Title
Application of a random walk model to geographic distributions of animal mitochondrial DNA variation.

Permalink
https://escholarship.org/uc/item/9sp6k341

Journal
Genetics, 135(4)

ISSN
0016-6731

Authors
Neigel, JE
Avise, JC

Publication Date
1993-12-01

License
https://creativecommons.org/licenses/by/4.0/ 4.0

Peer reviewed
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 to 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.

G ENETIC 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 $\sigma_H^2$. 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 $\sigma_H^2$, whereas a numerically large lineage could be restricted to a single locale and display a low $\sigma_H^2$). Furthermore, the actual distribution of single
Mutations Exceeding Dispersal

- Mutation
- Origin

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

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 \( \sigma_H^2 \) 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 geographic variance and a lineage's 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 "non-nested." 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 \( \left( \sigma_H^2 \right) \) of each lineage's distribution along the \( x \)-coordinate was calculated, along with the
There is a positive, but nonlinear correlation between dispersal distance and the carrying capacity of 100 females, a standard dispersal distance of about 200 m for 
P. maniculatus and the red-winged blackbird. The longitudinal width of the North American continent is shown at 4 x 10^2 km.

There is no correlation between \( \sigma^2_H \) and lineage age, so that \( \sigma_X \) 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. 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 x 10^2 km.

There is no correlation between \( \sigma^2_H \) and lineage age, so that \( \sigma_X \) 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. 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 x 10^2 km.
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 $\sigma_x$ 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, $\sigma_x$ 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 $\sigma^2_x$, the variance in position of the members of a lineage, vs. lineage age for nested 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 $\sigma^2_x$ data were also converted to $\sigma_x$ 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 $\sigma_x$ 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 $\sigma_x$ 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 $\sigma_x$ 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 $\sigma_x$ underestimated the

**TABLE 1**

<table>
<thead>
<tr>
<th></th>
<th>Young</th>
<th>Old</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum age</td>
<td>1</td>
<td>115</td>
</tr>
<tr>
<td>Maximum age</td>
<td>99</td>
<td>949</td>
</tr>
<tr>
<td>Mean age</td>
<td>47</td>
<td>574</td>
</tr>
<tr>
<td>Number of lineages</td>
<td>202</td>
<td>15</td>
</tr>
<tr>
<td>Mean $\sigma_x$</td>
<td>7.07</td>
<td>7.23</td>
</tr>
<tr>
<td>Standard deviation $\sigma_x$</td>
<td>2.95</td>
<td>2.77</td>
</tr>
<tr>
<td>Standard error $\sigma_x$</td>
<td>0.21</td>
<td>0.71</td>
</tr>
</tbody>
</table>

Wilcoxon rank-sum test $P = 0.46$. 

![Figure 5](image.png)
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 $\sigma_N$ was closer to the actual standard dispersal distance in simulations of larger populations. However, differences in $\sigma_N$ 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 $\alpha$ 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 $K_L = (K \times A_L)/A_T$, where $K$ is the entire population's carrying capacity, $A_L$ 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 $K_L/N_L$, where $N_L$ 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 2

<table>
<thead>
<tr>
<th>$K_1$</th>
<th>$K_2$</th>
<th>np</th>
<th>$\sigma_{x1}$</th>
<th>$\sigma_{x2}$</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>1,000</td>
<td>387</td>
<td>7.3</td>
<td>7.5</td>
<td>0.271</td>
</tr>
<tr>
<td>100</td>
<td>5,000</td>
<td>326</td>
<td>7.2</td>
<td>7.6</td>
<td>0.230</td>
</tr>
<tr>
<td>100</td>
<td>10,000</td>
<td>310</td>
<td>7.2</td>
<td>7.9</td>
<td>0.007</td>
</tr>
<tr>
<td>1,000</td>
<td>5,000</td>
<td>319</td>
<td>7.4</td>
<td>7.6</td>
<td>0.920</td>
</tr>
<tr>
<td>1,000</td>
<td>10,000</td>
<td>303</td>
<td>7.4</td>
<td>7.9</td>
<td>0.089</td>
</tr>
<tr>
<td>5,000</td>
<td>10,000</td>
<td>325</td>
<td>7.6</td>
<td>8.0</td>
<td>0.057</td>
</tr>
</tbody>
</table>

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, $\sigma_N$ is the standard deviation of dispersal distances. The actual standard deviation of dispersal distances, $\sigma_n$, was 9 in every case. $P$ is the probability for a Wilcoxon signed-rank test for lineages paired by age.

### Table 3

<table>
<thead>
<tr>
<th>$K_L$</th>
<th>$K_C$</th>
<th>$N_L$</th>
<th>$N_C$</th>
<th>$\sigma$</th>
<th>$w$</th>
<th>np</th>
<th>$\sigma_{xL}$</th>
<th>$\sigma_{xC}$</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>1000</td>
<td>997</td>
<td>1164</td>
<td>9</td>
<td>500</td>
<td>92</td>
<td>7.3</td>
<td>7.0</td>
<td>0.057</td>
</tr>
<tr>
<td>5000</td>
<td>5000</td>
<td>4769</td>
<td>4984</td>
<td>93</td>
<td>100</td>
<td>147</td>
<td>77.1</td>
<td>76.1</td>
<td>0.549</td>
</tr>
<tr>
<td>5000</td>
<td>2767</td>
<td>2767</td>
<td>2758</td>
<td>93</td>
<td>100</td>
<td>79</td>
<td>75.3</td>
<td>77.4</td>
<td>0.764</td>
</tr>
</tbody>
</table>

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, $\sigma$ 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, $\sigma_{xL}$ 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 $\sigma_N$ 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 $\sigma_N$ within each pair appear to be similar, and none were significantly different by Wilcoxon signed-rank tests with $\alpha$ 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 refuge, 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 $\sigma_N^2$ 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 $\sigma_N$ 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-
graphic clustering of mtDNA variants and a positive correlation between estimates of $\sigma^2_H$ 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 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 $\sigma^2_H$ 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 $\sigma^2_H$ 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,

**TABLE 4**

Data sets analyzed for this study

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

**TABLE 5**

PHYFORM estimates of dispersal distances

<table>
<thead>
<tr>
<th>Species</th>
<th>Generation time (yr)</th>
<th>Median lineage age (yr × 10^3)</th>
<th>Longitude (axis)</th>
<th>Latitude (axis)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Peromyscus maniculatus</em></td>
<td>0.2</td>
<td>200</td>
<td>0.14</td>
<td>0.10</td>
</tr>
<tr>
<td><em>Peromyscus polionotus</em></td>
<td>0.2</td>
<td>225</td>
<td>0.06</td>
<td>0.02</td>
</tr>
<tr>
<td><em>Geomys pinetis</em></td>
<td>1</td>
<td>150</td>
<td>0.18</td>
<td>0.12</td>
</tr>
<tr>
<td><em>Agelaius phoeniceus</em></td>
<td>3</td>
<td>125</td>
<td>3.2</td>
<td>3.2</td>
</tr>
<tr>
<td><em>Quiscalus quiscula</em></td>
<td>3</td>
<td>25</td>
<td>7.6</td>
<td>6.0</td>
</tr>
<tr>
<td><em>Melospiza melodia</em></td>
<td>2</td>
<td>75</td>
<td>6.1</td>
<td>3.6</td>
</tr>
</tbody>
</table>

**TABLE 6**

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

<table>
<thead>
<tr>
<th>Species</th>
<th>Linear correlation</th>
<th>Kendall's tau</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Peromyscus maniculatus</em></td>
<td>0.78</td>
<td>0.78*</td>
</tr>
<tr>
<td><em>Peromyscus polionotus</em></td>
<td>0.98</td>
<td>0.97*</td>
</tr>
<tr>
<td><em>Geomys pinetis</em></td>
<td>0.94</td>
<td>0.76*</td>
</tr>
<tr>
<td><em>Agelaius phoeniceus</em></td>
<td>0.23</td>
<td>0.53*</td>
</tr>
<tr>
<td><em>Quiscalus quiscula</em></td>
<td>0.23</td>
<td>0.81*</td>
</tr>
<tr>
<td><em>Melospiza melodia</em></td>
<td>0.08</td>
<td>0.67*</td>
</tr>
</tbody>
</table>

* Indicates values of $r_K$ that exceeded 95% of the distribution of $r_K$ for 100 replicates in which the data were geographically randomized data over the sample locales.
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 \( \sigma^2_H \) 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 \( \sigma^2_H \), with subsequent averaging of these \( \sigma^2_H \) 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 \( \sigma^2_H \).

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 percent 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 \( \sigma^2_H \) for each lineage age class in the original data set was then compared with the distribution of \( \sigma^2_H \) for 100 replicate randomized data sets. We refer to the range from the fifth lowest to the fifth highest of these \( \sigma^2_H \) values as the nonsignificant range. Only \( \sigma^2_H \) values below the nonsignificant range were considered to represent significant geographic clustering. To test for positive correlations between lineage age and \( \sigma^2_H \), values of the linear correlation coefficient (\( r^2 \)) and the nonparametric rank order correlation statistic, Kendall's tau (\( \tau_K \)) (BHATTACHARYYA and JOHNSON 1977), were calculated for each data set. Because estimates of \( \sigma^2_H \) for different age classes are not independent, the significance of an apparently high value of \( \tau_K \) was further evaluated by comparison with the distribution of \( \tau_K \) 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 $\sigma^2_H$ 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 $r_K$ was used to examine the significance of nonlinear rank order correlations. For those values of $r_K$ indicated as significant, less than 5% of the values calculated from geographically randomized data equaled or exceeded these observed values. These correlations are thus 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 $\sigma^2_H$ and mtDNA lineage age (shown as percent sequence divergence) for all nine taxa, along with the means and limits for the nonsignificant range of $\sigma^2_H$ 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 $\sigma^2_H$ 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 $\sigma^2_H$ 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 $\sigma^2_H$ 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 $\sigma^2_H$ 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 $\sigma^2_H$. Some plots suggest a weak correlation between estimates of lineage age and $\sigma^2_H$ 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 $\sigma^2_H$ 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 $\sigma^2_H$ because the locales of all individuals are sampled without replacement. Thus, the intersection of observed $\sigma^2_H$ 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 $\sigma^2_H$ increases with sequence divergence. Furthermore, estimates of $\sigma^2_H$ 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.
Random Walk mtDNA Distributions

**P. maniculatus** appear to have affected the spread of older lineages. Values of \( \sigma_n^2 \) 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 \( \sigma_n^2 \) 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 \( \tau_K \).
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 $\sigma_H^2$ 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 $\sigma_H^2$ 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 $\sigma_H^2$, and several values fell well below the nonsignificant range for randomized data. Furthermore, latitudinal values of $\sigma_H^2$ for M. melodia do not show excessive scatter and are below the nonsignificant range for the five youngest classes of lineages.

For the three avian species, both the linear and rank order correlations between lineage age and $\sigma_H^2$ 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 data sets produced $\tau_K$ 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 $\sigma_H^2$ 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 $\sigma_H^2$ were well below the nonsignificant range for randomized values, and there was a moderate linear correlation between estimates of $\sigma_H^2$ and lineage age. Although the value of $\tau_K$ for this correlation appears to be significant, over 5% of the geographically randomized data sets produced $\tau_K$ 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 been 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 a number of factors, 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 versus 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., C. A. REEB and N. C. SAUNDERS, 1987 Geographic population structure and species differences in mitochondrial DNA of mouthbrooding marine catfishes (Arriidae) and demersal spawning toadfishes (Batrachoididae). Evolution 41: 991–1002.


