

# Mutation at high rates reduces spatial structure within populations

BRYAN K. EPPERSON

126 Natural Resources Building, Michigan State University, East Lansing, Michigan 48824

## Abstract

The question of whether or not the high rates ( $\mu$ ) of mutation that occur for some hypervariable markers can affect commonly used empirical measures of spatial structure of genetic variation within populations is addressed. The results show that values of these measures are approximately halved when  $\mu$  is  $10^{-2}$ . Finest spatial-scale correlations, measured by either Moran's  $I$ -statistics or conditional kinship, are reduced by 30%–50%. When the mutation rate is 10 times lower, much smaller reductions result, e.g. averaging 7% for the finest scale correlations. Still smaller orders of magnitude of  $\mu$  cause negligible changes in spatial structure, where any effects normally would not be detectable. The reductions are caused by forward mutations, and when the reductions are measured as percentages, they are nearly independent of the amount of structure produced sans mutation, except when dispersal is nearly minimal. The percent reductions are also nearly independent of the number of alleles and of back mutations, hence of the nature of the mutation process (e.g. stepwise or not). The results demonstrate that some hypervariable loci should have reduced spatial structuring, and that marker choice may affect the values observed in experimental surveys. Moreover, if fine-scale correlations are used to indirectly estimate dispersal distances, then mutation at high rates could inflate estimates, easily up to two- to three-fold.

*Keywords:* Dispersal, hypervariable loci, microsatellite, mutation, spatial autocorrelation, spatial structure

*Received 11 June 2004; revision received 7 October 2004; accepted 4 November 2004*

## Introduction

Hypervariable loci are increasingly being used for population genetic studies. Often, loci are chosen by virtue of having the greatest amount of variation, in part because high variation can mean greater information on population parameters. For example, microsatellite markers having 50 or more alleles have been assayed in studies of patterns of genetic variation (e.g. Morand *et al.* 2002). Hypervariability is generally caused by exceptionally high mutation rates. Thus, while such molecular markers offer many advantages that support their popular uses, it seems possible that in some instances mutation rates can be high enough to violate assumptions made in analyses and interpretations. Mutation is generally considered to be a weak force in evolution, but it may not always be the case that direct effects of mutations are safely ignored.

Studies of the spatial distribution of genotypes within populations have increased rapidly in recent years (recent major reviews include Epperson 2003 and Escudero *et al.* 2003), in part because isolation by distance is an important population genetic process. Spatial distributions are also used to provide indirect measures of dispersal (reviewed in Heywood 1991), and sometimes to identify natural selection and other forces. Well over 50 studies have used either Moran's  $I$ -statistics or spatial covariances (Loiselle *et al.* 1995) to analyse spatial structures within populations. Hypervariable loci are increasingly being employed in such studies. As the number of alleles becomes large, the spatial patterns for each become nearly independent (Epperson 2004) realizations of a (usually highly stochastic) process. Thus, measures of precision such as the standard error of an average correlation (i.e. averaged over alleles) decrease by essentially the square root of the number of alleles.

There are several reasons to suspect that some hypervariable loci might exhibit less spatial structure, in particular degree of isolation by distance, and that their use could also lead to overestimation of dispersal. One theoretical

Correspondence: Bryan K. Epperson, Fax: (517) 432 1143; E-mail: epperson@msu.edu

study showed that mutation at rates on the order of  $10^{-2}$  causes significant reductions in values of one particular measure of spatial structure, but at rates of  $10^{-3}$  has little effect on spatial structure (Epperson 1990). Although the measure used in that study is different from those used in empirical studies of spatial distributions within populations, the results suggest that the critical range is  $10^{-3}$ – $10^{-2}$ . This range (as well as lower rates) is observed among hypervariable markers such as microsatellites (e.g. Bruford *et al.* 1992; Jarne & Lagoda 1996; Udupa & Baum 2001). Possible empirical evidence for effects of high mutation rates has been found in contrasts among loci in populations of *Pinus strobus*. Spatial autocorrelation was significantly lower for the one microsatellite that had by far the greatest diversity, compared to less variable microsatellites (Walter & Epperson 2004; Marquardt & Epperson 2004) and allozymes (Epperson & Chung 2001). However, in a study of *Quercus*, isozymes had less structure than microsatellites, although the authors attributed the differences to statistical error (Streiff *et al.* 1998). Nonetheless, it seems possible that choice of markers could affect inferences of structure and dispersal.

Most empirical studies map individual samples and then convert diploid genotypes into frequencies of a given allele (Escudero *et al.* 2003). Typically, genotypes are converted into the values 0, 0.5, or 1.0 according to the numbers (none, one, two) of a particular allele that are carried in the genotype (e.g. Heywood 1991; Epperson 2003). In many studies, Moran's *I*-statistics are calculated for the numerical values, separately for each allele, and then very precise averages over alleles and loci can be obtained. In other studies, either single allele or multiallele multilocus estimates of the genetic covariance, *R* (Cockerham 1969; Loiselle *et al.* 1995), also termed the conditional kinship measure, are obtained. Single allele *I*-statistics and *R*-statistics can be easily interconverted, if the fixation index is known (Hardy & Vekemans 1999). Most multiallele/multilocus estimates of *R* are linear functions of *I*-statistics (Epperson 2003, 2004).

*I*-statistics, *R*-statistics, and other methods based on pairs of individuals are highly efficient, and their theoretical properties have been well characterized under isolation by distance processes (sans mutation) (e.g. Sokal & Wartenberg 1983; Epperson 1995). Their sampling properties are also well known (Epperson *et al.* 1999). Shortest-distance correlations can be particularly useful. Precise and robustly predicted values are available for different processes, and differences can often be detected. For example, one study using large numbers of allozyme loci found that the average Moran's *I*-statistic for shortest distance differed between two palm species, even though both values were small, 0.0085 and 0.0387 (Luna *et al.* 2005). The difference was tied to contrasts in life history, density, and dispersal. Recently, the joint distribution of *I* (or *R*) statistics for alleles of the same locus has been elucidated, which allows multiallelic, multilocus average values to have known

distributions (Epperson 2004). This further improved ability to obtain precise estimates and to detect differences among loci (e.g. highly mutable vs. less mutable loci) and other contrasts. By combining multiple alleles and/or loci, even modest-sized samples can support high statistical precision (see Epperson *et al.* 1999). However precise these metrics can be, if mutation reduces the observed values, then misleading results will be obtained.

This article characterizes the influences of high mutation rates on the short-distance spatial correlations within populations undergoing isolation by distance processes (Wright 1943). A large number of space–time simulations are run in sets designed to determine the effects of mutation [jointly with the other major factor, the amount of dispersal as measured by Wright (1943) neighbourhood size], on the measures used in empirical studies. The focus is on Moran's *I*-statistics, but the results are extended to the other commonly used estimators in the Discussion.

Analytic studies of coalescence in isolation by distance processes in continuous populations that exist in two spatial dimensions have shown that spatial structure of genetic variation is created primarily by recent coalescences among spatially proximal genes (Barton & Wilson 1995; Barton *et al.* 2002). If forward mutations occur at high enough rates, they might substantially reduce identity by descent among genes that recently coalesce, and hence reduce spatial autocorrelation. Back mutations are expected to have little or no effect on spatial autocorrelations (as is discussed further in the Discussion). Hence, a *k*-allele mutation model should be sufficient. By conducting simulations with varying numbers of alleles, and thus varying rates of back mutation, lack of effect of back mutations is further demonstrated in this article. Because the nature of back mutations is what distinguishes various possible mutation models (including the stepwise mutation model) with respect to spatial genetic structure, the reductions observed are valid for all loci, including microsatellites.

## Methods

### Simulations

Each simulated population consisted of 10 000 individuals with diploid genotypes, located on a  $100 \times 100$  square lattice (a conceptual approximation to continuous space), and was initialized with a random distribution of diploid genotypes at an arbitrary locus, in Hardy–Weinberg proportions. Different sets of simulations had numbers of alleles,  $n_a$ , set equal to either two, five or 10, and each simulation had equal initial frequencies of all alleles. Allele frequencies changed very little during the course of a simulation. Previous simulation studies showed that numbers of alleles and a wide range of frequencies have no effects on values of Moran's *I*-statistics (Epperson *et al.* 1999), although

when frequencies are reduced to  $c. 0.02$ , some reduction of autocorrelation results (Epperson 2003). All sets of simulations included some level of restriction on dispersal (see succeeding discussion). Most also incorporated a 'random replacement' process, where each gene had probability  $\mu$  of being replaced and each of the other ( $n_a - 1$ ) alleles had equal probability of being the replacement. With respect to the spatial structure of a given allele, there are two conceivable influences of mutation, the effect of forward mutations of this allele to other alleles, and the effect of back mutations to it from other alleles with rate  $\mu/n_a$ . Sets of simulations had  $\mu$  equal to 0,  $10^{-3}$  or  $10^{-2}$ . The values  $10^{-2}$  and  $10^{-3}$  appeared to correspond to large and negligible effects, respectively, in previous work (Epperson 1990), and this was further evident in the present study, thus smaller nonzero mutation rates were not analysed. By varying  $\mu$  and  $n_a$  (and thus back mutation rate) among sets, separation of the effects of forward mutations from possible effects of back mutation was achieved. The mutation model ignores the stepwise nature of many mutations of microsatellites, but this is sufficient because: (i) the focus is on the spatial distribution of each allele; and (ii) the effect of back mutations is shown to be negligible.

Sets also varied according to four dispersal models that together represent a wide range of dispersal levels from near minimal to large neighbourhood sizes. [Other details of the simulation program, which uses Monte Carlo methods to simulate stochastic generations of life cycles, were described previously (Epperson 1990).] Limited dispersal is modelled in the following way. Either or both the female and male parents of an offspring were chosen at random (using two uniform pseudo-random numbers to choose the two coordinates for each parent) generally from one of the nearest  $N_f$  and  $N_m$  (respectively) neighbours including self. Thus, each individual within the group of size  $N_f$  and  $N_m$  had equal chances of being the female or male parent, respectively. This may be considered unrealistic for many species, in which the probability of dispersal decays with distance. However, it is justified by the fact that the form of the dispersal curve has very little effect on spatial structure, rather it is the standardized neighbourhood size that matters, over a wide range of conditions (e.g. Rohlf & Schnell 1971). A measure very similar to Wright's (1943) neighbourhood size  $N_e$  was calculated by the variance of dispersal summed over males and females. The four dispersal models simulated had, respectively: (i)  $N_f = 1$ ,  $N_m = 9$ ,  $N_e = 4.2$ ; (ii)  $N_f = 1$ ,  $N_m = 49$ ,  $N_e = 25.1$ ; (iii)  $N_f = 49$ ,  $N_m = 49$ ,  $N_e = 50.2$ ; and (iv)  $N_f = 1$ ,  $N_m = 225$ ,  $N_e = 115.2$ . Models with  $\mu = 0$  were run only with five alleles because earlier work showed that in such cases the number of alleles has no effect on  $I$ -statistics, nor does allele frequency (Epperson *et al.* 1999), unless it is about 0.02 or smaller (Epperson 2003). Each simulation was run for 200 generations, by which time a population has been at quasi-equilibrium

**Table 1** Values of Moran's  $I$ -statistics for distance class 1, for models with different mutation rates  $\mu$ , numbers of alleles  $n_a$ , and amounts of dispersal as measured by  $N_e$ . In the lower table are shown values as proportions of those for the model without mutation

$\mu$	$n_a$	$N_e$			
		4.2	25.1	50.2	115.2
0	5	0.451	0.147	0.112	0.036
$10^{-3}$	2	0.410	0.130	0.103	0.036
$10^{-3}$	5	0.423	0.134	0.105	0.033
$10^{-3}$	10	0.425	0.136	0.105	0.034
$10^{-2}$	2	0.267	0.074	0.062	0.017
$10^{-2}$	5	0.304	0.088	0.069	0.021
$10^{-2}$	10	0.311	0.090	0.071	0.023
$10^{-3}$	2	0.91	0.88	0.92	0.98
$10^{-3}$	5	0.94	0.92	0.94	0.90
$10^{-3}$	10	0.94	0.92	0.94	0.93
$10^{-2}$	2	0.59	0.50	0.56	0.48
$10^{-2}$	5	0.67	0.60	0.62	0.58
$10^{-2}$	10	0.69	0.61	0.63	0.62

for over 100 generations. One hundred simulations for each set were run on a SUN ULTRASPARC 10, and each set required about 4 h of cpu. The combinations of parameters are displayed in Table 1. In total, 28 sets or 2800 space-time simulations were run and analysed.

#### Statistical characterization of populations

The spatial distributions of genotypes during the period of the quasi-stationary phase were characterized by computing the statistics at generation 200 for each simulation run, as has been done in most previous studies. This choice is unbiased but largely arbitrary and efficient. Any other generation in the range from about 50 to several thousand would also have been adequate, because during this period the simulated populations exist in a quasi-stationary phase (e.g. Epperson 1990). It is much more informative to replicate over entire simulations rather than over generations. Moreover, because sampling schemes have very little effect on Moran's  $I$ -statistics apart from their effects through changes in spatial scale (Epperson & Li 1996, 1997; Epperson *et al.* 1999), all 10 000 genotypes were included.

In preparation for calculation of Moran's  $I$ -statistic for the individual genotypes, for each allele, the genotype at each location  $i$  was converted into the values  $X_i = 0, 0.5$ , or 1.0 according to the numbers (none, one, two) of that allele that were carried in the genotype. For each allele, Moran's  $I$ -statistics were calculated based on these numeric values. For each simulation, pairs or 'joins' were classified according to distance classes,  $D$ , in multiples of lattice units. For example, distance class 1 was 0–1.5 units, i.e. it included

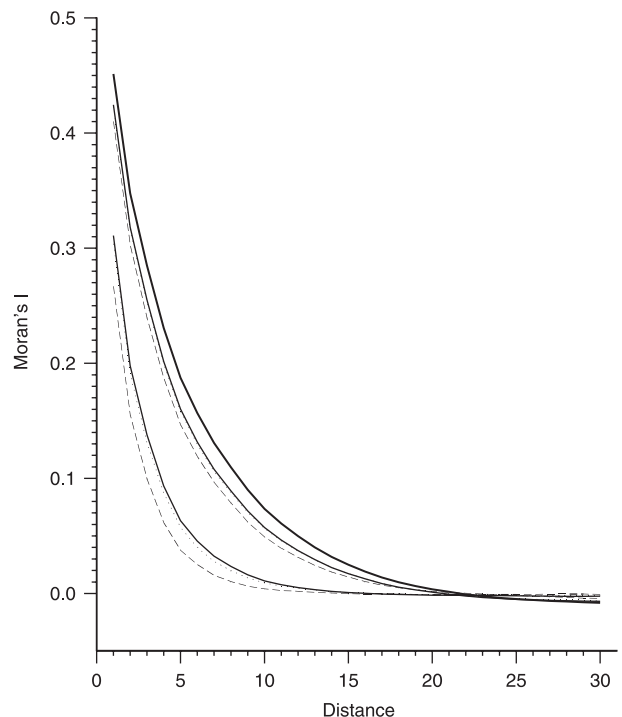
nearest neighbours (distance = 1.0), analogous to single space rook's moves in the game of chess, and diagonal neighbours (distance =  $2^{1/2}$  or *c.* 1.41). Distance class 2 included all pairs separated by distances in the range of (effectively) 1.5–2.5, and for example includes second nearest neighbours analogous to two-step rook moves, as well as other pairs. To reduce the amount of output to manage, only distance classes 1 through 30 are reported, and it is noted that by distance class 30 the correlograms are approaching asymptotic values. To calculate Moran's *I*-statistic for each distance class *D* let  $Z_i = X_i - q$ , where  $q$  is the mean of the genetic values also equal to the allele frequency in the simulation (generally very close to the initial values  $1/n_a$ , where  $n_a$  is the number of alleles). The equation for specific distance class *D* is written:  $I(D) = [n \sum_i \sum_j w_{ij}(D) Z_i Z_j] / W_D \sum_i Z_i^2$ , where  $w_{ij}(D)$  are binary (0,1) variables specifying the inclusion (1) or exclusion (0) of the pair of individuals *i* and *j* in class *D*, and  $W_D$  is the sum of the weights or twice the number of joins or pairs placed into class *D* (Sokal & Oden 1978; Cliff & Ord 1981). Under the randomization null hypothesis these statistics are completely free of assumptions about the underlying distributions (Cliff & Ord 1981).

For each simulation, alleles were indexed by a number from 1 to  $n_a$ . For each set of simulations, the means and variances of individual *I*-statistics were calculated, separately for each distance class and allele index, across the 100 values (100 simulations). For sets with more than two alleles, the means and variances were further averaged over alleles. Thus for example, the variances reported are the variances of allele-specific individual *I*-statistics over 100 simulations, averaged over all alleles.

## Results

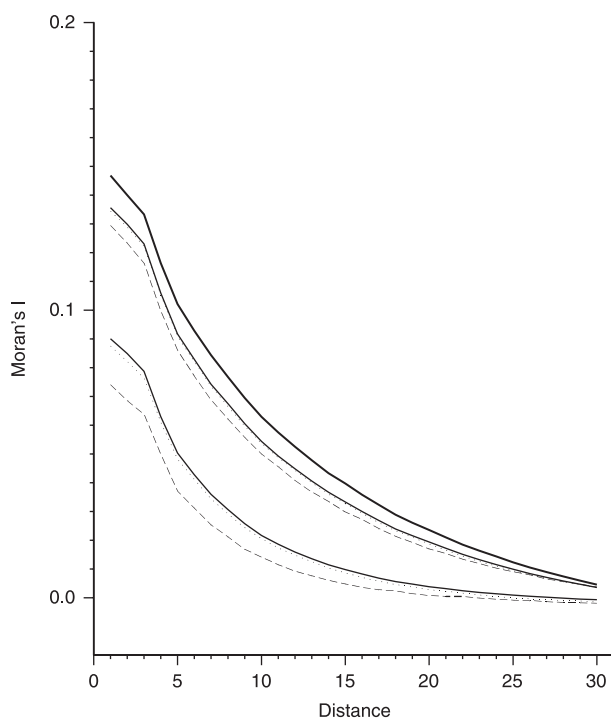
Correlograms averaged over all alleles (i.e. two, five or 10) and replications in a set (Figs 1–4, note the different scales of the *y*-axes), strongly diminished as dispersal increases, as has been shown in previous studies. The values for sets without mutation are nearly identical to those obtained in prior studies (e.g. Epperson 1995) and should be considered highly precise and robust. Mutation generally reduces spatial autocorrelation, and the effect is much stronger when the mutation rate  $\mu$  is  $10^{-2}$  than when  $10^{-3}$ . At the higher rate mutation roughly 'halves' the correlations, i.e. their differences from the expected value under the null hypothesis of a random distribution,  $-1/9999$ , effectively zero. In contrast, mutation has a minor effect when at the lower rate. The effects in terms of percents of reduction are quite consistent over levels of dispersal.

The effects of allele numbers are also clear. There is a negligible but consistent change in the 10-allele model relative to the five-allele model. Nearly point by point, i.e. for each distance class, values for the five-allele model are

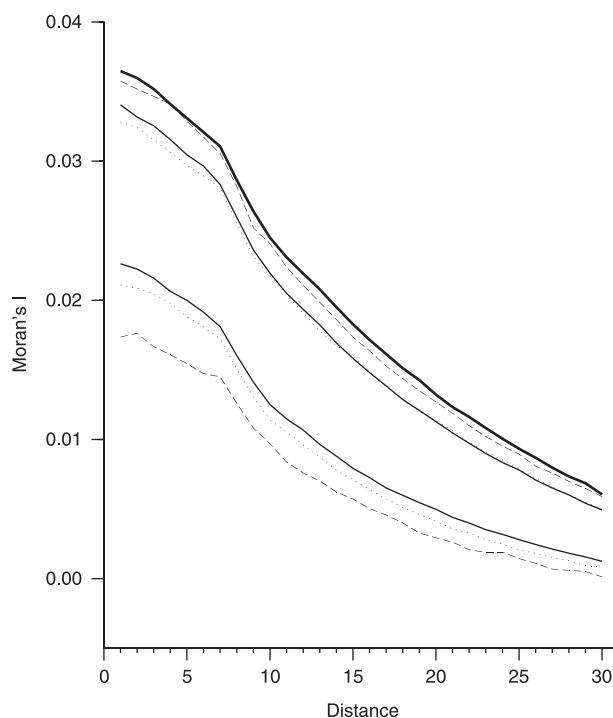


**Fig. 1** Correlograms of Moran's *I*-statistics for simulations with very low dispersal ( $N_e = 4.2$ ) for populations with various mutation rates  $\mu$  and numbers of alleles. The uppermost double thick solid line represents the model with  $\mu = 0$  and five alleles. The other three upper lines represent cases with  $\mu = 10^{-3}$  and the lower three have  $\mu = 10^{-2}$ . The dashed lines are for models with two alleles, dotted for five alleles and solid 10 alleles. Note that the curve for the five allele model with  $\mu = 10^{-3}$  is nearly invisible, because it is covered by that for the 10-allele model.

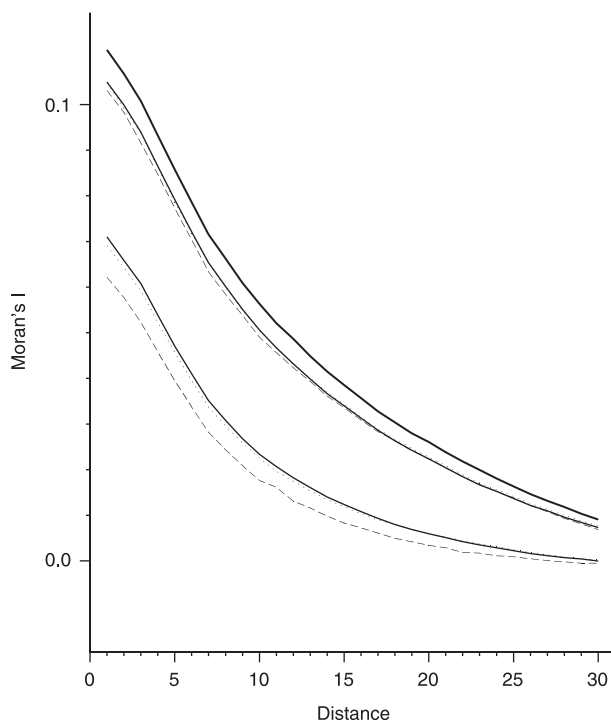
nearly identical to those for the corresponding 10-allele model, but slightly closer to zero. This also indicates robustness of the theoretical curves. However, the correspondence is generally closer when  $\mu$  is  $10^{-3}$  rather than  $10^{-2}$ , despite the smaller differentials from sets with  $\mu = 0$ . This pattern is consistent with the difference in back mutation rate between the five- and 10-allele models and its dependence on the total mutation rates. The back mutation rate for an allele *i* is the product of the frequencies of all other alleles, taken as  $(n_a - 1)/n_a$  (and ignoring the minor fluctuations of allele frequencies during the course of a simulation), times the probability of mutating  $\mu$ , times the probability that the mutation is to allele *i*, taken as  $1/(n_a - 1)$ . Thus when  $\mu = 10^{-2}$ , the back mutation rates  $\mu/n_a$  are  $0.2 \times 10^{-2}$  and  $0.1 \times 10^{-2}$  for five and 10 alleles, respectively. The difference,  $1.0 \times 10^{-3}$ , causes only a small reduction. The difference in back mutation rate,  $1.0 \times 10^{-4}$ , when  $\mu = 10^{-3}$ , has negligible effects. Thus, the fact that microsatellites often mutate in a stepwise fashion, and hence would back mutate in the same way is unimportant, as back mutations of all types do not significantly alter spatial structure of an allele in question. Moreover, the lack of spatial effects of back mutation, together



**Fig. 2** Correlograms of  $I$ -statistics for simulations with moderately low levels of dispersal ( $N_e = 25.1$ ). Legend for plots follows that in Fig. 1.



**Fig. 4** Correlograms of  $I$ -statistics for simulations with moderately high levels of dispersal ( $N_e = 115.2$ ). Legend for plots follows that in Fig. 1.



**Fig. 3** Correlograms of  $I$ -statistics for simulations with moderate levels of dispersal ( $N_e = 50.2$ ). Legend for plots follows that in Fig. 1.

with the smallness of the forward mutation effect itself at  $\mu = 10^{-3}$ , indicates that it is not worthwhile to simulate further models with  $\mu = 10^{-4}$ , because the effects would be negligible and largely inseparable from statistical noise in samples of realistic size. Just to further verify this, an additional set of 100 simulations was run with  $N_e = 4.2$ ,  $n_a = 5$ , and  $\mu = 10^{-4}$ . The correlations were nearly identical to those in the corresponding model with  $\mu = 0$ . For example, for the first 10 distance classes there were tiny reductions, ranging in numerical value from 0.002 to 0.005. For the especially important first distance class, the reduction represented a *c.* 1.0% change from models without mutation. The results also indicate that further increases in numbers of alleles beyond 10 will not substantially change the effects of mutation, at least when the forward mutation rate is  $10^{-2}$  or smaller. In other words, for all intents and purposes of this study, the five- and 10-allele models mimic the infinite alleles model.

Contrastingly, the two-allele model generally produces values that are somewhat more reduced, although the differences are still small. Again the differences are usually smaller when  $\mu = 10^{-3}$  than when  $\mu = 10^{-2}$ . There is one odd result for the model ( $N_e = 115.2$ ) with highest dispersal (Fig. 4), where the values for the two-allele model are larger than those for the five- and 10-allele models when  $\mu = 10^{-3}$ . However, the role of stochasticity in values of  $I$ -statistics appears greater in the simulations with higher

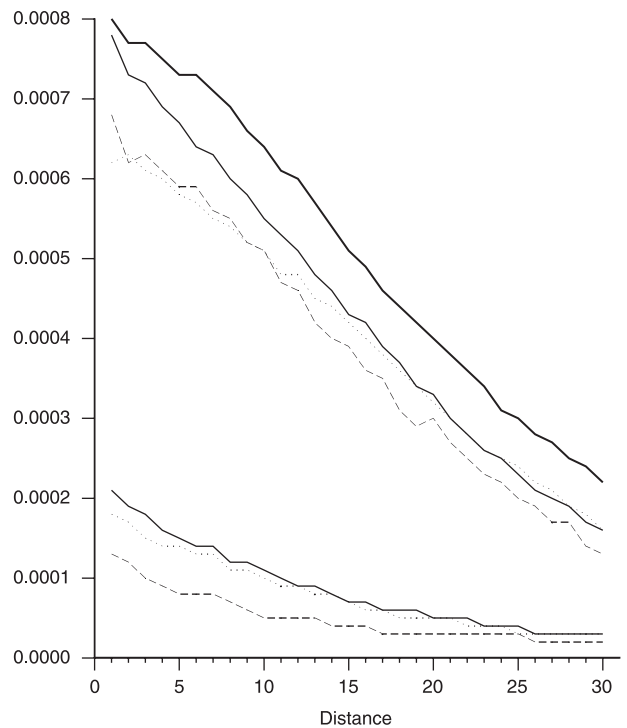
dispersal, because the values themselves are small. Also, the two-allele model does not allow averaging over alleles, i.e. the  $I$ -statistics for the two alleles of a diallelic locus are identical, and that is not the case for multiallelic loci. Thus there is greatest stochasticity in this case, although it is still fairly small.

It is of particular interest to examine the  $I$ -statistics for the first distance class (Table 1), because it contains considerable information about the amount of dispersal, and yet allows multiallele, multiple-locus summary statistical analyses that are straightforward, in contrast to correlograms (Oden 1984). Mutation at the rate of  $10^{-3}$  caused values to be as low as 88% those of the sets with no mutation (Table 1). If the value for the above-mentioned odd result for the model with two alleles and  $N_e = 115.2$  is ignored, the largest value is at 94%. On average, a 7% reduction in values of Moran's  $I$ -statistics for distance class 1 results from mutation at rate  $\mu = 10^{-3}$ . When the mutation rate is  $10^{-2}$  there appear to be lesser percent reductions of values of Moran's  $I$ -statistics for distance class 1 for the model with lowest dispersal, else little or no trends of reductions with dispersal. As for the correlograms, there are small changes from the two-allele model to the five-allele model, but almost no effect when allele number is further increased to 10. For a diallelic locus with reversible mutation (which is unlikely for a microsatellite with such high mutation rate) a reduction of *c.* 40%–50% occurs, and the results suggest that for numbers of alleles larger than five (in the present context, effectively infinite alleles) reductions of *c.* 30%–40% result.

The variances for sets with moderate dispersal ( $N_e = 50.2$ ) are illustrated in Fig. 5. Overall, the variances are quite small, supporting the high level of predictability seen in the repeatability of mean statistics over sets and the smoothness of the mean correlograms over distance classes. Mutation at the rate of  $10^{-2}$  caused an approximate four-fold reduction in variances for short distance classes. Mutation at the rate of  $10^{-3}$  caused much smaller percent reductions, although greater than the percent reductions in the mean. Very similar results in terms of percent reductions caused by mutation were found for the other dispersal models, and thus these are not shown.

## Discussion

The results show that when mutation has a rate  $\mu$  on the order of those of some hypervariable genetic markers, it reduces spatial autocorrelations by roughly half. Moran's  $I$ -statistics for the shortest distances of separation among individuals were reduced by *c.* 30%–50% when  $\mu = 10^{-2}$ . Similarly, for all distance classes studied, the deviations of values from the expected values under the null hypothesis, in this case essentially zero, were greatly reduced. In contrast, when mutation occurs at a rate of  $10^{-3}$ , only mild reductions, averaging 7% for the first distance class, were



**Fig. 5** Variances of allele-specific individual  $I$ -statistics over 100 replicate simulations, averaged over all alleles, separately for each distance class and set, for the sets with moderate dispersal ( $N_e = 50.2$ ). Legend for plots follows that in Fig. 1.

observed. These observations, as well as other evidence available with respect to rates and lack of effects of back mutations, indicate that when mutation occurs at lower orders of magnitude (i.e.  $10^{-4}$  or lower), effects are negligible. Hypervariable markers, including microsatellites and minisatellites, can have mutation rates on the order of  $10^{-2}$ , but may also range down to  $10^{-3}$  and  $10^{-4}$  (e.g. Bruford *et al.* 1992; Jarne & Lagoda 1996). Hence the range of studied rates distinguishes the critical range that may separate the most highly mutable hypervariable loci ( $\mu$  on the order of  $10^{-2}$ ) from less mutable ones ( $\mu = 10^{-3}$ ), as well as from genes that mutate at normal rates (typically  $10^{-5}$ – $10^{-6}$ ) such as isozymes. Choice of markers could affect the spatial structures observed.

The mutation models studied were all  $k$ -allele models, with fixed numbers of alleles and reversible mutation. Forward mutation in  $k$ -allele models behaves as in the infinite alleles model in terms of reducing identity by descent. As was noted in the Introduction, spatial structure of genetic variation is created primarily by recent coalescences among spatially proximal genes (Barton & Wilson 1995; Barton *et al.* 2002). The results show that forward mutations at rates on the order of  $10^{-2}$  significantly reduces spatial autocorrelation, and hence identity by descent among such genes. Naturally, mutations also occur among other pairs (i.e. not spatially proximal) of previously identical genes that do

not have recent coalescences. However, these do not substantially affect spatial structure, because such pairs contain almost no information on spatial distribution (Barton & Wilson 1995), i.e. the probability distribution of the distance between such pairs of genes is essentially uniform.

The results also showed that back mutation has little or no effect on spatial structure. Only the two-allele model showed any effect, and even that was small and negligible in a typical experimental setting. The same will be true for any other mutation model. For example, any potential importance of back mutation on identity in state, as opposed to identity by descent, is greatest for microsatellites, which to some degree mutate in a stepwise fashion. For the strict stepwise mutation model, the back mutations to any allele size  $i$  occur at rate  $\mu[p(i-1) + p(i+1)]/2$ , where  $p(i-1)$  and  $p(i+1)$  are the frequencies of allele size  $i-1$  and  $i+1$ , respectively. The largest possible value is thus  $\mu/2$ , although in most real cases it would be considerably smaller. The back mutation rate in the two-allele model simulated had precisely the same value, and there was negligible effect.

In essence, back mutations create identity in state in a pair of genes when one of the genes previously had a different allelic state than the other. The gene that mutated could be any spatial distance from the other gene, and as noted, this means there is essentially no information on spatial structure. There is little or no effect on spatial correlations. However, because spatial autocorrelation statistics are normalized by the variance of allele frequencies (i.e. genotypes converted to allele frequencies) in the population or sample (Cliff & Ord 1981), in theory the spatial autocorrelation statistics for shorter distances could be affected. Nonetheless, as noted even back mutation at the highest possible (and highly improbable) rate, has negligible effect (slight decrease) on spatial autocorrelations, even when the forward mutation rate is very high. Because back mutation at the highest possible rate had no appreciable effect, the nature of back mutations, which is what distinguishes various mutation models (including the stepwise mutation model), cannot have any effect. Thus, the reductions observed are valid for all loci, including microsatellites.

Importantly, the percent reductions in values of spatial correlations, caused by mutation, are independent of the amount of spatial structure. It is unnecessary to know a priori the amount of structure in order to predict or determine differences caused by variations in mutation rates. The only exception occurred when dispersal was nearly maximally limited and slightly greater reductions were observed. Thus, further increases in dispersal beyond those included in the study should not change the relationship between mutation rate and percent reduction.

The results on Moran's  $I$ -statistics can be extended to the conditional kinship measure  $R$  for pairs of individuals in the same distance class, because  $R = I(1 + F)/2$ , where  $F$  is the fixation index (see also Barbujani 1987; Hardy & Vekemans

1999). The fixation index is essentially a function of the correlation at 'zero' spatial distance, and thus responds similarly to mutation as do the short distance autocorrelations. Hence, the reductions for  $R$  can be derived from those for  $I$  and  $F$ . Mutation at high rates should reduce multiallele multilocus measures of the genetic covariance,  $R$ , as well as other closely related measures, such as the measure of Smouse & Peakall (1999). The measure of Loiselle *et al.* (1995), for example, is an average of  $I$ -statistics weighted by allele-specific values of  $(1 + F)/2$  (Epperson 2004).

Indirect estimates of dispersal based on  $I$  or  $R$  would be strongly affected by very high mutation rates. For example, using tables for the  $I$ -statistics for shortest distance class in Epperson *et al.* (1999), a 50% reduction corresponds to actual vs. estimated Wright's (1943) neighbourhood sizes of roughly: 8, 25; 25, 75; 100, 230, etc. for  $\mu = 10^{-2}$ . Hence a two- to three-fold inflation of neighbourhood size estimates should typically result from mutation at the rate of  $10^{-2}$ .

As noted in the Introduction, it appears that mutation may have affected estimates of spatial autocorrelation in populations of *Pinus strobus* in Michigan (Walter & Epperson 2004). In one seedling population, the average Moran's  $I$ -statistic for the first distance class for the most variable locus (Rps50 with 12 alleles in a sample of *c.* 100 seedlings, effective number of alleles 5.7) was  $-0.01$ , essentially equal to the expected value for the null hypothesis of a random distribution. The average  $I$ -statistic among alleles at all other microsatellite loci (2–5 alleles, average 3.8, average effective number 1.7) was 0.05, and the difference was statistically significant. Using the well-known relationship of effective number of alleles to the product of population size and  $\mu$ , in the genetic drift–infinite alleles mutation model (Ewens 1979), it appears that Rps50 mutates at a rate about seven times larger than the other microsatellites. Although direct estimates of absolute values of mutation rates are not available for these genes, the contrasts among relative rates (roughly one order of magnitude) are consistent with mutational effects. Similarly, the average observed for allozymes was 0.04. In addition, similar contrasts among loci were observed in another seedling population (Walter & Epperson 2004) and in adult populations in Michigan (Marquardt & Epperson 2004).

Effects of high rates of forward mutation on other spatial or nonspatial measures should depend on how much those measures are affected by recent coalescences. For example, in migration–drift models of systems of populations, genetic drift within populations tends to increase differentiation among populations and migration tends to make them homogeneous. Genes in different populations are unlikely to have a recent coalescence, unless they or their ancestors were involved in a recent migration (e.g. Nordborg 1997). Thus, the extent to which spatial correlations among populations or measures of  $F_{ST}$  are affected by high mutation rates is likely associated with their dependence on recent

migrations. Recently, Balloux & Lugon-Moulin (2002) discussed how high rates of mutation decrease estimates of population differentiation.

Hypervariable loci offer several important advantages in application to population genetics in general, including studies of structure. High levels of polymorphism per locus can often offset the added effort and costs in comparison to isozymes. Moreover, investigators often choose the more polymorphic microsatellites available, rather than run more microsatellites with fewer alleles, because of savings in costs and effort. While it is unusual for investigators to know the mutation rates of the markers employed, the results in this study indicate that some caution is in order, and that it is not always best to choose the most diverse markers available. In some cases, especially where mutation rates on the order of  $10^{-2}$  may be suspected, it would be better to increase the number of loci assayed.

## References

- Balloux F, Lugon-Moulin N (2002) The estimation of population differentiation with microsatellite markers. *Molecular Ecology*, **11**, 155–165.
- Barbujani G (1987) Autocorrelation of gene frequencies under isolation by distance. *Genetics*, **117**, 777–782.
- Barton NH, Wilson I (1995) Genealogies and geography. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **349**, 49–59.
- Barton NH, Depaulis F, Etheridge AM (2002) Neutral evolution in spatially continuous populations. *Theoretical Population Biology*, **61**, 31–48.
- Bruford MW, Hanotte O, Brookfield JFY, Burke T (1992) Multi- and single-locus fingerprinting. In: *Molecular Analysis of Populations: A Practical Approach* (ed. Hoelzel AR), pp. 225–269. IRL Press, Oxford, UK.
- Cliff AD, Ord JK (1981) *Spatial Processes*. Pion, London.
- Cockerham CC (1969) Variance of gene frequencies. *Evolution*, **23**, 72–84.
- Epperson BK (1990) Spatial autocorrelation of genotypes under directional selection. *Genetics*, **124**, 757–771.
- Epperson BK (1995) Fine-scale spatial structure: correlations for individual genotypes differ from those for local gene frequencies. *Evolution*, **49**, 1022–1026.
- Epperson BK (2003) *Geographical Genetics*. Princeton University Press, Princeton, New Jersey.
- Epperson BK (2004) Multilocus estimation of genetic structure within populations. *Theoretical Population Biology*, **65**, 227–237.
- Epperson BK, Chung MG (2001) Spatial genetic structure of allozyme polymorphisms within populations of *Pinus strobus* (Pinaceae). *American Journal of Botany*, **88**, 1006–1010.
- Epperson BK, Li T-Q (1996) Measurement of genetic structure within populations using Moran's spatial autocorrelation statistics. *Proceedings of the National Academy of Sciences of the United States of America*, **93**, 10528–10532.
- Epperson BK, Li T-Q (1997) Gene dispersal and spatial genetic structure. *Evolution*, **51**, 672–681.
- Epperson BK, Huang Z, Li T-Q (1999) Spatial genetic structure of multiallelic loci. *Genetical Research Cambridge*, **73**, 251–261.
- Escudero A, Iriondo JM, Torres ME (2003) Spatial analysis of genetic diversity as a tool for plant conservation. *Biological Conservation*, **113**, 351–365.
- Ewens WJ (1979) *Mathematical Population Genetics*. Springer, New York, NY.
- Hardy OJ, Vekemans X (1999) Isolation by distance in a continuous population: reconciliation between spatial autocorrelation analysis and population genetics models. *Heredity*, **83**, 145–154.
- Heywood JS (1991) Spatial analysis of genetic variation in plant populations. *Annual Review of Ecology and Systematics*, **22**, 335–355.
- Jarne P, Lagoda PJJ (1996) Microsatellites, from molecules to populations and back. *Trends in Ecology and Evolution*, **11**, 424–429.
- Loiselle BA, Sork VL, Nason J, Graham C (1995) Spatial genetic structure of a tropical understory shrub, *Psychotria officinalis* (Rubiaceae). *American Journal of Botany*, **82**, 1420–1425.
- Luna R, Epperson BK, Oyama K (2005) Spatial genetic structure of two sympatric neotropical palms with contrasting life histories. *Heredity*, in press.
- Marquardt PE, Epperson BK (2004) Spatial and population genetic structure of microsatellites in white pine. *Molecular Ecology*, **13**, 3305–3315.
- Morand M-E, Brachet S, Rossignol P, Dufour J, Frascaria-Lacoste N (2002) A generalized heterozygote deficiency assessed with microsatellites in French common ash populations. *Molecular Ecology*, **11**, 377–385.
- Nordborg M (1997) Structured coalescent processes on different time scales. *Genetics*, **146**, 1501–1514.
- Oden NL (1984) Assessing the significance of a spatial correlogram. *Geographical Analysis*, **16**, 1–16.
- Rohlf FJ, Schnell GD (1971) An investigation of the isolation-by-distance model. *American Naturalist*, **105**, 295–324.
- Smouse PE, Peakall R (1999) Spatial autocorrelation analysis of individual multiallele and multilocus genetic structure. *Heredity*, **82**, 561–573.
- Sokal RR, Oden NL (1978) Spatial autocorrelation in biology. 1. Methodology. *Biological Journal of the Linnean Society*, **10**, 199–228.
- Sokal RR, Wartenberg DE (1983) A test of spatial autocorrelation analysis using an isolation-by-distance model. *Genetics*, **105**, 219–237.
- Streiff R, Labbe T, Bacilieri R, Steinkellner H, Glössl J, Kremer A (1998) Within-population genetic structure in *Quercus robur* L. & *Quercus petraea* (Matt.) Liebl. assessed with isozymes and microsatellites. *Molecular Ecology*, **7**, 317–328.
- Udupa SM, Baum M (2001) High mutation rate and mutational bias at (TAA)<sub>n</sub> microsatellite loci in chickpea (*Cicer arietinum* L.). *Molecular Genetics and Genomics*, **265**, 1097–1103.
- Walter R, Epperson BK (2004) Microsatellite analysis of spatial structure among seedlings in populations of *Pinus strobus* (Pinaceae). *American Journal of Botany*, **91**, 549–557.
- Wright S (1943) Isolation by distance. *Genetics*, **28**, 114–138.

---

Bryan Epperson, Professor at Michigan State University, studies theoretical and statistical aspects of geographical genetics, and uses molecular markers to study the spatial population genetics of trees and other species.

---