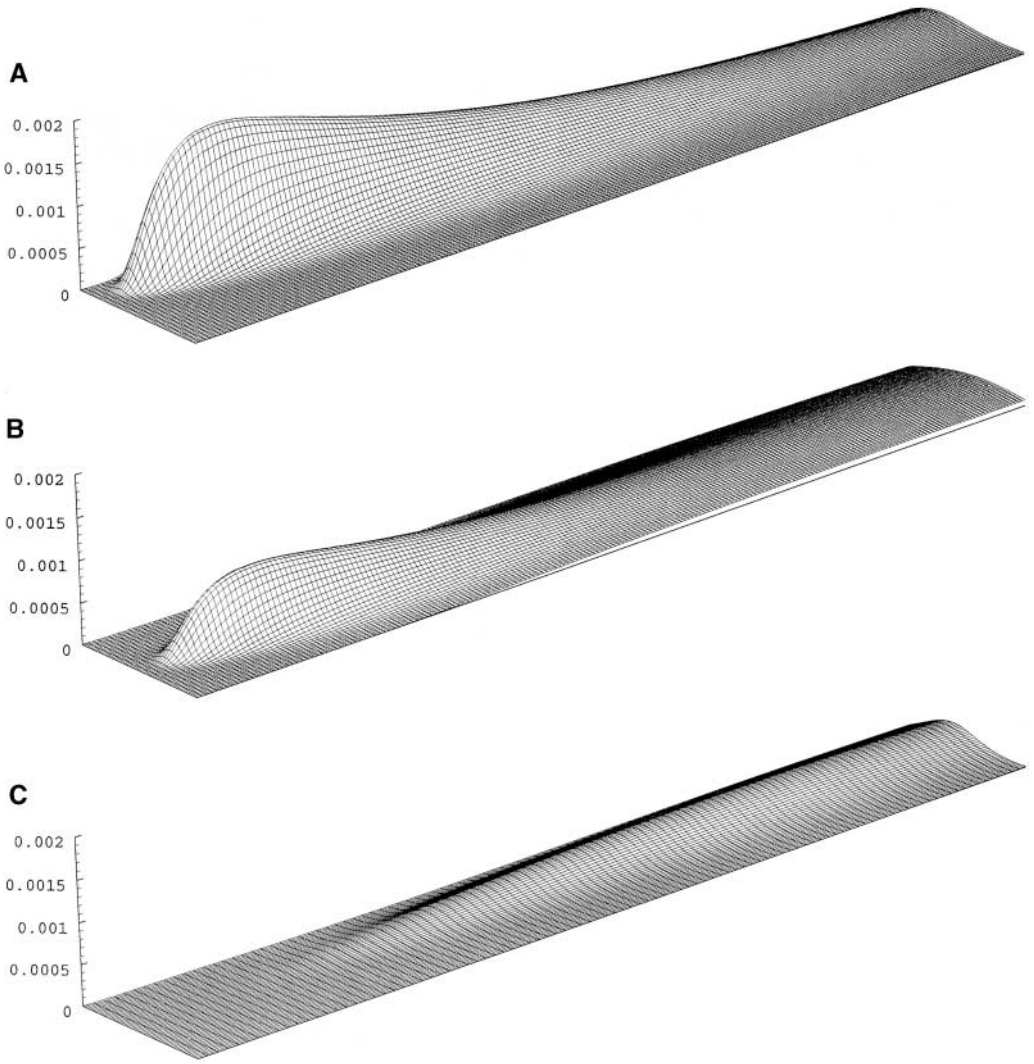

FIGURE 4.—Probability surfaces for the time and location of the most recent common ancestor. (A–C) The probability distributions for the times and locations of coalescent events for three different pairs of sampling locations. Location (spanning the entire habitat range) is plotted along the short axis, and time (from 0 to 2000 generations in the past) along the long axis; $2N = 2000$ and $\sigma_m^2 = 0.00005$. The sample locations (z_1^0, z_2^0) are (A) (0, 0.33), (B) (0.33, 0.67), and (C) (0, 0.67). Initially there is no chance of coalescence due to the separation of the sequences. In all three cases the most recent coalescence events are positioned right between the two sampling sites. Note that the surface is taller in A than in B, in spite of the fact that the two graphs represent the same separation between samples. This difference corresponds to the shorter average coalescence times for sequences taken from closer to the edges of the habitat range. Note also that the peak in C is smaller and shifted out on the time axis relative to A and B, as a result of the larger distance separating the sampling locations. As time in the past becomes very large, all three distributions approach a common shape, with coalescence locations biased toward the center of the range.

eastern Pacific may represent less of a genetic barrier than the equally large band of temperate water between North America and Japan.

This analysis is presented not to address specific issues in sardine biogeography or question the conclusions of Bowen and Grant, who attribute the observed pattern of genetic diversity to a range expansion of the sardine populations. Our purpose has been simply to show that patterns predicted by the model can, in fact, be found in natural populations and to illustrate how the model can be employed to estimate interesting demographic parameters. Furthermore, the analysis suggests how observed patterns of genetic diversity can be made more meaningful when compared to a more sophisticated null model.

DISCUSSION

We have presented a model for analyzing genetic diversity in a finite, continuous, linear population. Using a diffusion approximation, the model fully characterizes the distribution of possible genealogical histories of a pair of sequences sampled from such a population. Results derived from the model include the full distribution of coalescence times and locations, as well as a number of summary statistics, such as the mean time to the most recent common ancestor.

The analytic results derived from the model are in good agreement with simulations over a wide range of parameter values. This agreement extends even to extremely small neighborhood sizes (approaching one), allowing us to relax the usual coalescent theory assumption of a large local population size. In the other extreme, the strong migration limit where the neighborhood size approaches the population size, the model converges on well-established results for the coalescent process in a panmictic population.

The model makes several predictions regarding genealogies in a finite continuous habitat. In addition to the intuitive result that genetic divergence increases with distance, the model predicts that genetic diversity will

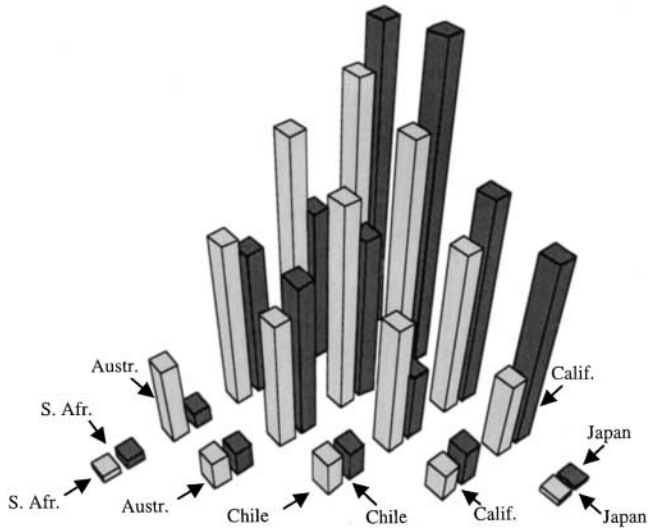

FIGURE 5.—Average pairwise differences for the mitochondrial control region in sardines. The average number of nucleotide differences between pairs of sequences sampled from five locations (dark shaded bars) are plotted here along with the values predicted from the model (light shaded bars). Parameter values for the expected results were chosen to minimize the sum of the squares of the differences between the expected and observed values. In the fourth column (California), the observed values are consistently lower than expected, suggesting that the barrier to gene flow between California and Chile is lower than might be suggested by distance alone. Similarly, the higher-than-expected values in the fifth column (Japan) suggest that the temperate zone in the northern Pacific represents a larger barrier than would be predicted by distance alone.

be greater near the center of the habitat than at the edges. Coalescent events in the recent past are most likely to occur between the sampling locations of the two sequences. In the distant past, the distribution of locations of coalescent events becomes independent of sampling location and is concentrated toward the center of the habitat. The locations of lineages (conditional on not having coalesced), on the other hand, are biased toward the edges of the habitat in the distant past. All of these effects are more pronounced under lower migration.

In this development of the model, we have assumed reflecting habitat boundaries, meaning that individuals suffer no loss of fecundity when they are situated at the edge of the habitat. It may be more reasonable to assume absorbing habitat boundaries. In the forward-time model, this would mean that gametes that dispersed outside the habitat range would be lost. The effect in the backward-time model would be to bias the distribution of lineage locations slightly toward the center of the habitat, decreasing slightly the time to common ancestry, but preserving the broad patterns described for the reflecting-boundaries case.

One property of the model not discussed above is the dependence of the coalescent process on neighborhood size. For a given pair of locations, the ratio of the expected time to coalescence to the total population size is determined primarily by the product $N\sigma_m^2$, which is analogous to Nm in demic models of population structure. However, N and σ_m enter into minor terms in the equations separately, meaning that it is possible, in principle, to estimate these two parameters independently without an independent estimate of the population size (*e.g.*, from θ and μ). In practice, however, it seems unlikely that sufficient data could be collected (or that a population could be found that conformed closely enough to the model) to separately estimate these parameters with any certainty.

It is also possible using this model to derive other values for a particular set of parameters. SLATKIN (1991) derived F_{ST} in relation to mean time to coalescence for pairs of genes as

$$F_{ST} = \frac{\bar{t} - \bar{t}_0}{\bar{t}}, \tag{10}$$

where $\bar{t}_0$ represents the mean time to coalescence for a pair of sequences drawn from the same deme, and $\bar{t}$ represents the mean time to coalescence for a pair drawn at random from the entire population. These two values can be derived from Equation 9, where $\bar{t}_0$ is the average value along the line ($x_0 = 1$, $-1 < y_0 < 1$), and $\bar{t}$ is the average value over the entire space ($-1 < x_0 < 1$, $-1 < y_0 < 1$). In fact, the value of $\bar{t}_0$ is very nearly N, independent of the value of σ_m^2, consistent with the observation that the mean within-deme coalescence time will average to N in any system with conservative migration (STROBECK 1987; NAGYLAKI 1998).

The distribution of coalescence times given by Equation A25 can also be combined with a particular mutational model. Integration of this probability against the mutational process will yield the probability of identity in state or the likelihood of a particular set of differences between the two sequences. Results such as these may be valuable in the analysis of sequence data.

Mathematica files for generating the results described in this article are available from the authors, as is a C program for estimating parameter values from sequence data.

We thank N. H. Barton, J. L. Cherry, T. Nagylaki, A. Platt, J. Wall, and three anonymous reviewers for helpful discussions and comments on the manuscript. This work was supported by a grant from the Howard Hughes Medical Institute to J.F.W. and in part by National Science Foundation grant no. DEB-9815367 to J.W.

LITERATURE CITED

BARTON, N. H., and I. WILSON, 1995 Genealogies and geography. Philos. Trans. R. Soc. Lond. B **349:** 49–59.

BARTON, N. H., and I. WILSON, 1996 Genealogies and geography, pp. 23–56 in *New Uses for New Phylogenies*, edited by P. H. HARVEY, A. J. LEIGH BROWN and J. MAYNARD SMITH. Oxford University Press, Oxford.

BOWEN, B. W., and W. S. GRANT, 1997 Phylogeography of the sar-

dines (*Sardinops* spp.): assessing biogeographic models and population histories in temperate upwelling zones. Evolution **51:** 1601–1610.

Butler, J. L., M. L. Granados, J. T. Barnes, M. Yaremko and B. J. Macewicz, 1996 Age composition, growth, and maturation of the Pacific sardine (*Sardinops sagax*) during 1994. Calif. Coop. Oceanic Fish. Invest. Rep. **37:** 152–159.

Churchill, R. V., and J. W. Brown, 1987 *Fourier Series and Boundary Value Problems.* McGraw-Hill, New York.

Durrett, R., and S. A. Levin, 1994 Stochastic spatial models: a user's guide to ecological applications. Philos. Trans. R. Soc. Lond. B **343:** 329–350.

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

Fleming, W. H., and C.-H. Su, 1974 Some one-dimensional migration models in population genetics theory. Theor. Popul. Biol. **5:** 431–449.

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

Herbots, H. M., 1994 *Stochastic Models in Population Genetics: Genealogy and Genetic Differentiation in Structured Populations.* Ph.D. Thesis, University of London.

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

Holley, R. A., and T. M. Liggett, 1975 Ergodic theorems for weakly interacting infinite systems and the voter model. Ann. Probab. **3:** 643–663.

Kimura, M., and G. H. Weiss, 1964 The stepping stone model of population structure and the decrease of genetic correlation with distance. Genetics **49:** 561–576.

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

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

Maruyama, T., 1970b On the rate of decrease of heterozygosity in circular stepping-stone models of populations. Theor. Popul. Biol. **1:** 101–119.

Maruyama, T., 1970c Analysis of population structure. I. One dimensional stepping-stone models of finite length. Ann. Hum. Genet. **34:** 201–219.

Maruyama, T., 1970d The rate of decrease of heterozygosity in a population occupying a circular or a linear habitat. Genetics **67:** 437–454.

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

Maruyama, T., 1972 Rate of decrease of genetic variability in a two-dimensional continuous population of finite size. Genetics **70:** 639–651.

Nagylaki, T., 1974a Genetic structure of a population occupying a circular habitat. Genetics **78:** 777–790.

Nagylaki, T., 1974b The decay of genetic variability in geographically structured populations. Proc. Natl. Acad. Sci. USA **71:** 2932–2936.

Nagylaki, T., 1976 The decay of genetic variability in geographically structured populations. Theor. Popul. Biol. **10:** 70–82.

Nagylaki, T., 1977 Genetic structure of a population occupying a circular habitat. Genetics **78:** 777–790.

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

Nagylaki, T., 1998 The expected number of heterozygous sites in a subdivided population. Genetics **149:** 1599–1604.

Nagylaki, T., 2000 Geographical invariance and the strong-migration limit in subdivided populations. J. Math. Biol. **41:** 123–142.

Nagylaki, T., and V. Barcilon, 1988 The influence of spatial inhomogeneities of neutral models of geographical variation. II. The semi-infinite linear habitat. Theor. Popul. Biol. **33:** 311–343.

Sawyer, S., 1976 Results for the stepping-stone model for migration in population genetics. Ann. Probab. **4:** 699–728.

Sawyer, S., 1977 Asymptotic properties of the equilibrium probability of identity in a geographically structured population. Adv. Appl. Probab. **9:** 268–282.

Slatkin, M., 1991 Inbreeding coefficients and coalescence times. Genet. Res. **58:** 167–175.

Soutar, A., and J. D. Isaacs, 1974 Abundance of pelagic fish during the 19th and 20th centuries as recorded in anaerobic sediments off the Californias. Fish. Bull. **72:** 257–273.

Strobeck, C., 1987 Average number of nucleotide differences in a sample from a single subpopulation: a test for population subdivision. Genetics **117:** 149–153.

Weiss, G. H., and M. Kimura, 1965 A mathematical analysis of the stepping stone model of genetic correlation. J. Appl. Probab. **2:** 129–149.

Wilkinson-Herbots, H. M., 1998 Genealogy and subpopulation differentiation under various models of population structure. J. Math. Biol. **37:** 535–585.

Wright, S., 1931 Evolution in Mendelian populations. Genetics **16:** 97–159.

Wright, S., 1943 Isolation by distance. Genetics **31:** 114–138.

Communicating editor: M. W. Feldman

APPENDIX

Derivation of formulas: Let the initial positions of the two sequences be denoted by z_1^0 and z_2^0, where z_1^0 and z_2^0 represent the relative locations along the length of the entire habitat, and therefore both lie between 0 and 1. Let x_0 and y_0 be the following transformations of these coordinates:

$$x_0 = 1 - |z_1^0 - z_2^0| \tag{A1a}$$

$$y_0 = z_1^0 + z_2^0 - 1. \tag{A1b}$$

A pair of coordinates (x, y) then fully describes the system of two identical particles along the line from 0 to 1. The state space in this coordinate system is a right triangle ranging from -1 to 1 in y and from $|y|$ to 1 in x. The state of the system at a time t generations in the past is given by a probability function $U(x, y, t) = U_x(x, t)\, U_y(y, t)$ on this state space. Migration of the two particles along the line is considered to be a diffusion process with constant σ_m^2, and the corresponding diffusion process in the transformed (x, y) coordinate system is a two-dimensional diffusion process with constant $2\sigma_m^2$ in each direction.

The two short sides of the triangular state space represent reflecting boundaries, which may be eliminated by the method of reflecting the state space across the boundary. Three such reflections generate a square state space ranging from -1 to 1 in each direction. Note that crossing over the diagonals of this square involves a transposition of x and y. However, since the diffusion process is isotropic, this transposition does not affect the analysis.

Because the x and y diffusion processes are now separable, further discussion focuses on a one-dimensional diffusion process. The two-dimensional process can be reconstructed by multiplication of two such one-dimensional processes. The derivation uses only x. The equations for the diffusion process in y are identical.

The long side of the triangular state space, which is now replicated four times as the boundary of the new square state space, represents a partially reflecting, par-

tially absorbing boundary. The diffusion process has now been reduced to a Sturm-Liouville-type problem, where the boundary conditions are set by a relationship between the flux rate across the boundary and the density function within the boundary.

The function $U_x(x, t)$ must satisfy the diffusion equation within the range $(-1, 1)$,

$$\frac{\partial U_x}{\partial t} = 2\sigma_m^2 \frac{\partial^2 U_x}{\partial x^2}, \tag{A2}$$

and is subject to the boundary conditions

$$J = -2\sigma_m^2 \frac{\partial U_x}{\partial x} = \frac{1}{N}\frac{1}{\sqrt{4\pi\sigma_m^2}}\int_0^\infty U_x(1 - s, t)e^{-s^2/4\sigma_m^2}ds \quad \text{at } x = 1 \tag{A3a}$$

$$J = 2\sigma_m^2 \frac{\partial U_x}{\partial x} = \frac{1}{N}\frac{1}{\sqrt{4\pi\sigma_m^2}}\int_0^\infty U_x(s - 1, t)e^{-s^2/4\sigma_m^2}ds \quad \text{at } x = -1. \tag{A3b}$$

In these equations, the flux J across the boundary is set equal to the average probability over the next generation that the separation between the two lineages is $<1/(2N)$. Since there are N individuals in the population and complete density-dependent population regulation, it is assumed that two lineages separated by less than one-half the "width" of an individual must be the same lineage. Thus the flux rate across the boundary is equal to the rate of coalescence. The probability that the two lineages coalesce in the previous generation is proportional both to this width and to the probability that the two lineages are separated by a distance s and that this separation decreases by s in one generation of dispersal, integrated over all possible values of s. In Equations A3a and A3b, a normal distribution has been assumed for the single-generation dispersal pattern. Other dispersal patterns are, of course, possible. However, most patterns will converge on a normal distribution after a relatively small number of generations. Reflections have been ignored in these two equations, as it is assumed that single-generation migration events much longer than the total habitat length are extremely rare.

The general solution to the diffusion equation is

$$U_x(x, t) = (C_1\cos(\alpha x) + C_2\sin(\alpha x))e^{-2\sigma_m^2\alpha^2 t} \tag{A4}$$

and we can incorporate the boundary conditions by using Taylor-series expansions

$$-2\sigma_m^2\frac{\partial U_x}{\partial x}\bigg|_{x=1} = \frac{1}{N}\frac{1}{\sqrt{4\pi\sigma_m^2}}\int_0^\infty \sum_{n=0}^\infty U_x^{(n)}(1, t)\frac{(-s)^n}{n!}e^{-s^2/4\sigma_m^2}ds \tag{A5a}$$

$$2\sigma_m^2\frac{\partial U_x}{\partial x}\bigg|_{x=-1} = \frac{1}{N}\frac{1}{\sqrt{4\pi\sigma_m^2}}\int_0^\infty \sum_{n=0}^\infty U_x^{(n)}(-1, t)\frac{s^n}{n!}e^{-s^2/4\sigma_m^2}ds, \tag{A5b}$$

where $U_x^{(n)}$ represents the nth derivative of U_x with respect to x. The right-hand sides of these equations can be integrated term by term to give

$$-2\sigma_m^2\frac{\partial U_x}{\partial x}\bigg|_{x=1} = \frac{1}{N}\frac{1}{\sqrt{4\pi\sigma_m^2}}\sum_{n=0}^\infty \frac{U_x^{(n)}(1, t)}{n!}(-1)^n 2^n(\sigma_m^2)^{(n+1)/2}\Gamma\left(\frac{n+1}{2}\right) \tag{A6a}$$

$$2\sigma_m^2\frac{\partial U_x}{\partial x}\bigg|_{x=-1} = \frac{1}{N}\frac{1}{\sqrt{4\pi\sigma_m^2}}\sum_{n=0}^\infty \frac{U_x^{(n)}(-1, t)}{n!}2^n(\sigma_m^2)^{(n+1)/2}\Gamma\left(\frac{n+1}{2}\right). \tag{A6b}$$

Substituting in expressions for the derivatives of U_x (derived from Equation A4), we get

$$\begin{aligned}2N\sigma_m^2\sqrt{4\pi\sigma_m^2}\alpha(C_1\sin(\alpha) - C_2\cos(\alpha)) &= \sum_{n=0}^\infty \frac{(C_1\cos(\alpha) + C_2\sin(\alpha))}{(2n)!}(-1)^n\alpha^{2n}2^{2n}(\sigma_m^2)^{n+1/2}\Gamma(n + \tfrac{1}{2}) \\ &\quad + \sum_{n=0}^\infty \frac{(C_1\sin(\alpha) - C_2\cos(\alpha))}{(2n+1)!}(-1)^n\alpha^{2n+1}2^{2n+1}(\sigma_m^2)^{n+1}\Gamma(n+1)\end{aligned} \tag{A7a}$$

and

$$\begin{aligned}-2N\sigma_m^2\sqrt{4\pi\sigma_m^2}\alpha(C_1\sin(-\alpha) - C_2\cos(-\alpha)) &= \sum_{n=0}^\infty \frac{(C_1\cos(-\alpha) + C_2\sin(-\alpha))}{(2n)!}(-1^n)\alpha^{2n}2^{2n}(\sigma_m^2)^{n+1/2}\Gamma(n + \tfrac{1}{2}) \\ &\quad - \sum_{n=0}^\infty \frac{(C_1\sin(-\alpha) - C_2\cos(-\alpha))}{(2n+1)!}(-1)^n\alpha^{2n+1}2^{2n+1}(\sigma_m^2)^{n+1}\Gamma(n+1).\end{aligned} \tag{A7b}$$

Collecting terms by C_1 and C_2 and simplifying, these conditions become

$$\begin{aligned}C_1[2N\sigma_m^2\sqrt{4\pi\sigma_m^2}\alpha\sin(\alpha) - \sum_{n=0}^\infty \frac{\cos(\alpha)}{(2n)!}(-1)^n\alpha^{2n}2^{2n}\sigma_m^{2n+1}\Gamma(n + \tfrac{1}{2}) \\ - \sum_{n=0}^\infty \frac{\sin(\alpha)}{(2n+1)!}(-1)^n\alpha^{2n+1}2^{2n+1}\sigma_m^{2n+2}\Gamma(n+1)] \\ = C_2[2N\sigma_m^2\sqrt{4\pi\sigma_m^2}\alpha\cos(\alpha) + \sum_{n=0}^\infty \frac{\sin(\alpha)}{(2n)!}(-1)^n\alpha^{2n}2^{2n}\sigma_m^{2n+1}\Gamma(n + \tfrac{1}{2}) \\ - \sum_{n=0}^\infty \frac{\cos(\alpha)}{(2n+1)!}(-1)^n\alpha^{2n+1}2^{2n+1}\sigma_m^{2n+2}\Gamma(n+1)]\end{aligned} \tag{A8a}$$

and

$$\begin{aligned}C_1[2N\sigma_m^2\sqrt{4\pi\sigma_m^2}\alpha\sin(\alpha) - \sum_{n=0}^\infty \frac{\cos(\alpha)}{(2n)!}(-1)^n\alpha^{2n}2^{2n}\sigma_m^{2n+1}\Gamma(n + \tfrac{1}{2}) \\ - \sum_{n=0}^\infty \frac{\sin(\alpha)}{(2n+1)!}(-1)^n\alpha^{2n+1}2^{2n+1}\sigma_m^{2n+2}\Gamma(n+1)] \\ = -C_2[2N\sigma_m^2\sqrt{4\pi\sigma_m^2}\alpha\cos(\alpha) + \sum_{n=0}^\infty \frac{\sin(\alpha)}{(2n)!}(-1)^n\alpha^{2n}2^{2n}\sigma_m^{2n+1}\Gamma(n + \tfrac{1}{2}) \\ - \sum_{n=0}^\infty \frac{\cos(\alpha)}{(2n+1)!}(-1)^n\alpha^{2n+1}2^{2n+1}\sigma_m^{2n+2}\Gamma(n+1)].\end{aligned} \tag{A8b}$$

Two classes of nontrivial solutions satisfy both of these conditions simultaneously. For the first class of solutions, $C_2 = 0$ and α satisfies the following:

$$\begin{aligned}4N\sigma_m^2\sqrt{\pi}\alpha\sin(\alpha) &= \cos(\alpha)\sum_{n=0}^\infty \frac{1}{(2n)!}(-1)^n(2\alpha\sigma_m)^{2n}\Gamma(n + \tfrac{1}{2}) \\ &\quad + \sin(\alpha)\sum_{n=0}^\infty \frac{1}{(2n+1)!}(-1)^n(2\alpha\sigma_m)^{2n+1}\Gamma(n+1).\end{aligned} \tag{A9}$$

The second class of solutions will have $C_1 = 0$ and values of α that satisfy

$$4N\sigma_m^2\sqrt{\pi}\alpha\cos(\alpha) = -\sin(\alpha)\sum_{n=0}^{\infty}\frac{1}{(2n)!}(-1)^n(2\alpha\sigma_m)^{2n}\Gamma(n+\tfrac{1}{2}) + \cos(\alpha)\sum_{n=0}^{\infty}\frac{1}{(2n+1)!}(-1)^n(2\alpha\sigma_m)^{2n+1}\Gamma(n+1). \quad \text{(A10)}$$

These relationships can be further simplified once we consider the following two relationships for the gamma function:

$$\Gamma(n+1) = n! \quad \text{(A11a)}$$

$$\Gamma(n+\tfrac{1}{2}) = \frac{\prod_{i=0}^{n-1}(2i+1)}{2^{2n}}\sqrt{\pi}. \quad \text{(A11b)}$$

The specific solution for the diffusion equation with these boundary conditions then becomes

$$U_x(x) = \sum_i C_1^i\cos(\alpha_i x) + \sum_j C_2^j\sin(\alpha_j^* x), \quad \text{(A12)}$$

where α_i satisfies

$$\cot(\alpha_i) = \frac{4N\sigma_m^2\alpha_i + 1/\sqrt{\pi}\sum_{n=1}^{\infty}(-1)^n(2\sigma_m\alpha_i)^{2n-1}((n-1)!/(2n-1)!)}{1+\sum_{n=1}^{\infty}[(-\sigma_m^2\alpha_i^2)^n/\prod_{m=1}^{n}2m]} \quad \text{(A13a)}$$

and α_j^* satisfies

$$-\tan(\alpha_j^*) = \frac{4N\sigma_m^2\alpha_j^* + 1/\sqrt{\pi}\sum_{n=1}^{\infty}(-1)^n(2\sigma_m\alpha_j^*)^{2n-1}((n-1)!/(2n-1)!)}{1+\sum_{n=1}^{\infty}[(-\sigma_m^2\alpha_j^{*2})^n/\prod_{m=1}^{n}2m]}. \quad \text{(A13b)}$$

The normalization values for the eigenfunctions are

$$\|X_i\|^2 = \int_{-1}^{1}\cos^2(\alpha_i x)\,dx = \frac{\alpha_i + \sin(\alpha_i)\cos(\alpha_i)}{\alpha_i} \quad \text{(A14a)}$$

$$\|X_j'\|^2 = \int_{-1}^{1}\sin^2(\alpha_j' x)\,dx = \frac{\alpha_j' - \sin(\alpha_j')\cos(\alpha_j')}{\alpha_j'}, \quad \text{(A14b)}$$

which makes the normalized solution at time $t = 0$,

$$U_x(x,0) = \sum_i\frac{C_{1_i}\sqrt{\alpha_i}\cos(\alpha_i x)}{\sqrt{\alpha_i + \sin(\alpha_i)\cos(\alpha_i)}} + \sum_j\frac{C_{2_j}\sqrt{\alpha_j^*}\sin(\alpha_j^* x)}{\sqrt{\alpha_j^* - \sin(\alpha_j^*)\cos(\alpha_j^*)}}, \quad \text{(A15)}$$

where the C_1 and C_2 terms are derived by integrating the product of the probability distribution at time $t = 0$ with each term:

$$C_{1_i} = \int_{-1}^{1}\frac{f(x)\sqrt{\alpha_i}\cos(\alpha_i x)}{\sqrt{\alpha_i + \sin(\alpha_i)\cos(\alpha_i)}}dx \quad \text{(A16a)}$$

$$C_{2_j} = \int_{-1}^{1}\frac{f^*(x)\sqrt{\alpha_j^*}\cos(\alpha_j^* x)}{\sqrt{\alpha_j^* - \sin(\alpha_j^*)\cos(\alpha_j^*)}}dx. \quad \text{(A16b)}$$

If we take the initial probability distribution to be the δ function $\delta(x - x_0)$, then the system in x is represented as

$$U_x(x,t) = \sum_i\frac{\alpha_i\cos(\alpha_i x)\cos(\alpha_i x_0)}{\alpha_i + \sin(\alpha_i)\cos(\alpha_i)}e^{-2\sigma_m^2\alpha_i^2 t} + \sum_j\frac{\alpha_j^*\sin(\alpha_j^* x)\sin(\alpha_j^* x_0)}{\alpha_j^* - \sin(\alpha_j^*)\cos(\alpha_j^*)}e^{-2\sigma_m^2\alpha_j^{*2}t}. \quad \text{(A17)}$$

The α_i series terms (from Equation A13a) are of the form $\alpha_i = (i - 1 + \varepsilon_i)\pi$, where $0 < \varepsilon_{i+1} < \varepsilon_i < \tfrac{1}{2}$. Similarly, the α_j^* terms (from Equation A13a) are of the form $\alpha_j^* = (i - \tfrac{1}{2} + \varepsilon_j^*)\pi$, where $0 < \varepsilon_{j+1}^* < \varepsilon_j^* < \tfrac{1}{2}$. Since both series are monotonically increasing, and α^2 terms are found in the exponential, the series can be truncated for purposes of making calculations without significant loss of accuracy. The number of terms required is determined by parameter values, with more terms needed for smaller neighborhood sizes, and for samples that are closer together. The calculations presented in Tables 1–3 were truncated after 40, 20, and 7 terms, respectively.

The equations for the α_i and α_j^* can be simplified in the case where the neighborhood size is very large ($\sqrt{4\pi}N\sigma_m \gg 1$), in which case these equations become approximately

$$\cot(\alpha_i) = 4N\sigma_m^2\alpha_i \quad \text{(A18a)}$$

and

$$-\tan(\alpha_j^*) = 4N\sigma_m^2\alpha_j^*. \quad \text{(A18b)}$$

The form of the solution in equation A17 works well so long as the probability of coalescence in the first generation in the past is very small, that is, when the two sequences are separated by a sufficient distance (when x_0 is not close to 1) or when the neighborhood size is large. If x_0 is close to 1 and the neighborhood size is not large, we must use a more complex form for the initial conditions. This results from the fact that the model assumes discrete time steps, and so the shortest possible coalescence time is one generation. The diffusion approximation solution, on the other hand, assumes continuous time, and coalescence is possible at any time $t > 0$. When the probability of coalescence occurring within the first time step is very small, this correction is negligible. However, under certain conditions, we must account for the fact that one generation of migration occurs prior to the first opportunity for coalescence. This migration effectively moves the initial condition probability peak away from the boundary, resulting in a longer predicted time to coalescence.

The more accurate form of the solution, valid over all values of x_0, takes as its initial conditions a normal distribution of variance $2\sigma_m^2$, centered at x_0 and reflected off of the boundary at $x = 1$. The initial probability distribution is given approximately by

$$f(x) = \frac{1}{\sqrt{4\pi\sigma_m^2}}(e^{-(x-x_0)^2/4\sigma_m^2} + e^{-(2-x-x_0)^2/4\sigma_m^2}). \quad \text{(A19)}$$

Then the probability distribution in x as a function of t becomes

$$U_x(x,\ t)\ =\ \sum_{i=1}^{\infty}\frac{\alpha_i f_i(x_0)\cos(\alpha_i x)}{\alpha_i\ +\ \sin(\alpha_i)\cos(\alpha_i)}e^{-2\sigma_m^2\alpha_i^2(t-1/2)}$$
$$+\ \sum_{j=1}^{\infty}\frac{\alpha_j^* f_j^*(x_0)\sin(\alpha_j^* x)}{\alpha_j^*\ +\ \sin(\alpha_j^*)\cos(\alpha_j^*)}e^{-2\sigma_m^2\alpha_j^{*2}(t-1/2)},\quad\text{(A20)}$$

where

$$f_i(x_0)\ =\ \frac{1}{\sqrt{4\pi\sigma_m^2}}\int_{-1}^{1}\cos(\alpha_i x)\,(e^{-(x-x_0)^2/4\sigma_m^2}\ +\ e^{-(2-x-x_0)^2/4\sigma_m^2})\,dx\quad\text{(A21a)}$$

$$f_j^*(x_0)\ =\ \frac{1}{\sqrt{4\pi\sigma_m^2}}\int_{-1}^{1}\sin(\alpha_j^* x)\,(e^{-(x-x_0)^2/4\sigma_m^2}\ +\ e^{-(2-x-x_0)^2/4\sigma_m^2})\,dx.\quad\text{(A21b)}$$

Only a single reflection (at the point $x = 1$) is considered in deriving the initial conditions. More accurately, reflection off of the $x = -1$ point would also be included, as well as higher-order reflections (off of both $x = 1$ and $x = -1$, etc.). However, unless the migration rate is extremely high (high enough to completely eliminate any geographical structure), these terms are negligible if $x_0 > 0$. We have assumed by definition that $x_0 > 0$, but y_0 is allowed to range between -1 and 1. It should be noted, however, that all solutions for y_0 will be identical under the transformation $y_0 = -y_0$. Therefore, by considering only $y_0 > 0$, we can retain a complete description of the system and ignore migratory reflections off of the $y_0 = -1$ boundary in the first generation.

Assuming this distribution for the initial conditions is equivalent to permitting one-half generation of migration to occur prior to initiation of the coalescent process. In this way, coalescent events that occur over the first generation take place in the range of times $\frac{1}{2} < t < \frac{3}{2}$. Note that $t = 1$, the first time when coalescence can occur in the discrete time model, lies at the center of this range. This "premigration," which is necessary to compensate for approximations made in the translation from discrete to continuous time, gives rise to the "$t - \frac{1}{2}$" terms in Equation A20 and in subsequent derivations. Once these factors have been taken into account, a full description of the state of the system at time t is given by $U(x, y, t) = U_x(x, t)\ U_y(y, t)$, which can be manipulated to yield a number of other results.

Cumulative distribution function of the coalescence time: First we derive the probability that the two lineages have coalesced prior to time t. Recall that, in this formulation, coalescence is equivalent to diffusion outside of the square state space. The formulas derived above have been normalized so that, at $t = 0$, the volume under the probability surface is equal to one. At a time t, the probability that the two lineages are still separate is simply given by the volume under the probability surface within the square:

$$P_2(t)\ =\ \int_{-1}^{1}\int_{-1}^{1}U(x,\ y,\ t)\,dxdy\ =\ \int_{-1}^{1}U_x(x,\ t)\,dx\int_{-1}^{1}U_y(y,\ t)\,dy.\quad\text{(A22)}$$

Due to symmetry, the sine terms in the series expansions of U_x and U_y integrate to zero, leaving only a product of the integrals of the cosine series, and the cumulative distribution function of the coalescence time is

$$1\ -\ P_2(t)\ =\ 1\ -\ 4\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\frac{f_i(x_0)\sin(\alpha_i)}{\alpha_i\ +\ \sin(\alpha_i)\cos(\alpha_i)}\frac{f_j(y_0)\sin(\alpha_j)}{\alpha_j\ +\ \sin(\alpha_j)\cos(\alpha_j)}e^{-2\sigma_m^2(\alpha_i^2+\alpha_j^2)(t-1/2)}.\quad\text{(A23)}$$

Since all α_i are nonzero, this value approaches one as t goes to infinity.

Probability density function of the coalescence time: The probability that the two lineages coalesce at time t is given by the time derivative of the cumulative distribution function,

$$C(t)\ =\ -\frac{\partial}{\partial t}P_2(t),\quad\text{(A24)}$$

which yields

$$C(t)\ =\ 8\sigma_m^2\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\frac{f_i(x_0)\sin(\alpha_i)}{\alpha_i\ +\ \sin(\alpha_i)\cos(\alpha_i)}\frac{f_j(y_0)\sin(\alpha_j)}{\alpha_j\ +\ \sin(\alpha_j)\cos(\alpha_j)}$$
$$\times\ (\alpha_i^2\ +\ \alpha_j^2)\,e^{-2\sigma_m^2(\alpha_i^2+\alpha_j^2)(t-1/2)}\quad\text{(A25)}$$

for values of $t > \frac{1}{2}$. For $t < \frac{1}{2}$, $C(t) = 0$.

Moments of the distribution: The expectation and variance for the distribution of coalescence times, as well as all higher moments, can be derived from the expression for $C(t)$. The pth moment of the distribution is given by

$$E[T^p]\ =\ \int_0^{\infty}t^p C(t)\,dt.\quad\text{(A26)}$$

Since $C(t) = 0$ for $t < \frac{1}{2}$, this becomes

$$E[T^p]\ =\ \int_0^{\infty}(t\ +\ \tfrac{1}{2})^p C(t\ +\ \tfrac{1}{2})\,dt.\quad\text{(A27)}$$

The solution to

$$\int_0^{\infty}t^p C(t\ +\ \tfrac{1}{2})\,dt\quad\text{(A28)}$$

has a simple form, which can be used to derive the moments:

$$\int_0^{\infty}t^p C(t\ +\ \tfrac{1}{2})\,dt\ =\ \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\frac{4p!}{(2\sigma_m^2(\alpha_i^2\ +\ \alpha_j^2))^p}$$
$$\times\ \frac{f_i(x_0)\sin(\alpha_i)}{\alpha_i\ +\ \sin(\alpha_i)\cos(\alpha_i)}\frac{f_j(y_0)\sin(\alpha_j)}{\alpha_j\ +\ \sin(\alpha_j)\cos(\alpha_j)}.\quad\text{(A29)}$$

The expectation for the coalescence time then becomes

$$E[T]\ =\ \frac{1}{2}\ +\ \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\frac{2}{\sigma_m^2(\alpha_i^2\ +\ \alpha_j^2)}\frac{f_i(x_0)\sin(\alpha_i)}{\alpha_i\ +\ \sin(\alpha_i)\cos(\alpha_i)}\frac{f_j(y_0)\sin(\alpha_j)}{\alpha_j\ +\ \sin(\alpha_j)\cos(\alpha_j)}\quad\text{(A30)}$$

and the variance is

$$\mathrm{Var}[T]\ =\ \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\frac{2}{(\sigma_m^2(\alpha_i^2\ +\ \alpha_j^2))^2}\frac{f_i(x_0)\sin(\alpha_i)}{\alpha_i\ +\ \sin(\alpha_i)\cos(\alpha_i)}\frac{f_j(y_0)\sin(\alpha_j)}{\alpha_j\ +\ \sin(\alpha_j)\cos(\alpha_j)}$$
$$-\ \left(\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\frac{2}{\sigma_m^2(\alpha_i^2\ +\ \alpha_j^2)}\frac{f_i(x_0)\sin(\alpha_i)}{\alpha_i\ +\ \sin(\alpha_i)\cos(\alpha_i)}\frac{f_j(y_0)\sin(\alpha_j)}{\alpha_j\ +\ \sin(\alpha_j)\cos(\alpha_j)}\right)^2.\quad\text{(A31)}$$

Simplified form: If the two samples are taken from locations separated by more than the single-generation dispersal range, or if the neighborhood size is large, we can neglect the premigration correction introduced above and use a simplified form of the solution where the initial state of the system is the δ function $\delta(x - x_0, y - y_0)$. The cumulative distribution function, probability density function, and pth moment of the distribution of the time to coalescence are given by

$$1 - P_2(t) = 1 - 4\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\frac{\cos(\alpha_i x_0)\sin(\alpha_i)}{\alpha_i + \sin(\alpha_i)\cos(\alpha_i)}\frac{\cos(\alpha_j y_0)\sin(\alpha_j)}{\alpha_j + \sin(\alpha_j)\cos(\alpha_j)}e^{-2\sigma_m^2(\alpha_i^2+\alpha_j^2)t} \tag{A32}$$

$$C(t) = 8\sigma_m^2\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\frac{\cos(\alpha_i x_0)\sin(\alpha_i)}{\alpha_i + \sin(\alpha_i)\cos(\alpha_i)}\frac{\cos(\alpha_j y_0)\sin(\alpha_j)}{\alpha_j + \sin(\alpha_j)\cos(\alpha_j)} \times (\alpha_i^2 + \alpha_j^2)e^{-2\sigma_m^2(\alpha_i^2+\alpha_j^2)t} \tag{A33}$$

$$E[T^p] = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\frac{4p!}{(2\alpha_m^2(\sigma_i^2 + \alpha_j^2))^p}\frac{\cos(\alpha_i x_0)\sin(\alpha_i)}{\alpha_i + \sin(\alpha_i)\cos(\alpha_i)}\frac{\cos(\alpha_j y_0)\sin(\alpha_j)}{\alpha_j + \sin(\alpha_j)\cos(\alpha_j)}. \tag{A34}$$

High migration limit: Considering the strong-migration limit, where $N\sigma_m^2$ becomes large, we expect the model to converge on a panmictic population. As $N\sigma_m^2$ becomes large, α_1 approaches zero, and only the first term in Equations A32–A34 contributes significantly to the sum. Equation A33 thus simplifies to

$$E[T^p] \approx \frac{4p!}{(4\sigma_m^2\alpha_1^2)^p}\frac{\cos(\alpha_1 x_0)\sin(\alpha_1)}{\alpha_1 + \sin(\alpha_1)\cos(\alpha_1)}\frac{\cos(\alpha_1 y_0)\sin(\alpha_1)}{\alpha_1 + \sin(\alpha_1)\cos(\alpha_1)}. \tag{A35}$$

Furthermore, $\sin(\alpha_1)$ approaches α_1, and $\cos(\alpha_1 x_0)$ and $\cos(\alpha_1 y_0)$ both approach one:

$$E[T^p] \approx \frac{p!}{(4\sigma_m^2\alpha_1^2)^p}. \tag{A36}$$

Finally, using a series expansion for $\cot(\alpha_1)$, we find that, for large $N\sigma_m^2$,

$$\cot(\alpha_1) \approx \frac{1}{\alpha_1} \approx 4N\sigma_m^2\alpha_1 \tag{A37}$$

$$\alpha_1 \approx \frac{1}{\sqrt{4N\sigma_m^2}}, \tag{A38}$$

which gives us

$$E[T^p] \approx p!N^p, \tag{A39}$$

which is exactly what is expected for a panmictic population of size N.

Locations of the lineages: The probability distribution for the locations of two lineages that have not yet coalesced is fully described by $U(x, y, t)$. The transformed (triangular) state space is represented four times within the square space over which we have been considering the values of U. Thus, each combination of lineage locations (z_1, z_2) corresponds to four pairs of (x, y) coordinates. The likelihood at time t that the two lineages are at positions z_1 and z_2 is given by

$$p(z_1, z_2, t) = U(x, y, t) + U(y, x, t) + U(-x, -y, t) + U(-y, -x, t), \tag{A40}$$

where

$$x = 1 - |z_1 - z_2|, \quad y = z_1 + z_2 - 1. \tag{A41}$$

This probability distribution is not conditional on the two lineages still being separate. That is, the total probability over all (z_1, z_2) between zero and one will not integrate to one, but rather to the probability that the two lineages have not yet coalesced.

Locations of the coalescence events: The instantaneous rate of coalescence at a particular location is equivalent to the flux across the boundary at the point corresponding to that location. The probability of a coalescent event occurring at a particular location in the habitat, z_0, corresponds to flux at four locations in the transformed state space, one on each of the four sides, and the differential width in the habitat, δz, is twice the corresponding differential width (δx or δy) in the transformed space. The coalescence rate within the range z_0 to $z_0 + \delta z$ is

$$C(z_0, z_0 + \delta z, t) = 2\delta z(J_x(1, 2z_0 - 1, t) + J_y(2z_0 - 1, 1, t) - J_x(-1, 1 - 2z_0, t) - J_y(1 - 2z_0, -1, t)), \tag{A42}$$

where J_x and J_y are the flux rates in the x and y directions at a particular point, which is related to the slope of the distribution function at that point:

$$J_x = -2\sigma_m^2\frac{\partial U}{\partial x} \tag{A43a}$$

$$J_y = -2\sigma_m^2\frac{\partial U}{\partial y}. \tag{A43b}$$

The coalescence rate as a function of location is then

$$C(z_0, z_0 + \delta z, t) = 8\sigma_m^2\,\delta z\left(\sum_{i=1}^{\infty}\sum_{i'=1}^{\infty}\frac{\alpha_i^2\sin(\alpha_i)}{\alpha_i + \sin(\alpha_i)\cos(\alpha_i)}\frac{\alpha_{i'}\cos(\alpha_{i'}(2z_0 - 1))}{\alpha_{i'} + \sin(\alpha_{i'})\cos(\alpha_{i'})} \times (f_i(x_0)f_{i'}(y_0) + f_i(y_0)f_{i'}(x_0))e^{-2\sigma_m^2(\alpha_i^2+\alpha_{i'}^2)(t-1/2)} - \sum_{j=1}^{\infty}\sum_{j'=1}^{\infty}\frac{\alpha_j^{*2}\cos(\alpha_j^*)}{\alpha_j^* - \sin(\alpha_j^*)\cos(\alpha_j^*)}\frac{\alpha_{j'}^*\sin(\alpha_{j'}^*(2z_0 - 1))}{\alpha_{j'}^* - \sin(\alpha_{j'}^*)\cos(\alpha_{j'}^*)} \times (f_j^*(x_0)f_{j'}^*(y_0) + f_j^*(y_0)f_{j'}^*(x_0))e^{-2\sigma_m^2(\alpha_j^{*2}+\alpha_{j'}^{*2})t-1/2)}\right). \tag{A44}$$

The total probability of the coalescent event occurring at a particular location is derived by integrating the coalescence rate over time,

$$C(z_0, z_0 + \partial z) = \int_0^{\infty} C(z_0, z_0 + \partial z, t)\,dt, \tag{A45}$$

which gives

$$C(z_0, z_0 + \delta z) = 2\delta z\left(\sum_{i=1}^{\infty}\sum_{i'=1}^{\infty}\frac{2\alpha_i^2\sin(\alpha_i)}{\alpha_i + \sin(\alpha_i)\cos(\alpha_i)}\frac{\alpha_{i'}\cos(\alpha_{i'}(2z_0 - 1))}{\alpha_{i'} + \sin(\alpha_{i'})\cos(\alpha_{i'})}\right.$$

$$\times \frac{f_i(x_0) f_{i'}(y_0) + f_i(y_0) f_{i'}(x_0)}{(\alpha_i^2 + \alpha_{i'}^2)}$$

$$- \sum_{j=1}^{\infty} \sum_{j'=1}^{\infty} \frac{2\alpha_j^{*2} \cos(\alpha_j^*)}{\alpha_j^* - \sin(\alpha_j^*)\cos(\alpha_j^*)} \frac{\alpha_{j'}^* \sin(\alpha_{j'}^*(2z_0 - 1))}{\alpha_{j'}^* - \sin(\alpha_{j'}^*)\cos(\alpha_{j'}^*)}$$

$$\times \frac{f_j^*(x_0) f_{j'}^*(y_0) + f_j^*(y_0) f_{j'}^*(x_0)}{(\alpha_j^{*2} + \alpha_{j'}^{*2})}\Bigg). \quad \text{(A46)}$$

Application to the discrete-demes stepping-stone model: This solution can be applied approximately to a finite-length stepping-stone model of population structure by treating the D demes as part of a continuous population. Assume a total population size of N (deme population size of N/D) and migration rate m (a fraction $m/2$ of a deme's population arrives from each of its two neighbors each generation). The distance between adjacent demes (from center to center), scaling the total population from 0 to 1 as in the continuous case, is $1/D$, and the migration variance is m/D^2, so σ_m^2 equals $m/(2D^2)$. The initial conditions are best represented as sampling sequences from a uniform distribution over the range of the deme. Two sequences from demes d_1 and d_2 (where $1 \le d_1, d_2 \le D$) are assumed to be uniformly distributed over the deme range:

$$P(z_1^0) = \frac{1}{D}, \quad \frac{d_1 - 1}{D} < z_1^0 < \frac{d_1}{D}, \quad P(z_1^0) = 0 \text{ elsewhere.} \quad \text{(A47)}$$

The initial distribution in the transformed (x, y) state space can be approximated by triangular functions in x and y. For $d_1 \neq d_2$, these functions are

$$P_x^0(x) = D\left(x - \left(1 - \frac{|d_1 - d_2| + 1}{D}\right)\right),$$

$$1 - \frac{|d_1 - d_2| + 1}{D} < x < 1 - \frac{|d_1 - d_2|}{D},$$

$$P_x^0(x) = D\left(1 - \frac{|d_1 - d_2| - 1}{D} - x\right),$$

$$1 - \frac{|d_1 - d_2|}{D} < x < 1 - \frac{|d_1 - d_2| - 1}{D},$$

$$P_x^0(x) = 0 \text{ elsewhere,} \quad \text{(A48a)}$$

$$P_y^0(y) = D\left(y - \left(\frac{d_1 + d_2 - 2}{D} - 1\right)\right),$$

$$\frac{d_1 + d_2 - 2}{D} - 1 < y < \frac{d_1 + d_2 - 1}{D} - 1,$$

$$P_y^0(y) = D\left(\frac{d_1 + d_2}{D} - 1 - y\right),$$

$$\frac{d_1 + d_2 - 1}{D} - 1 < y < \frac{d_1 + d_2}{D} - 1,$$

$$P_y^0(y) = 0 \text{ elsewhere.} \quad \text{(A48b)}$$

For $d_1 = d_2 = d$, we must factor the reflecting boundary into the distribution for x_0,

$$P_x^0(x) = 2D\left(x - \left(1 - \frac{1}{D}\right)\right), \quad 1 - \frac{1}{D} < x < 1,$$

$$P_x^0(x) = 0 \text{ elsewhere.} \quad \text{(A49)}$$

The expression for y_0 has the same form as in A48a when $d_1 = d_2$.

The functions $f_i(x_0)$ and $f_i^*(x_0)$ are derived by integration as

$$f_i(x_0) = \int_{-1}^{1} \cos(\alpha_i x) P_x^0(x)\, dx \quad \text{(A50a)}$$

$$f_i^*(x_0) = \int_{-1}^{1} \sin(\alpha_i^* x) P_x^0(x)\, dx, \quad \text{(A50b)}$$

where we use the large-neighborhood-size approximation to determine the α values:

$$\cot(\alpha_i) = \frac{2Nm\alpha_i}{D^2} \quad \text{(A51a)}$$

$$-\tan(\alpha_j^*) = \frac{2Nm\alpha_j^*}{D^2}. \quad \text{(A51b)}$$

The expressions for $f_i(y_0)$ and $f_i^*(y_0)$ are exactly analogous. These values can then be used to approximate the distribution of the time to coalescence using Equations A32–A34.