THEORETICAL POPULATION BIOLOGY 44, 110–127 (1993)

High Mutation Rate Loci in a Subdivided Population

NEIL O'CONNELL

Department of Statistics, University of California, Berkeley, California 94720

AND

MONTGOMERY SLATKIN*

Department of Integrative Biology, University of California, Berkeley, California 94720

Received June 16, 1992

Analytic and simulation studies were carried out in order to predict the average geographic area occupied by alleles in a continuously distributed population. The properties of three statistics were investigated: the sum of the squared distances between members of allelic classes, the sum of the root mean squared distances, and the sum of the squares of the numbers of alleles. The expectations of these quantities can be obtained analytically from both stepping-stone and branching diffusion models. The predictions of these two models are similar for wide ranges of parameter values and are consistent with the simulation results from a stepping-stone model. These results suggest that measures of the geographic distribution of alleles can be useful for estimating average dispersal distances at loci, such as minisatellite and microsatellite loci, at which mutation rates are high enough that they can be estimated with confidence.

1. INTRODUCTION

The use of new biochemical methods has led to the discovery of previously unsuspected levels of variation at individual genetic loci. Minisattelite (or VNTR) loci (Nakamura *et al.*, 1987) have alleles that are distinguished by the numbers of repeats of nearly conserved nucleotide sequencies of 30 or so nucleotides. Recently, Jeffreys, Neuman, and Wilson (1990) have further resolved alleles at VNTR loci by detecting the presence or absence of restriction sites within individual repeats. Microsattelite loci are similar but the repeats are of only 2, 3, or 4 base pairs (Boerwinkle

* To whom correspondence should be addressed.

0040-5809/93 $5.00

et al., 1989). Both classes of loci differ from other loci in being highly polymorphic, with heterozygosities approaching 95%, and in being highly mutable, with mutation rates on the order of 10^{-2} to 10^{-4} (Jeffreys *et al.*, 1991). As a consequence, they are of considerable practical importance in forensic applications and as markers in genetic mapping.

There have been extensive surveys of human populations for allele frequencies at both kinds of loci and VNTR loci are being used in surverys of animal populations as well (Dallas, in prep.). Nichols and Balding (1991) consider the effect of population subdivision on the allele frequency distributions at VNTR loci, but they concentrated on forensic applications. Highly mutable loci are also of interest from an evolutionary perspective. The high mutation rate suggests that the individual lifetime of an allele is very short. Therefore, the geographic area occupied by an allele is likely to indicate the extent of recent dispersal in a species. In this paper we investigate that possibility and show that certain measures of the geographic distribution of alleles generally depend in a relatively simple way on mutation rate and average dispersal distance. Because the mutation rates at highly mutable loci can be estimated with some confidence, our results suggest that geographic surveys of allele frequencies at such loci can be used to obtain rough estimates of average dispersal distances in relatively recent dispersal events.

We are concerned with a species that is continuously distributed in space. Our results are based on both analytic theory and computer simulations. The analytic theory is based on two kinds of models, stepping-stone and branching diffusion models. In the stepping-stone model, there is density regulation in each population so the population density remains uniform. The simulation model is also a stepping-stone model. The branching diffusion model has no local density regulation. Felsenstein (1975) has discussed the problems with using such a model in a population genetic context and Sawyer (1976a) showed that in general in a two-dimensional habitat, infinitely large clumps tend to form even if individuals are initially uniformly distributed on a plane. We show that despite this apparently fatal problem with the model, we can obtain sensible results from it. In fact, the results we obtain are nearly the same as those from the stepping-stone model, analyzed by Malécot (1968), Weiss and Kimura (1965), Sawyer (1977, 1979), and others. The simulation results are in agreement with the analytic theory and support the use of the analytic theory for parameter values that cannot be simulated.

2. Analytic Theory

Suppose we sample individuals from a geographically structured population, over a finite area A, and record the location and allelic type of each

individual sampled. If there are M distinct types present, denote by $k_1, ..., k_M$ the respective numbers of each type and for each j, by $x_1^j, ..., x_{k_j}^j$ the locations of the k_j type-j individuals. Denote by $|x-y|$ the distance between locations x and y. For our purposes A will be a bounded subset of $\mathbb{Z}^d$ or $\mathbb{R}^d$ and $|\cdot|$ the standard Euclidean distance. We are interested in the statistics

$$S_{A,\delta} = 2 \sum_{j=1}^{M} \sum_{l=1}^{k_j} \sum_{m=l+1}^{k_j} |x_l^j - x_m^j|^\delta, \tag{1}$$

for $\delta = 0, 1, 2$. For $\delta > 0$, $S_{A,\delta}$ can be thought of as the total variational spread of individual allelic types. Note that

$$S_{A,0} = \sum_{j=1}^{M} k_j(k_j - 1)$$

counts the number of terms in the summation. The ratio $S_{A,\delta}/S_{A,0}$ is therefore the average "area" occupied by an individual allelic type, or more precisely, the average squared/absolute distance between individuals of the same type. How do the properties of these statistics depend on (i) the model assumed, and (ii) the parameters of the model?

We study the behaviour of these statistics under the two seemingly different models for the underlying population dynamics. The first is the general stepping-stone model, where local density is assumed constant. The second is a Markov branching diffusion model. In both cases we assume the occurrence of selectively neutral mutations at a constant rate, where each mutation produces an entirely new allelic type. The key parameters in both models are the mutation rate and the variance of dispersal distances (the "migration variance"), although we see how the "shape" of the distribution of dispersal distance can also play a role. We observe that if the mutation rate is small and the migration variance sufficientlfy high, both models yield essentially the same results. We formulate both models in general terms and then specialize to the case in which the population density is unity and the distribution of dispersal distance is Gaussian. Results for other population densities are given for the stepping-stone model and can be obtained for the brancing difusion model by rescaling units of length.

2.1. *Stepping-Stone Model*

The model we assume here is described in detail by Sawyer (1976b, 1977), and is essentially the stepping-stone model of Malécot (1968) and Weiss and Kimura (1965), with general migrations. Suppose we have a population of individuals which is held at constant local density of N per

site in the infinite d-dimensional lattice. These sites are called "colonies." The population undergoes random mating within each colony, independent migration between colonies according to a migration density $g(x)$, and is subject to selectively-neutral mutations occuring at rate u per individual per generation, where each mutation produces an entirely new allelic type.

In what follows we assume that the creatures are diploid, although we approximate this by assuming that there are $2N$ independent haploid individuals at each site. Note therefore that results for haploid species can be deduced by replacing $2N$ by N. We say two individuals are identical by descent (i.b.d.) if and only if they are of the same allelic type.

Now suppose we sample one individual from each colony in a bounded region A. Set

$$S_{A,\delta} = \sum_{x,y \in A, x \neq y} |x-y|^{\delta} 1_B(x, y), \tag{2}$$

where 1_B is the indicator function of the set

$$B = \{(x, y)\text{: Individuals sampled at } x \text{ and } y \text{ are i.b.d.}\}.$$

The expectation of this quantity, at equilibrium, is given by

$$ES_{A,\delta} = \sum_{x,y \in A, x \neq y} |x-y|^{\delta} I(x-y, u), \tag{3}$$

where $I(x-y, u)$ is the probability that, at equilibrium, two randomly chosen individuals at locations x and y are i.b.d. The quantity $I(x-y, u)$ has been studied extensively in the literature. Exact formulas are known (Malécot, 1968; Sawyer, 1976b) and can be computed, but for our purposes it is convenient to appeal to the approximation results of Sawyer (1977) for $I(x, u)$ when $|x| > 0$ and u is small (or more precisely, when $|x|/\sigma = O(u^{-1/2})$). We first consider the two-dimensional case. Analogous results in one and three dimensions are given in the appendix.

For notational convenience assume that $g(x)$ is spherically symmetric with variance σ^2 (in each direction) and $\sum |x|^{2+a} g(x) < \infty$ for some $0 < a < 2$. Then (Sawyer, 1977, Th. 2),

$$I(x, u) = \frac{2K_0(|x|\sqrt{2u}/\sigma)}{4\pi(2N\sigma^2 + C_0) - \log(2u)}(1 + O(u^{a/4})) \tag{4}$$

uniformly for $0 < \varepsilon \leqslant |x|\sqrt{2u} \leqslant 1/\varepsilon$, for any $\varepsilon > 0$, where K_0 is the modified Bessel function of the second kind and the constant C_0 depends on the migration density. For Gaussian migrations $C_0 = 0$ is independent of σ. For Laplace (double-exponential) migrations with $\sigma \geqslant \sqrt{2}$ say, $C_0 \simeq -0.13$ does not contribute very much to $I(x, u)$. However, for any value of σ^2

there are migration densities for which C_0 can be arbitrarily large. For example, Malécot's "K-distribution" has this property (cf. Sawyer, 1977).

Let $A_q = [0, q]^2 \cap \mathbb{Z}^2$, and write $S_{q,\delta}$ for $S_{A_q,\delta}$. Using the above approximation we have for small u,

$$ES_{q,\delta} \simeq \sum_{x, y \in A_q,\, x \neq y} |x-y|^{\delta} \frac{2K_0(|x-y|\sqrt{2u}/\sigma)}{4\pi(2N\sigma^2 + C_0) - \log(2u)}. \tag{5}$$

We can now approximate this sum by an integral to get

$$ES_{q,\delta} \simeq \frac{2}{4\pi(2N\sigma^2 + C_0) - \log(2u)} \times \iint_{[0,q]^2 \times [0,q]^2} 1_{(|x-y| > \rho)} |x-y|^{\delta} K_0\left(|x-y|\frac{\sqrt{2u}}{\sigma}\right) dx\, dy, \tag{6}$$

for some $0 < \rho < 1$. The choice of ρ is somewhat arbitrary in this kind of approximation, but only really matters when q/u is small, which is not in the range of interest. We can simplify the above four-dimensional integral to get the two-dimensional one,

$$\begin{aligned} ES_{q,\delta} \simeq \frac{8}{4\pi(2N\sigma^2 + C_0) - \log(2u)} \Bigg(& \int_{\rho}^{q} r^{\delta+1} K_0\left(\frac{\sqrt{2u}}{\sigma} r\right) \\ & \times \left[\frac{\pi}{2} q^2 - 2qr + \frac{r^2}{2}\right] dr + \int_{q}^{\sqrt{2}q} r^{\delta+1} K_0\left(\frac{\sqrt{2u}}{\sigma} r\right) \\ & \times \left[\left(\frac{\pi}{2} - 2\cos^{-1}\frac{q}{r}\right) q^2 + 2q\sqrt{r^2 - q^2} - q^2 - \frac{r^2}{2}\right] dr \Bigg), \end{aligned}$$

from which we obtain the limiting result

$$\lim_{q \to \infty} q^{-2} ES_{q,\delta} \simeq \frac{4\pi}{4\pi(2N\sigma^2 + C_0) - \log(2u)} \int_{\rho}^{\infty} r^{\delta+1} K_0\left(\frac{\sqrt{2u}}{\sigma} r\right) dr.$$

But for small u,

$$\begin{aligned} \int_{\rho}^{\infty} r^{\delta+1} K_0\left(\frac{\sqrt{2u}}{\sigma} r\right) dr &\simeq \int_{0}^{\infty} r^{\delta+1} K_0\left(\frac{\sqrt{2u}}{\sigma} r\right) dr \\ &= 2^{\delta}\left(\frac{\sqrt{2u}}{\sigma}\right)^{-\delta-2} \Gamma\left(\frac{\delta+2}{2}\right)^2, \end{aligned}$$

(see, for example, Gradshteyn and Ryzhik (1964, p. 684)) so we get

$$\lim_{q \to \infty} q^{-2} ES_{q,\delta} \simeq \frac{2^{\delta+2}\pi\left(\frac{\sqrt{2u}}{\sigma}\right)^{-\delta-2} \Gamma\left(\frac{\delta+2}{2}\right)^2}{4\pi(2N\sigma^2 + C_0) - \log(2u)}. \tag{7}$$

In particular, when $N=1$ and the migration is Gaussian ($C_0=0$) we have

$$\lim_{q\to\infty} q^{-2}ES_{q,\delta} \simeq \frac{2^{\delta+2}\pi\left(\frac{\sqrt{2u}}{\sigma}\right)^{-\delta-2}\Gamma\left(\frac{\delta+2}{2}\right)^2}{8\pi\sigma^2-\log(2u)}. \tag{8}$$

Now that we have established the convergence in expectation of $q^{-2}S_{q,\delta}$, it is natural to ask for more. Do we in fact have almost sure convergence? It is difficult to prove rigorously, but we can argue heuristically that the exponential decay of genetic correlation over large distances is sufficient to allow a "strong law" effect, giving us almost sure convergence. If this were the case, then we could deduce a strong limit law for the average squared/absolute distance between individuals of the same type, $S_{q,\delta}/S_{q,0}$, namely, that with probability 1,

$$\frac{S_{q,\delta}}{S_{q,0}} \to 2^{\delta}\left(\frac{\sqrt{2u}}{\sigma}\right)^{-\delta}\Gamma\left(\frac{\delta+2}{2}\right)^2, \tag{9}$$

as $q\to\infty$, for $\delta=1, 2$. Note that the limiting quantity in (9) is independent of C_0, so this result is independent of the shape of the migration distribution.

It is also worth remarking that the above calculations (suitably normalized) are also valid if individuals are sampled from only a limited number of sites in the sampling area, provided this done with sufficient regularity (for example, every second or third site) and the sampling is sufficiently "dense" relative to u and σ^2 so that the integral approximation is justified. We have observed by simulation that the same is true if the individuals are selected at random from the sampling area, which is often a more realistic situation. If the proportion of individuals sampled is p, then the statistics should be divided by p^2 to have expectations given by the above. More generally, to apply the results given in this section or in the appendix for d dimensions, the appropriate normalizing factor is p^{2d}.

2.2. *Branching Diffusion Model*

In this section we assume that the individuals are haploid, are initially dispersed randomly throughout $\mathbb{R}^d$ according to a uniform Poisson random field, and reproduce according to a critical homogeneous Markov branching diffusion process with transition/migration density $p(x, y, t)$, and mean 1 exponential lifetimes. We also assume that this population is subject to selectively neutral mutations occurring at a constant rate u per individual per generation. This model is described in detail by Sawyer (1976a), and in some sense it is the natural analog of the stepping stone model without local density regulation.

It is notationally convenient to represent the individuals alive at each time t by an (infinite) random measure μ_t on $\mathbb{R}^d$, in such a way that for each bounded measurable subset $A \subset \mathbb{R}^d$, $\mu_t(A)$ is the number of individuals alive in the set A at time t. (In the notation of Sawyer (1976a), $\mu_t(A) = N_A(t)$.) The quantities of interest can now be defined as

$$T_{A,\delta}(t) = \iint_{A^2} |x-y|^{\delta} 1_{B_t}(x, y)\, \mu_t(dx)\, \mu_t(dy), \tag{10}$$

where

$$B_t = \{(x, y)\text{: there exist, at time } t\text{, i.b.d. individuals at locations } x \text{ and } y\}.$$

We can now proceed to take expectations of the above quantity and apply the results of Sawyer (1976a) on probabilities of identity by descent for this model. But first, it is important to point out that this model is not stationary, and hence cannot be studied in equilibrium like the stepping-stone model. As time passes, the population forms clusters; these clusters become more and more dense, and fewer and farther between. But we can still look at what happens for large times, and we find that the statistics of interest to us display a certain stability over time and converge in expectation to non-trivial limits.

To calculate the expectation of $T_{A,\delta}(t)$ we note that in the notation of Sawyer (1976a) we can write (informally)

$$\begin{aligned} E1_{B(t,x,y)}\mu_t(dx)\, \mu_t(dy) &= P(\exists \text{i.b.d. individuals at } x \text{ and } y \text{ at time } t) \\ &= C(t, x, y)\, w(t, x, y)\, dx\, dy, \end{aligned}$$

where $C(t, x, y)$ is the probability that individuals found at locations x and y are identical by descent, and $w(t, x, y)\, dx\, dy$ is the probability that there are individuals at locations x and y to be found, both formally defined in Sawyer (1976a) as the limiting quantities

$$\begin{aligned} C(t, x, y) &= \lim_{A \to x, B \to y} P(\text{some pair in } A, B \text{ is i.b.d.} \mid \mu_t(A), \mu_t(B) > 0), \\ w(t, x, y) &= \lim_{A \to x, B \to y} \frac{P(\mu_t(A) = \mu_t(B) = 1)}{\lambda(A)\, \lambda(B)}, \end{aligned} \tag{11}$$

where λ denotes Lebesgue measure on $\mathbb{R}^d$. So we have

$$ET_{A,\delta}(t) = \iint_{A^2} |x-y|^{\delta}\, C(t, x, y)\, w(t, x, y)\, dx\, dy. \tag{12}$$

We consider the case when the basic migration process is Brownian motion in $\mathbb{R}^d$ with infinitesimal variance σ^2, i.e.,

$$p(t, x, y) = (2\pi\sigma^2 t)^{-d/2} \exp[-(x-y)^2/2\sigma^2 t].$$

We assume that the mean population density is unity: results for other population densities can be obtained by rescaling units of length. It follows from the results of Sawyer (1976a) that in this case

$$C(t, x, y)\, w(t, x, y) = \int_0^t e^{-2us} p(2s, x, y)\, ds.$$

So if we set

$$g_d(2u, x-y) = \int_0^\infty e^{-2us} p(2s, x, y)\, ds,$$

we obtain

$$\lim_{t\to\infty} ET_{A,\delta}(t) = \iint_{A^2} |x-y|^\delta\, g_d(2u, x-y)\, dx\, dy. \tag{13}$$

Again, we restrict our attention to the two-dimensional case and, to avoid repetition, refer the reader to the Appendix for analogous results when $d=1$ or 3. In this case we have (see, for example, Sawyer, 1976a)

$$g_2(2u, x-y) = \frac{1}{2\pi\sigma^2} K_0\left(|x-y| \frac{\sqrt{2u}}{\sigma}\right).$$

Set $A_q = [0, q]^2$ and write $_{q,\delta}(t)$ for $T_{A_q,\delta}(t)$. Then

$$\lim_{t\to\infty} ET_{q,\delta}(t) = \frac{1}{2\pi\sigma^2} \iint_{A_q^2} |x-y|^\delta\, K_0\left(|x-y| \frac{\sqrt{2u}}{\sigma}\right) dx\, dy. \tag{14}$$

But this is the same integral with which we approximated $ES_{q,\delta}$ for the stepping stone model, up to a multiplying factor (and with the lower bound $\varepsilon = 0$). It can be simplified as before to get

$$\begin{aligned}\lim_{t\to\infty} ET_{q,\delta}(t) \simeq \frac{2}{\pi\sigma^2}\Bigg(&\int_0^q r^{\delta+1} K_0\left(\frac{\sqrt{2u}}{\sigma} r\right) \\ &\times\left[\frac{\pi}{2} q^2 - 2qr + \frac{r^2}{2}\right] dr + \int_q^{\sqrt{2}q} r^{\delta+1} K_0\left(\frac{\sqrt{2u}}{\sigma} r\right) \\ &\times\left[\left(\frac{\pi}{2} - 2\cos^{-1}\frac{q}{r}\right) q^2 + 2q\sqrt{r^2-q^2} - q^2 - \frac{r^2}{2}\right] dr\Bigg),\end{aligned}$$

and we get an analogous limiting result

$$\lim_{q\to\infty} q^{-2} \lim_{t\to\infty} ET_{q,\delta}(t) \simeq 2^{\delta}\sigma^{-2}\left(\frac{\sqrt{2u}}{\sigma}\right)^{-\delta-2}\Gamma\left(\frac{\delta+2}{2}\right)^{2}. \tag{15}$$

Note that when u is small and σ^2 is sufficiently large (what "sufficiently large" means will depend on the value of u), these results agree with the corresponding results for the stepping stone model with one haploid individual per colony and Gaussian migration (Eqs. (6) and (8)). To illustrate this, set

$$r(\sigma^2, u) = \frac{\lim_{t\to\infty} ET_{q,\delta}(t)}{ES_{q,\delta}}, \tag{16}$$

where we assume Gaussian migration in both cases, and for the stepping-stone model that we have one haploid individual per colony. Then comparing (6) and (14) we see that

$$r(\sigma^2, u) = \frac{4\pi\sigma^2 - \log(2u)}{4\pi\sigma^2}.$$

If it were possible to formulate a strong limit law for the ratio $T_{q,\delta}(t)/T_{q,0}(t)$ as q and t tend to infinity, simultaneously perhaps, then we would expect the same limit as in the analogous stepping stone model for small u and for any value of $\sigma > 0$.

3. Simulation Model

The simulation model assumes a two-dimensional ($L \times L$) lattice of locations each of which contains a single diploid individual. Each individual is characterized by the two alleles at a single diploid locus. Generations are non-overlapping and all alleles are assumed to be neutral. In each generation, each individual is assumed to "choose" both of its alleles from the individuals in the preceeding generation according to a specified distribution of dispersal distances. The simulation program used the coalescent approach described by Hudson (1990) and modified appropriately. Initially, there is a sample of genes drawn from specified geographic locations. Then the program simulated the ancestry of the genes in this sample to produce a gene genealogy. Once the genealogy is obtained, mutations then occur from an assigned state of the ancestor to yield the allelic states of all the genes in the sample. There are two advantages to this approach. First, it is unnecessary to deal with genes that are not ancestral to one of the genes in the sample, which makes the simulation much more efficient.

Second, because the entire gene genealogy is simulated in each replicate, there is no question that a stochastic equilibrium is reached. For the dispersal stage, we assumed that the probability of dispersal depended only on the distance separating two locations and used discretized versions of either an exponential or Gaussian distribution of dispersal distances in each direction. That is, for points that are i steps apart in one direction and j steps in the other, the probability of dispersing that distance is $f_i f_j$, where f_i is the one-dimensional distribution. At the edges of the lattice, we assumed a reflecting boundary. In the simulation, when the same gene was chosen as the ancestor of two genes in the next generation, a coalescent event occurred and the number of genes in the genealogy was reduced by one. If a third gene choose the same ancestor, another coalescent event would occur producing a trifurcation in the tree.

This simulation model differs from Hudson's (1990) because we did not assume that only one coalescent event could occur in each generation. That assumption would be incorrect for the parameter values we used. Once the gene genealogy was obtained in a replicate, we assigned an ancestral state to the root and then let mutations accumulate at each node, with the number of mutations occurring on a branch being a Poisson distributed random variable with mean uT, where u is the mutation rate per generation and T is the branch length in generations.

We assumed two different mutation models. The "infinite alleles model" assumed that each mutant was new to the population, which implies that all alleles in the same allelic class are identical by descent. We used this model for comparison with the analytic results described above, which assumed the infinite alleles model. We also assumed the "stepwise mutation" model which was originally used in the analysis of electrophoretic data (Ohta and Kimura, 1973; Wehrhahn, 1975). This model may be appropriate for microsatellite loci (Valdes, Slatkin, and Freimer, 1993). In the stepwise mutation model, the allele at the root is assigned to be of some state, which represents the repeat number; we can arbitrarily set that number to 0. When a mutation occurs, the repeat number changes by 1. A more general model would assume changes of more than one step, but the one-step model represents the extreme case in which it is most likely that alleles in the same allelic class might not be identical by descent. The question is whether this assumption leads to any significant differences from the infinite alleles assumption.

For each case we simulated, we assumed that genes were sampled from a square quadrat of length q on a side which was centered in the $L \times L$ lattice. We allowed for two possibilities for the distribution of alleles sampled from a quadrat, they could be either regularly spaced or randomly chosen. If they were regularly spaced, we assumed a minimum spacing of h, so the sample size is $n = (q/h)^2$. If $h = 1$ every individual in the $q \times q$ quadrat was

sampled. If individuals were randomly chosen, we specified, n, the sample size and sampled without replacement from the $q \times q$ quadrat. We assumed that one gene from in each individual was sampled from each individual in the sample.

For each case, the parameter values needed were L, the size of the lattice, q, the size of the quadrat from which individuals were sampled, u, the mutation rate, and σ^2, the variance in dispersal distances. In addition we had to specify the functional form of the dispersal distribution, either exponential or Gaussian, and we had to specify either h, the spacing of samples when samples were regularly spaced or, n, the number of individuals sampled when individuals were randomly sampled. At the end of each replicate the values of $(q^2/n^2) S_{q,\delta}$ for $\delta = 0, 1$, and 2 were computed. These values were then averaged over a specified number of replicates, usually 100, to obtain the results for that set of parameter values.

3.1. *Simulation Results*

We restricted our simulations to parameter values for which the value of L, the size of the lattice did not matter. We found that for relatively small values of σ^2, 2, 4, or 8, and values of q of 50, 100, or 200, that the results did not depend on L as long as $L \geqslant 2q$. Therefore we set $L = 2q$ in all the simulations. We also found that there was no detectable difference between the Gaussian and exponential distributions of dispersal distances, as is consistent with the analytic theory. Therefore, we used an exponential distribution for most of the simulations.

Table I shows the results for our simulations in the case in which every individual in the quadrat was sampled ($h = 1$). The predictions of the analytic theory for both the stepping-stone and branching diffusion models are included for comparison. As we can see, there is fair agreement between the simulation results and the analytic results for the stepping stone model (Eq. (6)) and the agreement is better for smaller mutation rates. The exact results for the branching diffusion model (Eq. (14)) are larger, as would be expected from Eq. (17). Clearly the asymptotic results (Eqs. (7) and (15)) are not very good for these parameter values.

Table II shows that the details of the sampling scheme are not important for the values of $(q^2/n^2) S_{q,\delta}$ for $\delta = 2$ or $\delta = 1$, as predicted by the analytic theory. However, the value of $(q^2/n^2) S_{q,0}$ does depend on n. That reveals a limitation of the analytic theory as an approximation for these parameter values. The dependence on n is less for smaller u.

Table III shows the results for one set of parameter values with different sample sizes under random sampling. The results for $\delta = 2$ and $\delta = 1$ are consistent with the analytic prediction that they should be independent of sample size. In effect, the values of an integral are being estimated by doing

TABLE I

$(1/q^2)\,E[S_{q,\delta}]$ from Simulations and Analytic Theory

					Stepping stone		Branching diffusion	
q	u	σ^2	δ	Simul.	Eq. (6)	Eq. (7)	Eq. (14)	Eq. (15)
50	0.01	2	2	2327.8	3000.41	9277.92	3467.44	10722
			1	140.2	172.13	364.34	198.92	421
			0	11.8	14.8	23.2	17.1	26.8
50	0.005	2	2	5897.6	7046.4	36642.9	8337.5	43357
			1	300.2	344.2	1017.5	407.3	1203.9
			0	21.5	24.5	45.8	29.04	54.2
50	0.001	2	2	27979.3	29340.9	889968	36596	1.1×10^6
			1	1123.7	1151.6	11051.8	1436.4	13785
			0	61.1	62	222.5	77.33	277.5
100	0.005	2	2	13242.6	17181.1	36642.9	20329.2	43357
			1	502.0	609.9	1017.5	721.7	1203.9
			0	28.69	33.6	45.8	39.8	54.2
100	0.005	4	2	19946.0	24738.1	76495.8	27604.5	83504
			1	612.4	710.4	1501.99	775.4	1639.6
			0	27.69	30.98	47.8	33.8	52.2
100	0.001	2	2	124750	140125	889968	174773	1.1×10^6
			1	3019.1	3284.1	11051.8	4096.2	13785
			0	106.34	112.2	222.5	140	277.5
100	0.001	4	2	136405	144773	1.9×10^6	162672	2.1×10^6
			1	3003.5	3057.6	16539.6	3435.6	18584
			0	92.92	91.3	235.5	102.6	264.5
200	0.01	2	2	5673.2	7308.4	9277.92	8445.9	10722
			1	239.2	307.7	364.34	355.6	421
			0	16.23	20.6	23.2	23.8	26.8
200	0.001	2	2	305,069	375,869	889,968	468,809	1.1×10^6
			1	5368.2	6198.5	11051.8	7731.2	13785
			0	146.56	158.8	222.5	198.1	277.5
200	0.04	4	2	856	1033.06	1219.36	1084.97	1280.6
			1	41.4	60.1	67.7	63.1	71.1
			0	3.06	5.52	6.1	5.8	6.4
200	0.02	4	2	2879.9	3816.4	4844.87	4060.77	5155.1
			1	118.3	160.67	190.26	170.96	202.44
			0	7.62	10.8	12.11	11.45	12.89

TABLE II

$(q^2/n^2)\, S_{q,\delta}$ for Random and Uniformly Distributed Samples

	$\delta = 2$	$\delta = 1$	$\delta = 0$
$\sigma^2 = 2$, $q = 200$, average of 100 replicates			
$u = 0.01$			
$h = 1$ ($n = 40{,}000$)	5673.2	239.2	16.2
$h = 4$ ($n = 2{,}500$)	5248.0	246.4	32.6
Random ($n = 2{,}500$)	5174.4	244.8	33.4
$u = 0.001$			
$h = 1$ ($n = 40{,}000$)	305,069	5368.2	146.6
$h = 4$ ($n = 2{,}500$)	301,587	5386.6	163.8
Random ($n = 2{,}500$)	299,152	5267.2	161.9

a Monte Carlo integration with fewer and fewer points. The expectation remains the same, although the variation among replicates increases with decreasing sample sizes, as expected. The results for $\delta = 0$ do not fit the analytic predictions. For these parameter values, smaller sample sizes result in missing low frequency alleles and overrepresenting the higher frequency alleles. That results in larger than expected values of $S_{q,0}$. There is the same bias in $S_{q,2}$ and $S_{q,1}$ but the low frequency alleles contribute so little to the sums of distances that the effect is not apparent.

We ran a few cases with the stepwise mutation model and found that there is a substantial difference between the results for that model and those for the infinite alleles model for the same parameter values. Some

TABLE III

Comparison of Simulation Results for Different Numbers of Individuals Randomly Sampled

Sample size	$\delta = 2$	$\delta = 1$	$\delta = 0$
$\sigma^2 = 2$, $\mu = 0.01$, $q = 50$ (averages over 100 replicates)			
2500	5174.4	244.8	33.4
2000	5162.0	242.0	37.2
1500	5306.7	250.7	44.3
1000	5184.0	248.0	57.8
500	5171.6	240.0	97.4
300	4711.0	231.3	152.0
100	5211.0	240.1	421.7

Note. The values given are $q^2 S_{q,\delta}/n^2$.

TABLE IV

Comparison of Simulation Results for $S_{q,\delta}/q^2$ in the Infinite Alleles and Stepwise Mutation Models

	$\delta = 2$	$\delta = 1$	$\delta = 0$
$\sigma^2 = 2$, $\mu = 0.01$, $q = 50$, $h = 1$ $(n = 2500)$			
Infinite alleles	2327.8	140.2	11.8
Stepwise	38,725.9	1318.4	60.3
$\sigma^2 = 2$, $\mu = 0.001$, $q = 50$, $h = 1$ $(n = 2500)$			
Infinite alleles	27,979.3	1123.7	61.1
Stepwise	118940.3	3874.7	160.0

Note. Averages over 100 replicates.

results are shown in Table IV. The values of all three statistics are much larger than in the infinite alleles model, as might be expected because that model allows the possibility that an allelic class can have a much wider geographic distribution than in the infinite alleles mode.

4. Discussion and Conclusions

Our results show that there are some relatively simple statistics describing the area occupied by different alleles. As our measure of area we used both the mean square distance between all copies of an allele and the root mean square distance. As statistics, we used the sums of mean square ($S_{q,2}$) or root mean square ($S_{q,1}$) distances and the difference between the sum of the squares of the numbers of alleles in each class and the total number of individuals sampled ($S_{q,0}$). The reason for using these statistics is that they can be predicted by the analytic theory. Either of the ratios $S_{q,2}/S_{q,0}$ and $S_{q,1}/S_{q,0}$ could be thought of as the average area occupied by alleles at a locus. It is important to realize, however, that these definitions of average area differ from the usual definition of average area, in which the average mean square or root mean square distances are computed for each allelic class and then the overall average is taken. That average cannot be easily related to analytic theory and differs substantially from what can be computed analytically. Although it would be possible to use simulations alone to find the dependence of the average area on the other parameters, we feel it is preferable to use statistics whose expectations can be found.

Our results are not restricted to high mutation rati loci but it seems likely that they will be most applicable to such loci. Relatively high mutation rates imply that relatively small areas will be occupied by different

alleles, which means that boundary effects will be less important. Furthermore, in many high mutation rate loci, u can be estimated so there is at least hope that observations of the geographic distributions of alleles can lead to estimates of the neighborhood size (which is proportional to the variance in dispersal distances multiplied by population density). The analytic theory shows that population density cannot be estimated separately using this approach. Although no such geographic surveys have yet been completed some are being carried out now.

One advantage to our results is that the analytic theory provides some tests of consistency of the underlying theory. In particular, the statistic $S_{A,0}$ should be inversely proportional to the mutation rate if the assumptions of the model are valid. That could be tested by making comparisons across loci. Also, given the value of $S_{q,0}$ the relationship between $S_{q,1}$ and $S_{q,2}$ is predicted by Eq. (9).

Our results also show that this approach to analyzing geographic patterns will be useful only in cases in which the infinite alleles model is reasonable. That may be true for VNTR loci in which variation between repeats can be detected. That is probably not true for microsatellite loci which might well fit the simple stepwise mutation model.

There are many potential problems with applying our approach to data. The assumption of a homogeneous population uniformly distributed in an region and having the same dispersal tendencies throughout the region is unlikely to be valid for any real species. Furthermore, except in plantations or other artificially constructed populations, it is unlikely that individuals can be sampled randomly or with even spacing. Instead samples are taken opportunistically and sample sizes for an area vary because of unforseen conditions. Nevertheless, our results show that the geographic distributions of alleles at loci for which mutation rates are known do provide information about dispersal. Generalizations of our results that do take spatial inhomogeneities into account could lead to further understanding of dispersal, particularly if information from different loci is combined.

APPENDIX: Analytic Results in One and Three Dimensions

For the Stepping-Stone Model

In one dimension, if we assume that the migration density has finite fifth moments and variance σ^2, then (Sawyer, 1977, Th. 1)

$$I(x, u) = \frac{\exp(-|x|\sqrt{2u}/\sigma)}{1 + \sqrt{2u}(4N\sigma + C_0)} + O(u) \tag{17}$$

uniformly for $0 < \varepsilon \leqslant |x|\sqrt{u} \leqslant M < \infty$, for any $\varepsilon > 0$, $M < \infty$, where the constant C_0 depends on the migration density. For Gaussian migrations

$C_0 = -0.8238$ is independent of σ. Proceeding as in the two-dimensional case, if we set $A_q = [0, q] \cap \mathbb{Z}$ and write $S_{q,\delta}$ for $S_{A_q,\delta}$, we can apply Sawyer's approximation (17) and replace the sum by an integral to get

$$ES_{q,\delta} \simeq \iint_{A_q^2} |x-y|^{\delta} \frac{\exp(-|x|\sqrt{2u}/\sigma)}{1+\sqrt{2u}\,(4N\sigma + C_0)} \tag{18}$$

which can be calculated explicitly. In particular,

$$\begin{aligned} q^{-1}ES_{q,0} &= q^{-1}[1+\sqrt{2u}\,(4N\sigma + C_0)]^{-1} \\ &\quad \times \sqrt{\frac{2}{u}}\,\sigma\left[q + \frac{\sigma}{\sqrt{2u}}(e^{-\sqrt{2u}\,q/\sigma} - 1)\right] \\ &\to 1+\sqrt{2u}\,(4N\sigma + C_0)]^{-1}\sqrt{\frac{2}{u}}\,\sigma, \end{aligned}$$

and

$$\begin{aligned} q^{-1}ES_{q,1} &= q^{-1}[1+\sqrt{2u}\,(4N\sigma + C_0)]^{-1} \\ &\quad \times \left[\frac{\sigma^2}{u}q(e^{-\sqrt{2u}\,q/\sigma}+1) + \sqrt{2}\,\sigma^3 u^{-3/2}(e^{-\sqrt{2u}\,q/\sigma} - 1)\right] \\ &\to [1+\sqrt{2u}\,(4N\sigma + C_0)]^{-1}\frac{\sigma^2}{u}. \end{aligned}$$

As in the two-dimensional case, we conjecture that $q^{-1}S_{q,0}$ and $q^{-1}S_{q,1}$ converge almost surely to the above quantities, in which case we could deduce that with probability 1,

$$\frac{q^{-1}S_{q,1}}{q^{-1}S_{q,0}} \to \frac{\sigma}{\sqrt{2u}}.$$

In three dimensions, if we assume that the migration density has finite fourth moments and (for notational convenience) is symmetric with variance σ^2 in each direction, then (Sawyer, 1977, Th. 3)

$$I(x,u) = \frac{\exp(-|x|\sqrt{2u}/\sigma)}{4\pi(|x|/\sigma)(2N\sigma^3 + C_0)}(1 + O(u^{1/2})) \tag{19}$$

uniformly for $0 < \varepsilon \leqslant |x|\sqrt{u} \leqslant 1/\varepsilon$, for any $\varepsilon > 0$, where the constant C_0 depends on the migration density. For Gaussian migrations $C_0 = \sum_1^{\infty}{}' (4\pi u)^{-3/2}$ is independent of σ, As before, we can use (19) in (3) and approximate by an integral to calculate $ES_{A,\delta}$.

For the Branching Diffusion Model

In one dimension we have

$$g_1(2u, x-y) = \frac{\exp(-|x-y|\sqrt{2u}/\sigma)}{2\sigma\sqrt{2u}},$$

so if we set $A_q = [0, q]$ we get

$$\lim_{t\to\infty} ET_{q,\delta}(t) = \iint_{A_q^2} |x-y|^\delta \frac{\exp(-|x-y|\sqrt{2u}/\sigma)}{2\pi\sigma^2\sqrt{2u}}\,dx\,dy,$$

which again, up to a multiplying factor, is the same integral we obtained for the stepping stone model (18), and just as in two dimensions, the results we obtain here will approximately agree with those of the stepping-stone model with one haploid individual per site and Gaussian migration provided u is small and σ^2 sufficiently large.

In three dimensions,

$$g_3(2u, x-y) = \frac{\exp(-|x-y|\sqrt{2u}/\sigma)}{4\pi|x-y|\,\sigma^2}$$

and similar remarks apply.

ACKNOWLEDGMENTS

This research was supported in part by NIH Grant GM40282 to M. S. and by NSF Grant MCS90-01710 to N.O'C. We thank F. Bonhomme and J. F. Dallas for asking questions that led to the analysis in this paper. We also thank S. Sawyer for very helpful discussions of this topic, and S. A. Frank for letting us use one of his computers to run some of the very large simulations.

REFERENCES

BOERWINKLE, E., XIONG, W., FOUREST, E., AND CHAN, L. 1989. Rapid typing of tandemly repeated hypervariable loci by the polymerase chain reaction: Application to the apolipoprotein B3′ hypervariable region, *Proc. Natl. Acad. Sci., U.S.A.* **86**, 212–216.

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

GRADSHTEYN, I. S., AND RYZHIK, I. W. 1965. "Table of Integrals, Series and Products," Academic Press, New York.

HUDSON, R. R. 1990. Gene genealogies and the coalescent process, *in* (D. J. Futuyma and J. Antonovics, Eds.), Vol. 7, pp. 1–44, "Oxford Surveys in Evolutionary Biology" Oxford Univ. Press, Oxford.

JEFFREYS, A. J., MACLEOD, A., TAMAKI, K., NEIL, D. L., AND MONCKTON, D. G. 1991. Minisatellite repeat coding as a digital approach to DNA typing, *Nature* **354**, 204–209.

JEFFREYS, A. J., NEUMANN, R., AND WILSON, V. 1990. Repeat unit sequence variation in minisatellites: A novel source of DNA polymorphism for studying variation and mutation by single molecule analysis, *Cell* **60**, 473–485.

MALÉCOT, G.1968. "The Mathematics of Heredity," Freeman, San Francisco.

NICHOLS, R. A., AND BALDING, D. J. 1991. Effects of population structure on DNA fingerprint analysis in forensic science, *Heredity* **66**, 297–302.

NAKAMURA, Y., LEPPERT, M., O'CONNELL, M., WOLFF, P., HOLM, R., CULVER, T., MARTIN, M., FUJIMOTO, E., HOFF, M., KUMLIN, E., AND WHITE, R. 1987. Variable number of tandem repeat (VNTR) markers for human gene mapping, *Science* **235**, 1616–1622.

OHTA, T., AND KIMURA, M. 1973. The model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a genetic population, *Genet. Res.* **22**, 201–204.

SAWYER, S. 1967a. Branching diffusion processes in population genetics, *Adv. Appl. Prob.* **8**, 659–689.

SAWYER, S. 1976b. Results for the stepping stone model for migration in population genetics, *Ann. Prob.* **4**, 699–728.

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

SAWYER, S. 1979. A limit theorem for patch sizes in a selectively-neutral migration model, *J. Appl. Prob.* **16**, 482–495.

VALDES, A. M., SLATKIN, M., AND FREIMER, N. B. 1993. Allele frequencies at microsatellite loci: the stepwise mutation model revisited, *Genetics* **133**, 737–749.

WEHRHAHN, C. 1975. The evolution of selectively similar electrophoretically detectable alleles in finite natural populations, *Genetics* **84**, 639–659.

WEISS, G. H., AND KIMURA, M. 1965. A mathematical analysis of the stepping stone model of genetic correlation, *J. Appl. Prob.* **2**, 129–149.

Printed in Belgium
Uitgever: Academic Press, Inc.
Verantwoordelijke uitgever voor België:
Hubert Van Maele
Altenastraat 20, B-8310 Sint-Kruis