Application of a Random Walk Model to Geographic Distributions of Animal Mitochondrial DNA Variation

Joseph E. Neigel* and John C. Avise†

**Department of Biology, University of Southwestern Louisiana, Lafayette, Louisiana 70504, and †Department of Genetics, University of Georgia, Athens, Georgia 30602*

Manuscript received July 8, 1993
Accepted for publication August 18, 1993

ABSTRACT

In rapidly evolving molecules, such as animal mitochondrial DNA, mutations that delineate specific lineages may not be dispersed at sufficient rates to attain an equilibrium between genetic drift and gene flow. Here we predict conditions that lead to nonequilibrium geographic distributions of mtDNA lineages, test the robustness of these predictions and examine mtDNA data sets for consistency with our model. Under a simple isolation by distance model, the variance of an mtDNA lineage's geographic distribution is expected be proportional to its age. Simulation results indicated that this relationship is fairly robust. Analysis of mtDNA data from natural populations revealed three qualitative distributional patterns: (1) significant departure of lineage structure from equilibrium geographic distributions, a pattern exhibited in three rodent species with limited dispersal; (2) nonsignificant departure from equilibrium expectations, exhibited by two avian and two marine fish species with potentials for relatively long-distance dispersal; and (3) a progression from nonequilibrium distributions for younger lineages to equilibrium distributions for older lineages, a condition displayed by one surveyed avian species. These results demonstrate the advantages of considering mutation and genealogy in the interpretation of mtDNA geographic variation.

GENETIC drift, migration, mutation and selection all have the potential to influence the geographic distributions of animal mitochondrial (mt) DNA lineages. In addition, contemporary distributions may reflect historical patterns of population subdivision that can no longer be observed directly. With this diversity of possible influences, observed distributions may be compatible with several alternative explanations. However, even minimal explanations of geographic variation must acknowledge mutation and dispersal. The existence of variation establishes that mutation has occurred, and the presence of individuals in more than one location is a demonstration of dispersal. Thus, distributions of mtDNA lineages that would be generated from mutation and dispersal alone may be viewed as the hypotheses against which other, more elaborate explanations are tested. A general scenario is illustrated in Figure 1. Because each unique mtDNA mutation demarcates a lineage descended from a single female, it follows that the geographic distribution of a lineage represents the result of dispersal from a single place of origin. Initially, a lineage will be concentrated around this point of origin, but later generations may disperse the lineage from this point until it is fully distributed throughout the available geographic range of the species (if extinction does not occur first).

This view of lineage expansion is suggested by maps of the geographic distributions of animal mtDNA variants. For example, mtDNA variation in the deer mouse *Peromyscus maniculatus* displays a phylogenetic hierarchy that appears to correspond to a geographic hierarchy (LANSMAN *et al.* 1983): younger lineages occupy more restricted ranges than older lineages. Many other species similarly exhibit hierarchical associations between mtDNA phylogeny and geographic range (AVISE *et al.* 1987; MORITZ, DOWLING and BROWN 1987). This suggests a general model in which the geographic distributions of surviving lineages broaden over time.

NEIGEL, BALL and AVISE (1991) introduced a random walk model for the estimation of single generation migration distances from geographic variation in mtDNA. For this model it was assumed that each mtDNA lineage originates at a unique geographic location and spreads outward at a rate dependent on dispersal distances, as defined by the distances along a single dimension between birth sites of mothers and daughters. The relevant parameter is the variance of the geographic positions of the individuals that comprise the lineage, which we abbreviated as σ^2_H. This parameter is a measure of dispersion and is not dependent on the size of a lineage in terms of numbers of individuals (thus, in principle, a rare lineage might nonetheless have a very broad geographic distribution and high σ^2_H, whereas a numerically large lineage could be restricted to a single locale and display a low σ^2_H). Furthermore, the actual distribution of single

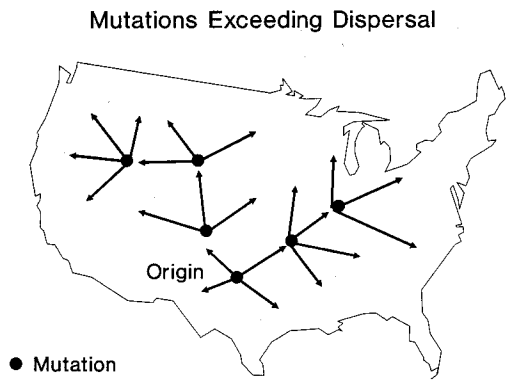

FIGURE 1.—Conceptual model of the branching and geographic expansion of mtDNA lineages.

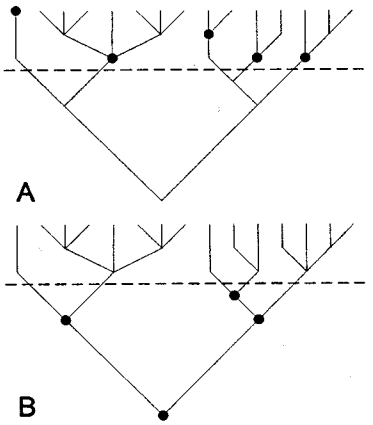

FIGURE 2.—Two methods for selecting lineages from a phylogeny. The method illustrated by (A) yields nonnested lineages. Each lineage includes all individuals that share a common ancestor more recently than the time indicated by the dashed line. The ages of these lineages are defined by the time since the common ancestors indicated by filled circles (the coalescence time). The method illustrated by (B) yields nested lineages. All lineages that correspond to nodes below the time indicated by the dashed line are selected. The ages of these lineages are also defined by their coalescence times.

generation dispersal distances is not critical. For example, the overall distribution may include both interpopulation and intrapopulation movements. The model was intended to focus on the processes that determine the locations and geographic distributions of specific lineages (mutation and dispersal) rather than the processes that govern their relative frequencies (drift and selection). Here we extend this model to examine the conditions under which nonequilibrium distributions of mtDNA lineages may be predicted, and test these predictions with published data.

SIMULATION STUDIES

The model: Under our basic random walk model, the expected σ^2_H for a lineage is proportional to its age. However, this relationship could be altered by barriers to dispersal, alterations in population demography and other forces that would violate the assumptions of the model. To explore the effects of some of these departures from the model, we have employed simulations that allow us to examine the robustness of the relationship between lineage age and geographic variance. Although this approach does not lead to general formulations for each effect, it does allow us to consider specific situations that would be difficult to treat analytically. Thus, we have examined effects of population size, density regulation and range limits on the relationship between the geographic variance of a lineage and its age. These simulations were based on a model presented earlier (NEIGEL, BALL and AVISE 1991), with some modifications. Individuals occupied a two-dimensional geographic space, with both x and y axes divided into 10,000 increments. Generations were nonoverlapping, and only females were represented in the population. The number of progeny for each individual followed a Poisson distribution, with a mean determined by some form of population regulation (see below). The geographic locations of the progeny along each axis were sampled from a discrete approximation of a normal distribution centered on the location of the parent, with these distributions truncated at the limits of the geographic space. A representation of the phylogenetic relationships within the population was updated each generation, so that all values for lineage ages were exact. Thus, these simulations were not subject to errors associated with the estimation of phylogenies. Each simulation was begun with an initial set of individuals of undefined phylogenetic relationships. Simulations were run for a sufficient number of generations to insure that in the final generation, all individuals were descended from just one of the initial founders. For the analysis of simulation results, lineages were selected by either of two methods, illustrated in Figure 2. By the first method, lineages below a specified age were selected, and once selected, all sublineages within them were excluded from selection. This method produces lineages that may be regarded as independent with respect to ancestry, but limits the number of lineages that can be selected from each genealogy, especially if older lineages are included. Lineages selected by this method will be referred to as "nonnested." For the second method, all lineages above a specified age were selected. Although this method can yield a large number of lineages, the selected lineages are phylogenetically nested, and so cannot be treated as independent samples. Lineages selected by this method will be referred to as "nested."

Analysis of simulations: After a specified number of generations, the geographic distributions of lineages within the population were analyzed. The geographic variance (σ^2_H) of each lineage's distribution along the x-coordinate was calculated, along with the

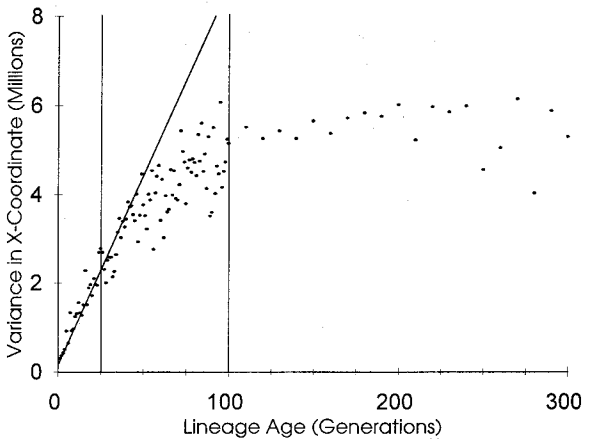

FIGURE 3.—Simulation illustrating three stages of lineage dispersal: 1,000 replicates were run of a population initiated at the carrying capacity of 100 females, a standard dispersal distance of 500 units within a range of 10,000. Each point represents an average of either all lineages of a specific age (lineages less than 100 generations old) or all lineages within an age class that spans 10 generations (lineagesover 100 generations old). Transitions betweenstages are indicated, as is the initial slope during stage 1.

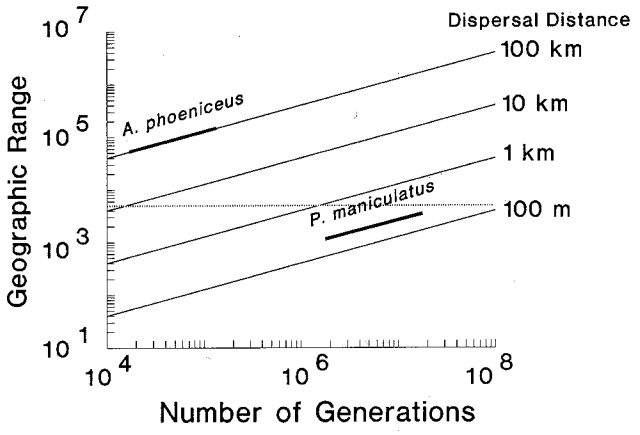

FIGURE 4.—Expected geographic ranges for lineages dispersed by a multigeneration random walk process. The expected geographic range of a lineage (defined as the boundaries that enclose 95% of the expected geographic distribution) is given on the vertical axis for lineages that have dispersed from a single point of origin for the number of generations shown on the horizontal axis, with standard single generation dispersal distances of 100 m, 1 km, 10 km or 100 km. Expected geographic ranges are also given for the red-winged blackbird and the deer mouse. The longitudinal width of the North American continent is shown at 4×10^3 km.

ratio of this variance to the lineage's age, σ^2_{Xt}, and the square root of this ratio, σ_{Xt}. One of the goals of the study was to evaluate the suitability of σ_{Xt} as an estimate of the standard deviation of single generation dispersal distances, which we refer to as the standard dispersal distance. Simulations that were run under different conditions were tested for differences in the distribution of σ^2_{Xt} for age-matched, nonnested lineages by Wilcoxon rank-sum tests. Simulations and data analysis programs were written in C++, compiled with BORLANDC++ 3.1, and run on a Gateway 2000 486/33E microcomputer. Details of the programs are available from JEN.

Conditions for nonequilibrium distributions: Without constraints, the expected variance of a lineage dispersed by a random walk process would increase at a linear rate, so that σ_{Xt} would be independent of lineage age. In reality, the dispersal of a lineage must eventually be constrained by the geographic limits that correspond to the range of the species. As those limits are approached, the rate of increase in σ^2_H should decline, until a maximum, equilibrium level is reached. This effect is illustrated by the simulation results shown in Figure 3, which suggests three stages of lineage dispersal:

Stage 1: For sufficiently young lineages, geographic constraints have not been encountered, and σ_{Xt} is independent of lineage age.

Stage 2: For lineages of intermediate ages, dispersal continues, but is limited by geographic constraints. There is a positive, but nonlinear correlation between σ^2_H and lineage age, so that σ_{Xt} declines with older lineages.

Stage 3: For sufficiently old lineages, an equilibrium between genetic drift and migration has been reached. There is no correlation between σ^2_H and lineage age, so that σ_{Xt} declines with lineage age.

For a given taxon, the time scales over which these stages would be reached can be estimated by comparison of the geographic range of the taxon with expected random walk distributions. Although random walk distributions are expected to approach a Gaussian form without defined spatial boundaries, we can consider the spatial limits that enclose 95% of a random walk distribution as a practical definition of its geographic boundaries. These 95% limits should span about four standard deviations. Thus, the transition from stage 1 to stage 2 might be expected to occur about when the standard deviation of a lineage's distribution along a geographic axis approaches one-fourth of the maximum available range along the same coordinate.

Figure 4 shows expected geographic ranges for lineages dispersed by a multigeneration random walk process. The expected geographic range of a lineage (defined as the boundaries that enclose 95% of the expected geographic distribution) is given on the vertical axis for lineages that have dispersed from a single point of origin for the number of generations shown on the horizontal axis, with standard single generation dispersal distances of 100 m, 1 km, 10 km or 100 km. Specific cases for the deer mouse (*P. maniculatus*) and the red-winged blackbird (*Agelaius phoeniceus*) are also shown on this graph. Both species range across North America, a longitudinal distance of about 4,000 km. Mark-and-recapture estimates correspond to standard dispersal distances of about 200 m for *P. maniculatus* (BLAIR 1940; STICKEL 1968; DICE and HOWARD 1951) and 100 km for *A. phoeniceus* (MOORE and DOLBEER

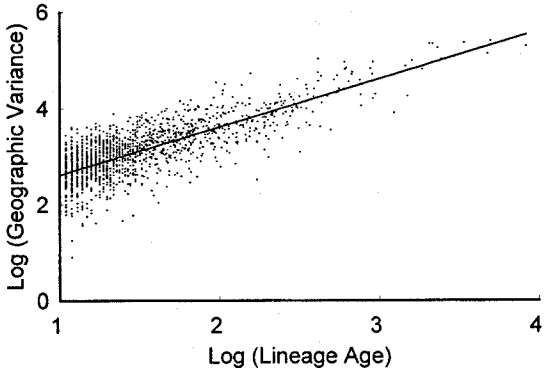

FIGURE 5.—Variance of geographic distribution along x axis *vs.* lineage age for a simulation of nonequilibrium conditions. The population was initiated at its carrying capacity of 10,000 females, and the standard dispersal distance was 10 units within a range of 10,000. Each point corresponds to an individual node representing a lineage of at least 10 generations in age with at least two members. A logarithmic scale for lineage age is used in this plot for visual clarity.

1989). Based on an mtDNA sequence evolution rate of 1% per million years and generation time estimates of 0.2 years for *P. maniculatus* (LAYNE 1968) and 3 years for *A. phoeniceus*, the lineages in these analyses range in age from 1.75 million to 17.5 million generations for *P. maniculatus* (LANSMAN *et al.* 1983) and from 17,000–130,000 generations for *A. phoeniceus* (BALL *et al.* 1988). These combinations of mark and recapture dispersal distances and mtDNA lineage ages on Figure 4 show that the expected geographic ranges for the *P. maniculatus* mtDNA lineages fall just below the 4,000-km continental limit (shown by a horizontal line), while even the youngest mtDNA lineages of *A. phoeniceus* are expected to be well above this limit. From this assessment we can predict that *P. maniculatus* lineages should exhibit stage 1 geographic distributions, whereas *A. phoeniceus* should exhibit stage 3 distributions.

In practice, the transitions between these stages of lineage dispersal will be gradual, and must be defined arbitrarily. The equilibrium distribution of stage 3 will be approached asymptotically (Figure 3), and so must be defined with respect to the resolving power of the data. For our present purposes, we considered the null hypothesis to be an equilibrium stage 3 distribution, and we tested for departures from that distribution.

Linearity of age-variance relationship: If a decline in σ_{Xt} for older lineages is interpreted as an effect of range limits, as suggested above, then it must be assumed that in the absence of range limits, σ_{Xt} is independent of lineage age. This assumption was tested in a simulation with lineages that spanned several orders of magnitude in age, but which were all expected to be within stage 1 of the dispersal process. Figure 5 is a scatter plot of σ^2_H, the variance in position of the members of a lineage, *vs.* lineage age for nested

TABLE 1

Comparison of distributions of σ_{Xt} between lineage age groups

	Young	Old
Minimum age	1	115
Maximum age	99	949
Mean age	47	574
Number of lineages	202	15
Mean σ_{Xt}	7.07	7.23
Standard deviation σ_{Xt}	2.95	2.77
Standard error σ_{Xt}	0.21	0.71

Wilcoxon rank-sum test $P = 0.46$.

lineages. Details of the simulation are given in the legend. A least-squares regression line, with the intercept forced to zero, is drawn through these points. No departures from linearity are evident. The σ^2_H data were also converted to σ_{Xt} values and regressed on lineage age. As expected, no age related trend was evident: the least-squares line was essentially horizontal (slope = −0.00023). However, because the lineages were nested, these data could not be used for statistical tests that assume independence. Therefore, two age ranges of nonnested lineages were used to test the hypothesis that σ_{Xt} is independent of lineage age. The lower range was from 1–99 generations with a mean of 47. The upper range was from 115–949 generations with a mean age of 574. The distributions of σ_{Xt} are very similar as shown in Table 1, and were not different at the 0.05 level in a Wilcoxon rank-sum test. Thus, over this range of lineage ages, which should all be within stage 1, we saw no evidence of a dependence of σ_{Xt} on lineage age.

Population size effects: If the dispersal and growth of each maternal lineage occurred independently of other lineages, the relationship between lineage age and geographic variance would be independent of population size. However, if either demography or dispersal were influenced by population size (or density), this influence could extend to the relationship between lineage age and geographic variance. To assess the sensitivity of our model to population size effects, we used a series of simulations that differed in both population size and the mechanism by which population size was regulated. For simulation of global population regulation, the total population size, N, was regulated about a specified carrying capacity K, by setting the expected fecundity of every individual to K/N each generation. Because the distribution of lineage ages has been shown to be dependent on population size (AVISE, Neigel and ARNOLD 1984), only relatively young lineages can be used for paired comparisons with small populations. Table 2 shows the results of a comparison of populations globally regulated at carrying capacities of 100, 1,000, 5,000 and 10,000, with a standard dispersal distance of nine geographic units. In every case σ_{Xt} underestimated the

TABLE 2

Effect of carrying capacity on σ_{Xt}

K_1	K_2	np	σ_{Xt1}	σ_{Xt2}	P
100	1,000	387	7.3	7.5	0.271
100	5,000	326	7.2	7.6	0.230
100	10,000	310	7.2	7.9	0.007
1,000	5,000	319	7.4	7.6	0.920
1,000	10,000	303	7.4	7.9	0.089
5,000	10,000	325	7.6	8.0	0.057

Simulations are compared in pairs by the Wilcoxon signed-rank test. Parameters for either member of a pair are denoted by a subscript of 1 or 2. K is the population carrying capacity, np is the number of pairs of non-nested lineages compared, σ_{Xt} is the standard dispersal distance. The actual standard deviation of dispersal distances, σ_x, was 9 in every case. P is the probability for a Wilcoxon signed-rank test for lineages paired by age.

actual standard dispersal distance by 10–20%. This bias has been noted before and was interpreted as an effect of the covariance in the geographic positions of those individuals within a lineage that share a more recent common ancestor than the single ancestor of the entire lineage (NEIGEL, BALL and AVISE 1991). Paired comparisons of nonnested lineages up to 50 generations in age revealed that σ_{Xt} was closer to the actual standard dispersal distance in simulations of larger populations. However, differences in σ_{Xt} between runs were generally small, and only a comparison of simulations with the smallest carrying capacity ($K = 100$) and the largest ($K = 10{,}000$) was found to be significantly different in a Wilcoxon signed-rank test with α set at 0.05.

Local density regulation: For species with limited dispersal, it is unlikely that population regulation would operate on a global scale that greatly exceeded typical dispersal distances. We therefore examined simulations in which the expected fecundity of every individual was based on the density of individuals within a surrounding zone, rather than total population size. In this local regulation model, each individual was at the center of its own zone, so that the zones of nearby individuals overlapped. The local carrying capacity (the number expected within a zone-sized area) was calculated as $K_Z = (K \times A_Z)/A_T$, where K is the entire population's carrying capacity, A_Z is the area within a zone and A_T is the total geographic space (10^8 units). The individual's expected fecundity was calculated as K_Z/N_Z, where N_Z is the actual number of individuals within the zone. With this form of local regulation, total population size was often regulated considerably below the entire population's carrying capacity. In extreme cases, in which dispersal was not sufficient to distribute individuals away from crowded areas, the entire population would decline to extinction. Thus, it was necessary to use combinations of standard dispersal distances, zone widths and local carrying capacities that would sustain populations.

TABLE 3

Effect of local density regulation on σ_{Xt}

K_L	K_G	N_L	N_G	σ_x	w	np	σ_{XtL}	σ_{XtG}	P
1000	1000	997	1164	9	500	92	7.3	7.0	0.057
5000	5000	4769	4984	93	100	147	77.1	76.1	0.549
5000	2767	2767	2758	93	100	79	75.3	77.4	0.764

Parameters for simulations with local population regulation are subscripted with L, for corresponding simulations with global regulation, with G. K is the population carrying capacity, N is the mean population size, σ_x is the actual standard deviation of dispersal distances, w is the width of the zone for local regulation, np is the number of pairs of lineages compared, σ_{Xt} is the standard dispersal distance, P is the probability for a Wilcoxon signed-rank test for nonnested lineages paired by age.

Table 3 compares the distributions of σ_{Xt} for simulations that were paired so that they differed in the mode of population regulation (global or local) but were alike in most other respects. In general, the distributions of σ_{Xt} within each pair appear to be similar, and none were significantly different by Wilcoxon signed-rank tests with α set at 0.05.

Cyclic range contraction: The most complex form of regulation that we considered was intended to simulate gradual range contraction and expansion, such as those which occurred during Pleistocene climate cycles (GRAHAM 1986). This process is illustrated in Figure 6. The population was initiated at its carrying capacity of 10,000 females and the standard dispersal distance was 186 units. Each individual was also subject to local density regulation within 200 units. During the course of the simulation, five cycles of range contraction and expansion occurred, with the simulation beginning and ending during phases of maximum range expansion. During maximum contraction, the range was reduced to 20% of the area available during maximum expansion. To simulate a refugium, the fecundity of individuals with y-coordinates less than 1,000 were subject only to normal local density regulation at all times. Figure 7 is a scatter plot of σ^2_H *vs.* lineage age. A least-squares regression line, with the intercept forced through zero, is drawn through these points. No departures from linearity are evident. Furthermore, there was no significant difference in the distribution of σ_{Xt} in a paired comparison with lineages from a comparable simulation without cycles (158 paired lineages, Wilcoxon signed rank test, $P = 0.92$)

METHODS FOR ANALYSIS OF MITOCHONDRIAL DNA DATA

Tests for nonequilibrium distributions: The form of lineage expansion predicted by our model suggests tests for departures from equilibria between genetic drift and gene flow. For a nonequilibrium distribution, it should be possible to demonstrate both geo-

TABLE 4

Data sets analyzed for this study

Species and common name	Individuals	Haplotypes	Localities	Reference
Peromyscus maniculatus Deer mouse	136	61	41	LANSMAN *et al.* (1983)
Peromyscus polionotus Old-field mouse	68	22	15	AVISE *et al.* (1983)
Geomys pinetis Pocket gopher	87	23	26	AVISE *et al.* (1979)
Agelaius phoeniceus Red-winged blackbird	127	34	20	BALL *et al.* (1988)
Quiscalus quiscula Common grackle	35	29	8	ZINK, ROOTES and DITTMANN (1991)
Melospiza melodia Song sparrow	170	58	29	ZINK (1991)
Anguilla rostrata American eel	108	21	7	AVISE *et al.* (1986)
Arius felis Hard-head catfish	60	11	11	AVISE, REEB and SAUNDERS (1987)
Crassostrea virginica American oyster	212	82	14	REEB and AVISE (1990)

TABLE 5

PHYFORM estimates of dispersal distances

Species	Generation time (yr)	Median lineage age (yr × 10^3)	Dispersal distance (km) Longitude axis	Latitude axis
Peromyscus maniculatus	0.2	200	0.14	0.10
Peromyscus polionotus	0.2	225	0.06	0.02
Geomys pinetis	1	150	0.18	0.12
Agelaius phoeniceus	3	125	3.2	3.2
Quiscalus quiscula	3	25	7.6	6.0
Melospiza melodia	2	75	6.1	3.6
			Along shoreline	
Anguilla rostrata	10	175	11	
Arius felis	2	175	2.8	
Crassostrea virginica	1	100	3.3	

TABLE 6

Correlation coefficients for estimates of lineage ages *vs.* lineage geographic variances

Species	Linear correlation Longitude	Latitude	Kendall's tau Longitude	Latitude
Peromyscus maniculatus	0.78	0.92	0.78*	0.78
Peromyscus polionotus	0.98	0.97	0.87*	1.0*
Geomys pinetis	0.94	0.70	0.71*	0.62
Agelaius phoeniceus	0.23	0.53	0.07	0.43
Quiscalus quiscula	0.23	0.81	0.05	0.71
Melospiza melodia	0.08	0.67	0.11	0.67
	Along shore		Along shore	
Anguilla rostrata	−0.70		−0.52	
Arius felis	0.47		0.4	
Crassostrea virginica	0.88		0.60	

* Indicates values of τ_K that exceeded 95% of the distribution of τ_K for 100 replicates in which the data were geographically randomized data over the sample locales.

graphic clustering of mtDNA variants and a positive correlation between estimates of σ^2_H and lineage age. To apply these criteria with the statistical procedures described earlier (NEIGEL, BALL and AVISE 1991), it was necessary to provide suitable null models for comparison. Because the geographic range of any species is finite, a cluster defined by range limits will always be present and must be distinguished from clustering of lineages within those limits. In addition, biases may be inherent in our estimation procedures that would produce apparent correlations between the σ^2_H of lineage distributions and their ages. For these reasons, we have compared analyses of each data set with replicate analyses in which the geographic coordinates of individuals were randomly exchanged.

A linear relationship between σ^2_H and lineage age is expected during stage 1 if the standard dispersal distance is constant over time. However, since the time scales relevant to the history of mtDNA lineages generally span periods of major climatic change, departures from linearity are anticipated. To remove assumptions of linearity we used rank correlations in addition to linear correlations to detect positive correlations between σ^2_H and lineage age.

Data and methods of analysis: The data sets analyzed were from published studies of mtDNA restriction fragment length polymorphisms (RFLPs) in nine species native to North America (Table 4). Percent sequence divergence between mtDNA haplotypes was estimated by the method of NEI and LI (1979), with a computer program written by J. NEIGEL. Phylogenetic trees were estimated by the UPGMA method,

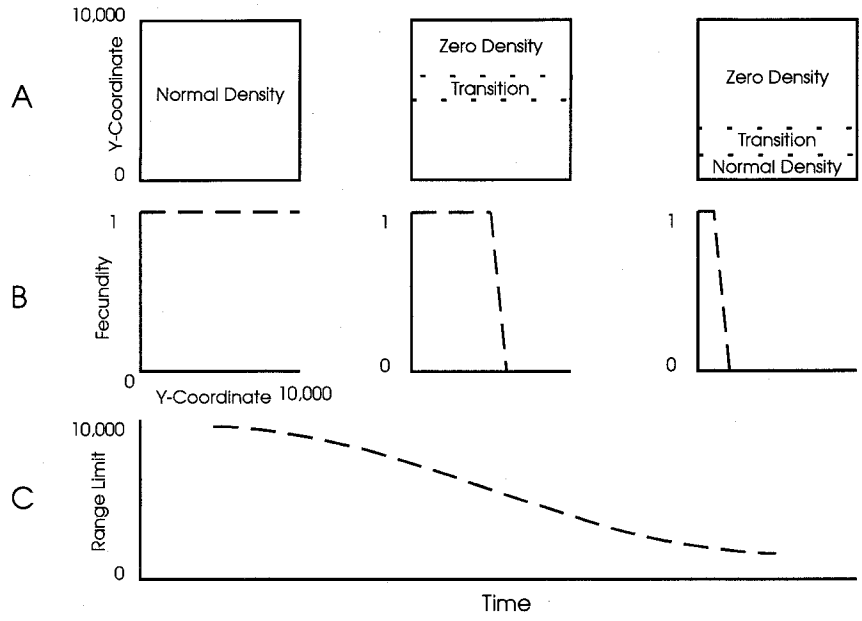

FIGURE 6.—Simulation of cyclical range contraction. Part A represents a map of the geographic space at three points during the first half of a cycle; a total of five cycles occurred during the simulation. As each cycle progressed, the range limit moved along the y-coordinate as shown. Between the region of normal population density and the region of zero density there was a transition zone in which expected fecundity decreased linearly. In part B, graphs of expected fecundity *vs.* the y-coordinate are shown below the maps. Part C graphs the sinusoidal movement of the range limit as a function of time.

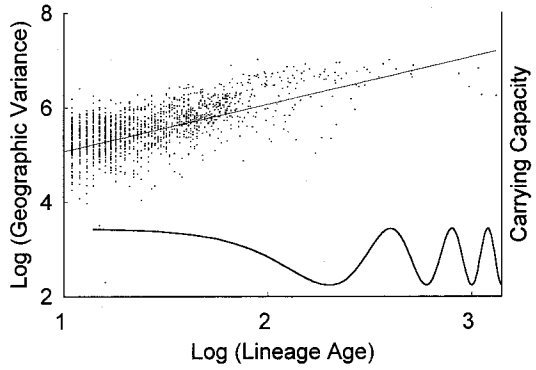

FIGURE 7.—Variance of geographic distribution along y axis *vs.* lineage age for a simulation of cyclical range contraction. The details of the simulation are given in Figure 6. A logarithmic scale for lineage age is used in this plot for visual clarity.

with the NTSYS numerical taxonomy package (ROHLF 1988). These data, with geographic coordinates of collection locales, were analyzed by the computer program PHYFORM (available from J. NEIGEL). Lineages were divided into ten age classes of equal width and σ^2_H was calculated for composite lineages formed from each age class. This is a slight modification of PHYFORM as described previously (NEIGEL, BALL and AVISE 1991), in which lineages of the same exact age (as estimated by the UPGMA analysis) were overlaid for calculations of σ^2_H, with subsequent averaging of these σ^2_H values among lineages within the same age class. This present modification was intended to provide a consistent procedure for lumping lineages of different ages among data sets that differed in the extent to which the ages of these lineages were resolved. Data sets with empty age classes produced fewer than 10 estimates of σ^2_H.

For all taxa analyzed in this study, it was assumed that mtDNA sequences evolve at a rate of 1% per million years, which is the equivalent of 2% divergence between lineages per million years. For those portions of our analysis that did not require the conversion of sequence divergence to units of time, we have used the upper limit of per cent sequence divergence within a lineage to indicate its relative age.

To test for clustering, the geographic distributions of mtDNA lineages were compared with distributions that would result if individuals were randomly placed among the collecting locales. A modification of the PHYFORM program was used to make these comparisons. The modified program (RANDFORM) sampled the geographic locations of all individuals in a data set, then randomly shuffled these locations among individuals. This randomization provided data sets with the same phylogenetic and geographic distributions as the original, but without any correlation between these distributions. The estimated σ^2_H for each lineage age class in the original data set was then compared with the distribution of σ^2_H for 100 replicate randomized data sets. We refer to the range from the fifth lowest to the fifth highest of these σ^2_H values as the nonsignificant range. Only σ^2_H values below the nonsignificant range were considered to represent significant geographic clustering. To test for positive correlations between lineage age and σ^2_H, values of the linear correlation coefficient (r^2) and the nonparametric rank order correlation statistic, Kendall's tau (τ_K) (BHATTACHARYYA and JOHNSON 1977), were calculated for each data set. Because estimates of σ^2_H for different age classes are not independent, the significance of an apparently high value of τ_K was further evaluated by comparison with the distribution of τ_K values for 100 replicates of geographically randomized data.

RESULTS OF mtDNA ANALYSIS

Estimated dispersal distances: Estimates of standard dispersal distance derived from PHYFORM analysis of mtDNA data and the generation times that these estimates were based upon are listed in Table 5. For the three rodent taxa, all of which appear to meet the criteria of nonequilibrium distributions, these estimates can be directly compared with other measures of dispersal, such as mark and recapture estimates. We have shown previously that an estimate of dispersal distance for *P. maniculatus* calculated by PHYFORM from mtDNA data is within the range of mark and recapture estimates for this species (NEIGEL, BALL and AVISE 1991). For the other taxa, which do not meet one or both of our criteria for nonequilibrium distributions, these estimates may still be usefully interpreted as lower bounds on actual standard dispersal distances. In this context, it is interesting that our estimates of minimum single generation dispersal distances for *A. phoeniceus* and *Quiscalus quiscula* are below those obtained from an analysis by MOORE and DOLBEER (1989) of continent-wide USFWS band-and-recover studies, whereas our estimate for *Melospiza melodia* exceeds the estimate obtained from a finite-area band-and-recover study by BARROWCLOUGH (1980).

Tests for equilibrium lineage distributions: Correlation coefficients between estimates of σ^2_H and lineage age are listed in Table 6. High positive linear correlations (*i.e.*, greater than 0.7), such as those observed for *Peromyscus polionotus*, suggest that dispersal of lineages has not been constrained by geographic boundaries. These high correlations are consistent with expectations for stage 1 lineage distributions. Low linear correlation coefficients (*i.e.*, less than 0.5), such as those observed for *A. phoeniceus* and *Q. quiscula*, suggest that the distributions of lineages are effectively at equilibrium within the geographic range sampled, as expected for stage 3 distributions. For several taxa, moderate positive linear correlations (between 0.5 and 0.7) were observed for one or both geographic axes. The value of τ_K was used to examine the significance of nonlinear rank order correlations. For those values of τ_K indicated as significant, less than 5% of the values calculated from geographically randomized data equaled or exceeded these observed values. These correlations are thus above those that could be explained from any inherent bias in the statistical procedures employed. Based on correlation statistics alone, the null hypothesis of equilibrium distributions (stage 3) can be rejected only for the rodent taxa.

Figures 8-10 show the relationships between estimates of σ^2_H and mtDNA lineage age (shown as percent sequence divergence) for all nine taxa, along with the means and limits for the nonsignificant range of σ^2_H values generated from geographically randomized data. For the six terrestrial species, longitude and latitude are shown separately. For the three marine species, only a single dimension is shown, which corresponds to position along a linear coastline. Because the selected data sets differed considerably in the ranges of sequence divergence and σ^2_H values, it was not practical to graph all of them with a common scale. However, the same scale was used for longitude and latitude plots of each species, and a single scale was used for all three species of birds.

Some general patterns can be seen among these plots. First, there are typically higher values of σ^2_H for longitude than latitude. This was an expected consequence of the shapes of the geographic ranges of these species, which are generally broader in longitude than in latitude. Usually, values of σ^2_H appear to either remain constant or increase with sequence divergence, which are the two alternative outcomes expected from our model. Some plots are more "noisy" than others, as expected from differences in sample size and associated sampling errors. However as shown in Table 1, the data set for *M. melodia*, which produced what appears to be a random scatter of σ^2_H values for the longitude (but not latitude) plot, was among the largest of the data sets in terms of numbers of individuals, numbers of haplotypes and numbers of localities sampled, whereas, the relatively small data set for *Agelaius felis* shows little variation in the values of σ^2_H. Some plots suggest a weak correlation between estimates of lineage age and σ^2_H for the geographically randomized data. This implies that our methods of estimation may introduce some bias, so that observed correlations should be tested against those that appear in randomized data sets.

For many of the plots, the value of σ^2_H for the oldest lineage age class appears at a point where the range of randomized values converges. This convergence occurs when the oldest age class includes only a single lineage. In this case, geographic randomization does not affect σ^2_H because the locales of all individuals are sampled without replacement. Thus, the intersection of observed σ^2_H values with randomized values at this point indicates a lack of replicates for lineages of this age, rather than geographic constraints.

Based on dispersal distances, estimated lineage ages, and the continental range of *P. maniculatus*, we did not expect mtDNA lineages in this species to have reached equilibrium distributions (see Figure 4). This expectation is consistent with the plot shown in Figure 8, in which σ^2_H increases with sequence divergence. Furthermore, estimates of σ^2_H were below the nonsignificant range for randomized data, with the anticipated exception of the value for the oldest lineage class which is unaffected by geographic randomization (see above). In contrast, latitudinal constraints in *P.*

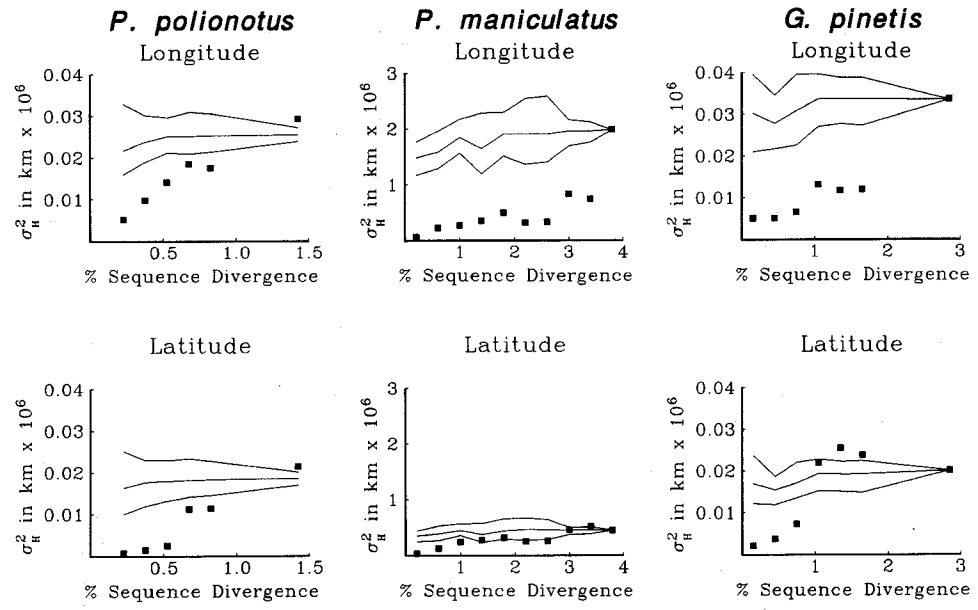

FIGURE 8.—Estimates of the longitudinal and latitudinal geographic variance, σ^2_H, of mtDNA lineages and their relative ages indicated by percent sequence divergence for three rodent species. Points indicate estimates of σ^2_H for different age classes of lineages. Lines indicate the means and upper and lower 95% limits for values of σ^2_H obtained from the data after random reshuffling of the geographic locations of individuals.

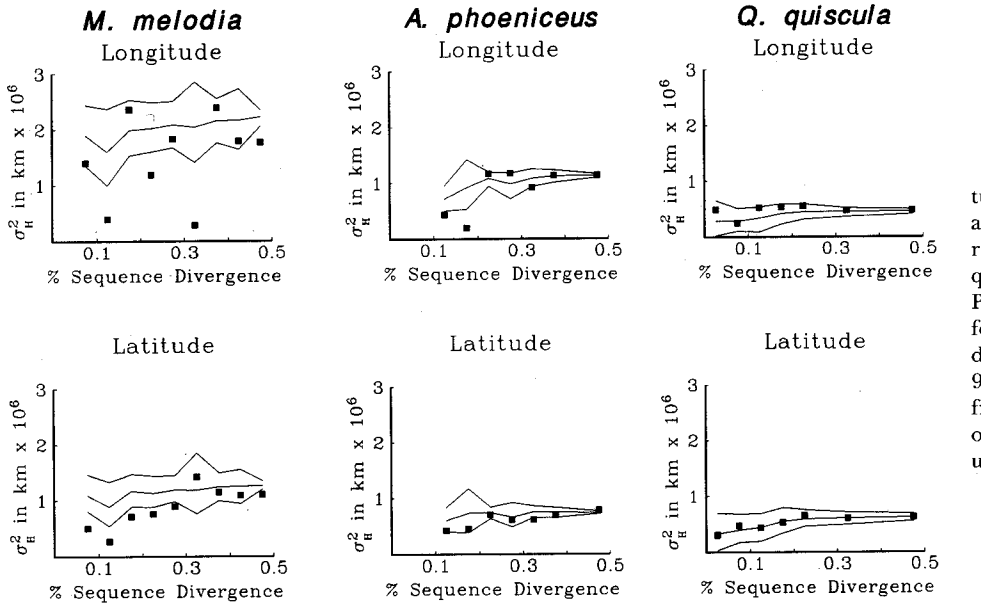

FIGURE 9.—Estimates of the longitudinal and latitudinal geographic variance, σ^2_H, of mtDNA lineages and their relative ages indicated by percent sequence divergence for three avian taxa. Points indicate estimates of σ^2_H for different age classes of lineages. Lines indicate the means and upper and lower 95% limits for values of σ^2_H obtained from the data after random reshuffling of the geographic locations of individuals.

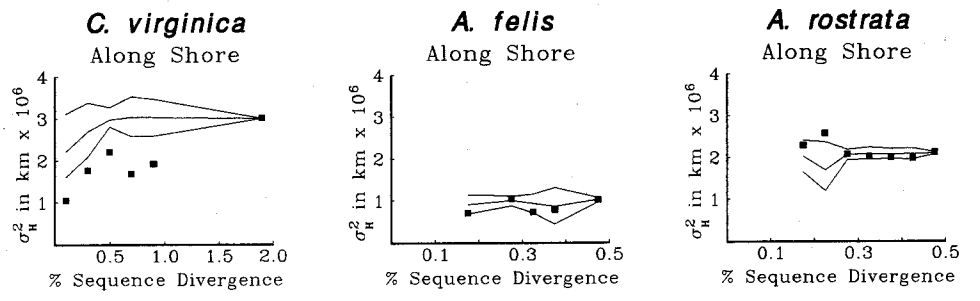

FIGURE 10.—Estimates of along shore geographic variance, σ^2_H, of mtDNA lineages and their relative ages indicated by percent sequence divergence for three marine species. Points indicate estimates of σ^2_H for different age classes of lineages. Lines indicate the means and upper and lower 95% limits for values of σ^2_H obtained from the data after random reshuffling of the geographic locations of individuals.

maniculatus appear to have affected the spread of older lineages. Values of σ^2_H on the latitudinal plot enter the nonsignificant range of randomized values for lineages with over 1% sequence divergence.

The data for *P. polionotus* represent less geographic area and a lower limit on sequence divergence than that for *P. maniculatus*. The high linear correlations between σ^2_H and lineage age for both longitude and latitude components (0.98 and 0.97; see Table 3) are in remarkably good agreement with the model's prediction for stage 1 of lineage expansion. The *P. polionotus* data also yielded the highest values for τ_K,

including a perfect rank order correlation between lineage age and latitudinal geographic variance.

Sequence divergence among lineages of *Geomys pinetis* was near that of *P. maniculatus*, whereas their overall geographic range was similar in size to that of *P. polionotus*. As with the other rodent species, the analysis indicates both a significant departure of the observed lineage distributions from randomized data and significant correlations between lineage age and geographic variance. An intriguing feature of this analysis was a sharp rise in both latitudinal and longitudinal components of geographic variance for lineages with over 1% sequence divergence.

As expected, the σ^2_H values estimated for the avian species, *A. phoeniceus*, *Q. quiscula* and *Melospiza melodia* are generally higher than those estimated for rodents, especially when comparisons are made for lineages with similar estimated ages (Figures 8 and 9). Comparisons of estimates of σ^2_H with those for randomized data suggest that avian mtDNA lineages are at or near stage 3, at which an equilibrium between gene flow and genetic drift has been achieved. However, exceptions to this generalization appear in the data for *M. melodia*. As mentioned above, there was an unusual degree of scatter among the estimated values for longitudinal σ^2_H, and several values fell well below the nonsignificant range for randomized data. Furthermore, latitudinal values of σ^2_H for *M. melodia* do not show excessive scatter and are below the nonsignificant range the five youngest classes of lineages.

For the three avian species, both the linear and rank order correlations between lineage age and σ^2_H were uniformly positive, although generally very low (Table 3). Two species, *Q. quiscula* and *M. melodia*, showed what appear to be significant rank order correlations, but, over 5% of the geographically randomized data sets produced τ_K values that were higher than those observed, and so they were considered nonsignificant.

Analysis of data for the marine fishes, *Anguilla rostrata* and *Arius felis*, produced plots of σ^2_H *vs.* lineage age (Figure 10) that look much like those obtained for the birds *A. phoeniceus* and *Q. quiscula* (Figure 9). Values are close to the means observed for randomized data, and there are no apparent trends. Both linear and rank order correlations for *A. rostrata* were negative, although not at a significant level. However, analysis of the data for the oyster, *Crassostrea virginica*, suggests a nonequilibrium distribution. Estimated values of σ^2_H were well below the nonsignificant range for randomized values, and there was a moderate linear correlation between estimates of σ^2_H and lineage age. Although the value of τ_K for this correlation appears to be significant, over 5% of the geographically randomized data sets produced τ_K values that were higher. However, it should be pointed out that there were only six nonempty age classes for this species, which reduced the statistical power for the rank order comparison.

DISCUSSION

Not all theoretical models of the relationship between gene flow and population structure include the effects of mutation. Differences in the assumptions of gene flow models have been considered by FELSENSTEIN (1976) and SLATKIN (1985). The use of WRIGHT's (1951) F_{ST} and related quantities have provided measures of population structure that are relatively insensitive to mutation rates and details of the model, provided that mutation rates are much lower than migration rates (SLATKIN 1985). Another class of models, based on genetic distance measures such as NEI's (1972), provide a more direct link between mutation and the divergence of populations. However, this approach has not been developed as a tool for the estimation of migration parameters to the same extent as methods based on F_{ST}.

Applications of gene flow theory to allozyme data appear to have relied mostly on models in which the effects of mutation are ignored. This may be a reasonable assumption for those allozyme loci that exhibit only a few common alleles and have low mutation rates (*i.e.*, on the order of 10^{-6} per generation). With some additional assumptions, these models predict a relationship between Nm (the product of the effective population size and migration rate), and the correlation of alleles drawn from the same population relative to alleles drawn from different populations. This has led to the use of Wright's F_{ST} and related statistics to derive estimates of Nm (reviewed in SLATKIN 1987).

Statistics that are analogous to F_{ST} have also be used in the analysis of animal mtDNA data. These include G_{ST} (NEI 1973; TAKAHATA and PALUMBI 1985) and N_{ST} (LYNCH and CREASE 1990). However, mtDNA variation as detected by either restriction site analysis or by direct sequencing differs from allozyme variation in several important respects: (1) Animal mtDNA is typically transmitted maternally as a haploid genome. Thus, models based on biparental Mendelian inheritance may be inappropriate. (2) The mitochondrial genome appears to be transmitted as a single linkage unit. It is therefore not possible to sample multiple, independently segregating loci to obtain more accurate estimates of parameters such as F_{ST}. (3) The mutation rate for the entire mtDNA molecule appears to be relatively high, on the order of 10^{-4} mutations per generation. The assumption that mutation rate is much lower than migration rate is more likely to be violated. And (4), the total number of mtDNA variants observed generally depends on the amount of mtDNA sequence surveyed. Thus, the rate of detectable mutations may vary considerably among studies.These properties suggest that it may inappro-

priate to estimate Nm by the same approach used for allozyme variation.

TAKAHATA and PALUMBI (1985) modified WRIGHT's (1943) finite island model for mtDNA by including maternal transmission and sex-dependent gene flow. Their results are essentially equivalent to those obtained by CROW and AOKI (1984) for nuclear loci, if the effective migration rate for mitochondrial genomes is divided by two. However, because the finite island model assumes that migration rates between demes are independent of geographic distance, their model may be inappropriate for the large scale, hierarchical patterns of geographic differentiation (particularly in rodents) that we have considered here. Furthermore, although their model includes mutation, they treat different sites within the mitochondrial genome as independent loci and assume that mutation can be ignored in the estimation of Nm from G_{ST}.

More recently, SLATKIN (1991) has used the relationship between the probability of identity by descent and coalescence time (time to a common ancestral mtDNA) to find the equilibrium distributions of coalescence times for Wright's finite island model as well as the geographically structured one- and two-dimensional stepping stone models. These models are well suited for considering the effects of mutation because the probability of mutation should be correlated with coalescence time. Extension of this approach to Wright's isolation by distance model represents an alternative, but complementary, approach to our own. Which of these approaches would be more appropriate could depend on the historical biogeography of a given species, as well as the specific objectives of the analysis. Like coalescence models, our model for nonequilibrium distributions of mtDNA variants would also apply to unique and nonrecombining nuclear variants that arose at a sufficient rate. Such rapidly evolving components of the nuclear genome may be restricted to synonymous nucleotide substitutions, or substitutions that occur in noncoding sequences.

The results of our simulation studies suggest that the relationship between the age of a lineage and its geographic variance is fairly robust over a range of population sizes and with respect to several models of population regulation. The case of local density regulation is of particular importance, because it has been difficult to develop analytical models of continuously distributed populations (FELSENSTEIN 1975). However, it has not been feasible to conduct these simulations on large enough scales to be directly comparable with the species we have considered. Thus, it would still be useful to develop analytical treatments of these models that could be extended to larger populations and longer time scales.

Disruptions of gene flow that geographically partition a species in the course of its history could also generate phylogeographic patterns that meet our criteria for nonequilibrium situations. Under vicariance scenarios, mutations that postdate the formation of a geographic barrier to gene flow would be restricted to subregions within a species' range. Indeed, we have argued previously that such vicariant events have played a major role in shaping geographic variation in animal mtDNA (AVISE *et al.* 1987). However, for continuously distributed species it may not always be possible to distinguish the effects of vicariance *vs.* isolation by distance through simple examination of the geographic distributions of mtDNA lineages alone. Even such a distinct feature as a sharp geographic boundary for a lineage may represent nothing more than the "snapshot" of a steadily expanding front of dispersal. This line of reasoning implies that specific explanations for every feature of an empirical mtDNA distribution may be unwarranted. Vicariant explanations would be strongly favored only for phylogeographic patterns that clearly were: (a) concordant with longstanding geographic barriers; (b) shared by several taxa (AVISE 1992); or (c) displayed concordantly by independently segregating genetic markers within a single taxon (AVISE and BALL 1990; KARL and AVISE 1992).

Perhaps the most surprising result of our mtDNA analyses is the near linearity of the age-variance relationship for rodent mtDNA lineages that presumably have experienced several Pleistocene glaciations (GRAHAM 1986). It might be expected that range contractions and expansions, which are believed to have accompanied glacial cycles, would obliterate this feature of lineage distributions. However, we have found that the age-variance relationship can be maintained in simulations of cyclical range contractions. From visual inspection of the distributions of the lineages during the simulations, it appears that while some lineages became extinct during range contractions, others were compressed and reexpanded in the same relative positions. Thus, it appears that nested phylogeographic distributions can arise in the absence of barriers to gene flow, and that these distributions can be remarkably resilient. Furthermore, if vicariant events have contributed to the geographic structure of mtDNA or other genetic lineages within a species, the imprints of these events may no longer be superimposed exactly over the geographic features that initially forged them.

Research support was provided under The Louisiana Education Quality Support Fund grant LEQSF(1990–92)-RD-A-30 (J. NEIGEL), EPSCoR NSF/LEQSF (1992–1996)-ADP-02 (J. NEIGEL) and by grants from the National Science Foundation to J. AVISE.

LITERATURE CITED

AVISE, J. C., 1992 Molecular population structure and the biogeographic history of a regional fauna: a case history with lessons for conservation biology. Oikos **63:** 62–76.

AVISE, J. C., and R. M. BALL, JR., 1990 Principles of genealogical concordance in species concepts and biological taxonomy. Oxf. Surv. Evol. Biol. **7:** 45–67.

AVISE, J. C., J. E. NEIGEL and J. ARNOLD. 1984, Demographic influences on mitochondrial DNA lineage survivorship in animal populations. J. Mol. Evol. **20:** 99–105.

AVISE, J. C., C. A. REEB and N. C. SAUNDERS, 1987 Geographic population structure and species differences in mitochondrial DNA of mouthbrooding marine catfishes (Arridae) and demersal spawning toadfishes (Batrachoididae). Evolution **41:** 991–1002.

AVISE, J. C., J. ARNOLD, R. M. BALL, E. BERMINGHAM, T. LAMB, *et al.*, 1987 Intraspecific phylogeography: the mitochondrial DNA bridge Between population genetics and systematics. Annu. Rev. Ecol. Syst. **18:** 489–522.

AVISE, J. C., C. GIBLIN-DAVIDSON, J. LAERM, J. C. PATTON and R. A. LANSMAN, 1979 Mitochondrial DNA clones and matriarchal phylogeny within and among geographic populations of the pocket gopher, *Geomys pinetis*. Proc. Natl. Acad. Sci. USA **76:** 6694–6698.

AVISE, J. C., G. S. HELFMAN, N. C. SAUNDERS and L. S. HALES, 1986 Mitochondrial DNA differentiation in North Atlantic eels: population genetic consequences of an unusual life history pattern. Proc. Natl. Acad. Sci. USA **83:** 4350–4354.

AVISE, J. C., J. F. SHAPIRA, S. W. DANIEL, C. F. AQUADRO and R. A. LANSMAN, 1983 Mitochondrial DNA differentiation during the speciation process in *Peromyscus*. Mol. Biol. Evol. **1:** 38–56.

BALL, R. M. JR., S. FREEMAN, F. C. JAMES, E. BERMINGHAM and J. C. AVISE, 1988 Phylogeographic population structure of Red-winged Blackbirds assessed by mitochondrial DNA. Proc. Natl. Acad. Sci. USA. **85:** 1558–1562.

BARROWCLOUGH, G. F., 1980 Gene flow, effective population sizes, and genetic variance components in birds. Evolution **34:** 789–798.

BHATTACHYARYYA, G. K., and R. A. JOHNSON, 1977 *Statistical Concepts and Methods*. John Wiley and Sons, New York.

BLAIR, W. F., 1940 A study of prairie deer-mouse populations in southern Michigan. Am. Midland Nat. **24:** 273–305.

CROW, J. F., and K. AOKI, 1984 Group selection for a polygenic behavioral trait: Estimating the degree of population subdivision. Proc. Natl. Acad. Sci. USA **81:** 6073–6077.

DICE, L. R., and W. E. HOWARD, 1951 Distance of dispersal by prairie deer mice from birthplaces to breeding sites. Cont. Lab. Vert. Biol. Univ. Mich. **50:** 1–15.

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

FELSENSTEIN, J., 1976 The theoretical population genetics of variable selection and migration. Annu. Rev. Genet. **10:** 253–280.

GRAHAM, R. W. 1986 Response of mammalian communities to environmental changes during the Late Quaternary, pp. 300–313 in *Community Ecology*, edited by J. DIAMOND and T. J. CASE. Harper and Row, New York.

KARL, S. A., and J. C. AVISE, 1992 Balancing selection at allozyme loci in oysters: implications from nuclear RFLPs. Science **256:** 100–102.

LANSMAN, R. A., J. C. AVISE, C. F. AQUADRO, J. F. SHAPIRA and S. W. DANIEL, 1983 Extensive genetic variation in mitochondrial DNAs among geographic populations of the deer mouse, *Peromyscus maniculatus*. Evolution **37:** 1–16.

LAYNE, J. N., 1968 Ontogeny, pp. 148–253 in *Biology of Peromyscus (Rodentia)*, edited by J. A. King. Am. Soc. Mammal. Spec. Publ. No. 2.

LYNCH, M., and T. J. CREASE, 1990 The analysis of population survey data on DNA sequence variation. Mol. Biol. Evol. **7:** 377–394.

MOORE, W. S., and R. A. DOLBEER, 1989 The use of banding recovery data to estimate dispersal rates and gene flow in avian species: case studies in the red-winged blackbird and common grackle. Condor **91:** 242–253.

MORITZ, C., T. E. DOWLING and W. M. BROWN, 1987 Evolution of animal mitochondrial DNA: relevance for population biology and systematics. Annu. Rev. Ecol. Syst. **18:** 269–292.

NEI, M, 1972 Genetic distance between populations. Am. Nat. **106:** 283–292.

NEI, M., 1973 Analysis of gene diversity in subdivided populations. Proc. Natl. Acad. Sci. USA **70:** 3321–3323.

NEI, M., and W. H. LI, 1979 Mathematical model for studying genetic variation in terms of restriction endonuclease. Proc. Natl. Acad. Sci. **76:** 5269–5273.

NEIGEL, J. E., R. M. BALL and J. C. AVISE, 1991 Estimation of single generation migration distances from geographic variation in animal mitochondrial DNA. Evolution **45:** 423–432.

REEB, C. A., and J. C. AVISE, 1990 A genetic discontinuity in a continuously distributed species: mitochondrial DNA in the American oyster, *Crassostrea virginica*. Genetics **124:** 397.

ROHLF, F. J., 1988 NTSYS-pc Numerical Taxonomy and Multivariate Analysis System, Version 1.50. Exeter Publishing, Ltd., New York.

SLATKIN, M., 1987 Gene flow and the geographic structure of natural populations. Science **236:** 787–792.

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

SLATKIN, W. M., 1985 Gene flow in natural populations. Annu. Rev. Ecol. Syst. **16:** 393–430.

STICKEL, L. F., 1968 Home range and travels, pp 373–411 in *Biology of Peromyscus (Rodentia)*, edited by J. A. KING. Am. Soc. Mammal. Spec. Publ. No. 2.

TAKAHATA, N., and S. R. PALUMBI, 1985 Extranuclear differentiation and gene flow in the finite island model. Genetics **109:** 441–457.

WRIGHT, S., 1943 Isolation by distance. Genetics **28:** 114–138.

WRIGHT, S., 1951 The genetical structure of populations. Ann. Eugen. **15:** 323–354.

ZINK, R. M., 1991 The geography of mitochondrial DNA variation in two sympatric sparrows. Evolution **45:** 329–339.

ZINK, R. M., D. L. DITTMANN and W. L. ROOTES, 1991 Mitochondrial DNA variation and the phylogeny of Zonotrichia. Auk **108:** 578–584.

ZINK, R. M., W. L. ROOTES and D. L. DITTMANN, 1991 Mitochondrial DNA variation, population structure, and evolution of the common grackle (*Quiscalus quiscula*). Condor **93:** 318–329.

Communicating editor: A. G. CLARK