Influence of Mutational and Sampling Factors on the Estimation of Demographic Parameters in a "Continuous" Population Under Isolation by Distance

Raphaël Leblois, Arnaud Estoup, Francois Rousset

Influence of Mutational and Sampling Factors on the Estimation of Demographic Parameters in a "Continuous" Population Under Isolation by Distance

Raphaël Leblois,† Arnaud Estoup,* and François Rousset†*

*Laboratoire Modélisation et Biologie Evolutive, CBGP-INRA, Montferrier sur Lez, France; and †Laboratoire Génétique et Environnement, CNRS-UMR 5554, Montpellier, France

In numerous species, individual dispersal is restricted in space so that "continuous" populations evolve under isolation by distance. A method based on individual genotypes assuming a lattice population model was recently developed to estimate the product $D\sigma^2$, where D is the population density and σ^2 is the average squared parent-offspring distance. We evaluated the influence on this method of both mutation rate and mutation model, with a particular reference to microsatellite markers, as well as that of the spatial scale of sampling. Moreover, we developed and tested a nonparametric bootstrap procedure allowing the construction of confidence intervals for the estimation of $D\sigma^2$. These two objectives prompted us to develop a computer simulation algorithm based on the coalescent theory giving individual genotypes for a continuous population under isolation by distance. Our results show that the characteristics of mutational processes at microsatellite loci, namely the allele size homoplasy generated by stepwise mutations, constraints on allele size, and change of slippage rate with repeat number, have little influence on the estimation of $D\sigma^2$. In contrast, a high genetic diversity ($\approx$0.7–0.8), as is commonly observed for microsatellite markers, substantially increases the precision of the estimation. However, very high levels of genetic diversity ($>$0.85) were found to bias the estimation. We also show that statistics taking into account allele size differences give unreliable estimations (i.e., high variance of $D\sigma^2$ estimation) even under a strict stepwise mutation model. Finally, although we show that this method is reasonably robust with respect to the sampling scale, sampling individuals at a local geographical scale gives more precise estimations of $D\sigma^2$.

Introduction

Dispersal rates and population sizes or densities are important demographic parameters in evolutionary processes. Many studies have attempted to estimate such parameters using either direct methods (e.g., mark-recapture methods) or indirect methods (e.g., genetic markers). A number of indirect methods for demographic parameter estimation using genetic data at neutral loci or clines of selected markers have been defined (see Slatkin (1994) and Rousset (2001*b*) for reviews). Discrepancies between estimations made with direct and indirect methods have often been attributed to inadequacies of the assumptions of the genetic models made in indirect methods (Hastings and Harrison 1994; Koenig et al. 1996; Slatkin 1994). The kinds of assumptions usually considered to be inadequate are those related to (1) the modalities of dispersal (e.g., the island model), (2) the demographic stability in space and time, (3) the mutation rates and mutation processes of genetic markers, and (4) the selective neutrality of genetic markers.

In numerous species, individual dispersal is restricted in space. This means that there is a higher probability that individuals mate with individuals born in close proximity to themselves than to individuals born far away. Several studies on animals or plants have shown such restricted dispersal (e.g., for plant data, see Crawford 1984; and for animal data, Rousset 1997, 2000; Spong and Creel 2001; Sumner et al. 2001). Isolation by distance models taking into account this biological feature were introduced by Wright (1943 and 1946). Under these models the genetic differentiation at neutral loci is expected to increase with geographical distance (e.g., Malécot 1950, 1967; Sawyer 1977). Empirical data indicate that such a relationship holds for many species (Endler 1977; Slatkin 1993). Recently, a method of analysis was developed based on the increase, at a local scale, of genetic differentiation between individuals with geographical distance in a "continuous" population evolving under isolation by distance (Rousset 2000). The method makes use of the regression of estimators of a parameter analogous to the parameter $F_{ST}/(1 - F_{ST})$, calculated between individuals, and the logarithm of the geographical distance, to estimate the product $D\sigma^2$, where D is the density of adults and σ^2 the average squared axial parent-offspring distance. It is expected to perform better than previous methods for several reasons. First, the demographic model on which the method is based makes weak assumptions about the shape of the distribution of dispersal distances. In particular, the method is valid for leptokurtic distributions of dispersal distance (Rousset 2000), a feature commonly observed in natural populations (for review and data, see Endler 1977; Portnoy and Willson 1993; Clark et al. 1999). Second, analysis of genetic differentiation is made at a small (local) geographical scale so that heterogeneity of demographic parameters such as dispersal or density is reduced and hence its influence on genetic differentiation is also reduced (Slatkin 1993; Rousset 2001*b*). In a similar way, influence of non-neutrality of the genetic markers may be less problematic for studies at local scale because selection parameters may be less heterogeneous at a small geographical scale. On the other hand, the theory on which the method is based shows that only estimations from analysis over short distances will be accurate (Rousset 1997). These expectations have been confirmed by several comparisons of direct and indirect estimates of $D\sigma^2$ (Rousset 1997, 2000; Sumner et al. 2001). Although the geographical scale at which the sampling has been done is

Key words: coalescence, dispersal, isolation by distance, microsatellite DNA, nonparametric ABC bootstrap.

E-mail: leblois@isem.univ-montp2.fr.

Mol. Biol. Evol. 20(4):491–502. 2003
DOI: 10.1093/molbev/msg034

expected to influence the quality of the estimation of $D\sigma^2$, very few analytical or simulation studies have formally addressed this question.

Since their discovery in the 1980s, microsatellite loci have been increasingly used as genetic markers. Rapid progress in molecular biology technologies, especially the development of the polymerase chain reaction, and attractive evolutionary features (e.g., high level of polymorphism), explain why this category of markers are progressively replacing, or at least complementing, classical markers such as allozymes for numerous applications in molecular systematics, population genetics, and ecology (reviewed in Estoup and Angers 1998; Estoup, Jarne, and Cornuet. 2002). However, the mutation processes (i.e., the nature of mutations) at microsatellite loci are complex and not yet well understood (e.g., Estoup and Cornuet 1999). The effect of the mutation processes on evolutionary inferences depends in large part on the method, the statistics, and the evolutionary time scale considered (e.g., Estoup, Jarne, and Cornuet 2002). Some authors have discussed the effect of the nature of the mutation on F_{ST} values (Slatkin 1995; Rousset 1996). Because a stepwise mutation process occurs at microsatellite loci, several statistics taking into account the allele size have been proposed (Goldstein et al. 1995; Slatkin 1995; Michalakis and Excoffier 1996). Their utility, however, has often been criticized (e.g., Takezaki and Nei 1996; Gaggiotti et al. 1999). Overall, the potential interest of the different statistics has never been addressed in the context of the estimation of demographic parameters under isolation by distance.

In this study, we developed an original simulation algorithm based on the coalescent theory in order to study the sensitivity of the estimation of $D\sigma^2$ to different factors: (1) the sampling scale of individuals, (2) the mutation model of markers and (3) their mutation rate, with particular reference to microsatellite markers for the two latest points. This algorithm was also used to test a nonparametric ABC bootstrap procedure allowing the construction of confidence intervals on the $D\sigma^2$ estimation. Finally, we draw guidelines that could be useful for empirical investigators using the individual-based method of Rousset (2000).

Models and Methods

Demographic Model and Population Cycle

The model that we considered for "continuous" populations is the lattice model with each lattice node corresponding to one diploid individual. This model without demic structure is viewed as an approximation for truly continuous populations with infinite local competition (Malécot 1975; Rousset 2000). More realistic continuous models would incorporate the feature that individuals could settle in any position in a continuous space. Although such models have been formulated (e.g., Malécot 1967; Sawyer 1977), it is known that they do not follow a well-defined set of biological assumptions (Maruyama 1972; Felsenstein 1975; see Barton et al. 2002 for an alternative approach for continuous populations). Individuals are assumed to be diploids by a model with two independent genes per node. To avoid edge effects, the lattice is represented on a circle for a one-dimensional model or a torus for a two-dimensional model. Edge effects have little influence on local differentiation when the habitat area (i.e., the lattice size) is large when compared to the mean dispersal. Finally, we considered that dispersal occurs through gametes only.

The life cycle is divided into four steps: (1) at each reproductive event, each individual gives birth to a great number of gametes, and then dies; (2) gametes undergo the effect of mutations; (3) gametes disperse; (4) diploid individuals are formed, and (5) competition brings back the number of adults in each deme to one.

Coalescent Algorithm

The genealogical tree of a sample of n genes taken from a panmictic population of constant size N can be modeled using a stochastic process known as the n-coalescent. This process was introduced by Kingman (1982*a*, 1982*b*) as an approximation of a gene genealogy under the "Wright-Fisher" neutral model (see also Hudson 1990, Tajima 1983). More sophisticated models have since been developed for analysis of more complex evolutionary scenarios with recombination, selfing, and variable population size (reviewed in Nordborg 2001).

The n-coalescent approximation can be used in the same context as diffusion equations (Nordborg 2001). It is thus valid for a restricted numbers of models of population structure, e.g., panmictic populations or the infinite island model. In the present work, we focused on isolation by distance. For this category of models, no analytical treatment of coalescence time or coalescence probabilities has been done for more than two genes. Algorithms such as those developed for likelihood estimation by Griffiths and collaborators (see Nath and Griffiths 1996; Bahlo and Griffiths 2000) could in principle deal with continuous models; however, they are not ready for demographic inferences (De Iorio and Griffiths, personal communication). The coalescent algorithm we developed is not based on the n-coalescent theory; rather it is an algorithm for which coalescence and migration events are considered "generation by generation" until the common ancestor of the sample has been found. The idea of tracing lineages back in time generation by generation is fundamental in the coalescence theory, and is well described in Nordborg (2001). At least one study already used this simple concept for simulations (i.e., Pope, Estoup, and Morris 2000). Although such a generation-by-generation algorithm leads to less efficient simulations in terms of computation time than those based on the n-coalescent theory, it is much more flexible when complex demographic and dispersal features are considered. The algorithm described below and the program used in this study were checked at every step during elaboration by comparison with exact analytical results for probabilities of identity in models of isolation by distance on finite lattice (e.g., Malécot 1975 for the lattice model, adapted to different mutation models following Rousset 1996). These comparisons show that estimates of identity probabilities from our program and

analytical expectations differ by less than one per thousand for sufficiently long runs.

Let us consider, at a given time and on a two-dimensional lattice, a sample of $n(0)$ genes numbered 1 to $n(0)$. The position of each gene on this lattice is given by a pair of coordinates (x,y). The set of coordinates of sampled genes is given by the two vectors $X(0) = [x_1(0), \ldots, x_{n(0)}(0)]$, $Y(0) = [y_1(0), \ldots, y_{n(0)}(0)]$, where $x_i(0)$ and $y_i(0)$ are the coordinates of the gene i at $G = 0$, with G corresponding to the number of generations since sampling.

This algorithm goes backward in time, generation by generation (considering discrete generations). At $G = 1$, parents of our $n(0)$ sampled genes have coordinates $x_i(1) = x_i(0) + dx$, $y_i(1) = y_i(0) + dy$, where dx and dy are random variables representing dispersal distance in one dimension, expressed in number of steps on the lattice. Under a two-dimensional model, the density function of the random variable (dx,dy) is given by $b_{dx,dy}$, the "backward" dispersal function. The term *backward* is used because the position of the parental gene is determined knowing the position of its descendant gene. This function is calculated using $f_{dx,dy}$, the forward dispersal density function describing where descendants go. The dispersal functions are detailed in the next section. We assume that dispersal is independent in each direction, so that $f_{dx,dy} = f_{dx} \times f_{dy}$. Considering that density is homogenous in space, backward dispersal functions are equal to forward dispersal functions, so that $b_{dx,dy} = f_{dx,dy} = f_{dx} \times f_{dy}$.

Once the position of the parents on the lattice is known, the coalescence events occurring at $G = 1$ are assessed. In other words, we determine whether some genes share a common parent at $G = 1$. This step corresponds to the idea of "individuals picking their parents at random from the previous generation" (Nordborg 2001). A coalescence event occurs if genes are both on the same lattice node and if they originate from the same parental gene. Multiple coalescences are allowed. The probability for a coalescence of k genes in a given parental gene is $1/2^{k-1}$ under the model with one individual per lattice node. In this case, the remaining j genes from the same lattice node coalesce in the other parental gene. For convenience, we keep the numbering ($i \in [1, \ldots, n(0)]$) of descendant genes for their parents when these genes do not coalesce and attribute new numbers ($i \in [n(0) + 1, \ldots, n(1)]$) for the parents of the coalesced genes. A gene i at $G = 0$ and its parent at $G = 1$ have the same number if there was no coalescence event between the gene i and another gene at $G = 0$. Thus our numbering refers more to the branches of the coalescent tree than to the genes themselves. This particular numbering of branches, nodes, and genes is illustrated in figure 1. At G = 1, we have $X(1) = (x_1(1), ., x_{n(1)}(1))$, $Y(1) = (y_1(1), ., y_{n(1)}(1))$, the $n(1)$ geographic coordinates at $G = 1$ for each branch corresponding to a lineage of our sample. We keep in memory the ages of the tree "nodes" (corresponding to coalescence events) and the labels of the branches descending from this "node." The entire process is repeated over generations until the most recent common ancestor of our entire gene sample has been found.

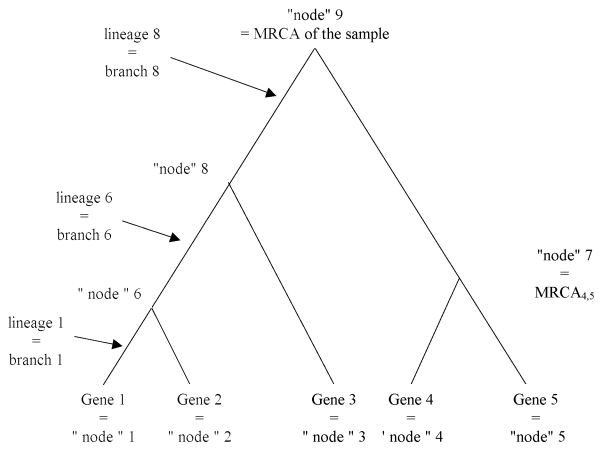

FIG. 1.—Numbering of branches, genes, and nodes of a genealogical tree for a sample of five genes as described by our coalescence algorithm.

Dispersal Functions

Biologically realistic dispersal functions often have a high kurtosis (Endler 1977; Kot, Lewis, and van den Driessche 1996). Forward dispersal distributions for which the probability of moving k steps (for $0 < k \leq K_{max}$) in one direction is of the form $f_k = f_{-k} = M/k^n$ were considered, with parameters M and n controlling the total dispersal rate and the kurtosis, respectively.

By suitable choice of the two parameter values, large kurtosis can be obtained with high migration rates (Rousset 2000). For all of our simulations, we used a dispersal distribution with a moderate σ^2 value ($\sigma^2 = 4$), corresponding to a dispersal distribution with parameters:

$$f_1 = f_{-1} = 0.06, \quad f_2 = f_{-2} = 0.03 \quad \text{and for}$$
$$2 < k < 49, \quad M = 0.802 \quad \text{and} \quad n = 2.518. \quad (1)$$

With such a dispersal distribution the product $4\pi D\sigma^2$ is 50.26. This value corresponds to a relatively strong isolation by distance, which appears biologically reasonable for many species (see references cited in the *Introduction*).

Mutation Processes

One interesting feature of the coalescent-based approach is that, for neutral loci, genealogical and mutation processes are totally independent, so that the effects of mutation are simply superimposed on the genealogical tree obtained for the gene sample.

Two theoretical mutation models, the infinite allele model (IAM: Kimura and Crow 1964) and the K-allele model (KAM: Crow and Kimura 1970), have sometimes been used for microsatellite loci. However, the most widely adopted model for microsatellite mutation is the stepwise mutation model (SMM: Ohta and Kimura 1973) in which the mutant allele differs from its parent by one repeat. Direct and indirect studies have shown that mutations of several repeats also occurred, indicating that a strict one-step model is inappropriate (Estoup and Angers 1998; Gonser et al. 2000; Ellegren 2000). In

practice, modeling assumptions are commonly limited to the SMM (e.g., Reich and Goldstein 1998; Wilson and Balding 1998), and sensitivity of the final inferences to this assumption may be substantial, although this is rarely investigated. In several studies (e.g., Pritchard et al. 1999), a generalization of the SMM was adopted in which the change in the number of repeat units forms a geometric random variable. This generalization was named the GSM (generalized stepwise mutation) model. The geometric distribution in our GSM model refers to a change expressed in an (absolute) number of repeat units subsequently added or withdrawn to the mutating allele with equal probability. Under this model, the large data set of microsatellite mutations of Dib et al. (1996) in humans suggests an estimate of the variance of the geometric distribution near 0.36 (Estoup et al. 2001). The GSM does not capture all the complexity of the mutation process at microsatellite loci. In particular, constraints on allele size occur at some microsatellite loci (reviewed in Amos 1999; Estoup and Cornuet 1999; Ellegren 2000) and potentially affect various statistics in population genetics (Estoup et al. 2002). This evolutionary feature, particular to microsatellite loci, was thus tested on our method. Allele size constraints were included in our simulations by imposing reflecting boundaries to the allele size range (e.g., Feldman et al. 1997; Estoup et al. 1999). Another outstanding feature of the microsatellite mutation process is that within-loci mutation rate increases with allele length (Ellegren 2000; Huang et al. 2002). Whether this increase is linear with the number of repeats remains subject to further investigation (Schlötterer 2000; Stumpf and Goldstein 2001; Brohede et al. 2002). In our simulations, we considered a linear model in which (1) the mutation rate was fixed to 5×10^{-4} for the allelic state of the root of the tree (fixed at 100 repeats units and considered the "middle size allele"); (2) a decrease in mutation rate with allele size of 0.1% or 1% per repeat unit for a weak or a strong variation, respectively is simulated for alleles shorter than 100 repeat units; (3) a similar increase is simulated for alleles longer than 100 repeat. In other words, this leads to the linear form: $\mu(L) = \mu_0 + s*L$, where $\mu(L)$ is the mutation rate for an allele of size L, μ_0 the mutation rate for the smallest allele, and s the increase per repeats unit. We set $s = 0.1\%$ or 1% for a weak or a strong variation, respectively, to be close to the value given in Brohede et al. (2002).

Interlocus variability in the mutation rate potentially decreases the precision of parameter estimation in population genetics (Takezaki and Nei 1996; Gonser et al. 2000). The effect of variable mutation rate was thus tested as well. Little information is available on the interlocus variance of the mutation rate at microsatellite loci. Several pedigree studies show that the mutation rates can differ across loci in important respects (reviewed in Schlötterer 2000). Without more information, we modeled variable mutation rates at microsatellite loci by drawing single locus mutation rate values in a gamma distribution with parameters (shape, scale) being (2, 2.5 10^{-4}). This distribution has a mean equal to 5×10^{-4}, a value considered as the average mutation rate in many species (reviewed in Estoup and Angers 1998), and 2.5% and 97.5% quantiles equal to 6×10^{-5} and 1.4×10^{-3}, respectively. These values are similar to the mean and 95% confidence interval values typically considered for autosomal microsatellites in humans (Weber and Wong 1993).

The following step-by-step procedure was used to add mutations to the genealogical tree. Take at random two genes i, j and their most recent common ancestor, the gene l, and let $state_i$, $state_j$, $state_l$ be their respective allelic states. The number of mutations that occurred in lineage i is proportional to the length L_i (expressed in number of generations) of branch i (from l to i) and is given by a binomial distribution with parameters (μ, L_i), which can be approximated by a Poisson process with parameter μL_i. Let m_i be the number of mutations that occurred on branch i. One can easily deduce $state_i$ from $state_l$ through m_i successive steps, each step corresponding to a mutation event under the chosen mutation model. The allelic states of the various genes of the sample were obtained starting from a given state for the common ancestor of the sample (root of the genealogical tree) and going forward in time on each branch.

Method of Analysis

Each simulation iteration gave the genotypes at l polymorphic loci for ($n \times n$) individuals denoted by their coordinates on the lattice. l independent coalescent trees were used to simulate multi-locus genotypes. This process was repeated 1,000 times giving 1,000 multilocus samples sharing the same demographic conditions. We computed estimates of the parameter

$$a_r \equiv \frac{Q_w - Q_r}{1 - Q_w}$$

for each pair of individuals, where Q_w is the probability of identity in state for two genes taken from the same individual, and Q_r the probability of identity in state for two genes at geographical distance r (Rousset 2000). The statistic a_r is a parameter analogous to the parameter $F_{ST}/(1 - F_{ST})$, calculated between individuals (and not between populations, as in Rousset 1997). An estimator of a_r for a pair π of individuals taken from the P different possible pairs is:

$$\hat{a} \equiv \frac{SS_{b(\pi)}P}{\sum_{k=1}^{P} SS_{w(k)}} - \frac{1}{2}$$

with

$$SS_{b[etween](\pi)} \equiv \sum_{i,u} (X_{i.:u} - X_{..:u})^2$$

and

$$SS_{w[ithin](\pi)} \equiv \sum_{i,j,u} (X_{ij:u} - X_{i.:u})^2,$$

where $X_{ij:u}$ is an indicator variable taking the value 1 if gene i of individual j is of allelic type u and the value 0 otherwise (Rousset 2000).

To test the effect of using a statistic that takes into account the allele length differences (and hence the stepwise mutational process occurring at microsatellite

loci), we defined another parameter b_r, equivalent to a_r, except that it is defined in terms of squared differences in microsatellite allele lengths (SD) instead of probabilities of non- identity in state ($1 - Q$). Thus, we have

$$b_r \equiv \frac{SD_r - SD_w}{SD_w},$$

where SD_r is the expectation of the squared length differences between two genes at geographical distance r and SD_w is the expectation of the squared length differences between two genes taken in the same individual. b_r was estimated for a pair π of individuals taken from the P different possible pairs in a way similar to a_r:

$$\hat{b} \equiv \frac{SSD_{b(\pi)}P}{\sum_{k=1}^{P} SSD_{w(k)}} - \frac{1}{2}$$

with

$$SSD_{b[etween](\pi)} \equiv \sum_i (S_{i.} - S_{..})^2$$

and

$$SSD_{w[ithin](\pi)} \equiv \sum_{i,j} (S_{ij} - S_{i.})^2,$$

where S_{ij} is a variable representing the size of gene i of individual j, expressed in number of repeat units.

For each of the 1,000 repetitions, the value of the slope of the regression line between $\hat{a}$ (or $\hat{b}$) and the logarithm of geographical distance was computed. In the limit of low mutation rates, the inverse of the slope is an estimate of the product $4\pi D\sigma^2$, where D is the density of adults and σ^2 the average squared axial parent-offspring distance (Rousset 1997). It is worth noting that high mutation rates should not result in an asymptotic bias as long as the focus is on local processes involving distances between sampled individuals

$$r \ll \frac{\sigma}{\sqrt{2\mu}}.$$

Beyond this limit, the linear relationship between a_r (or b_r) and the logarithm of the distance holds less well (for details, see Rousset 1997). Thus, if the analysis is done at a small geographical scale, the use of highly variable loci such as microsatellite loci should not bias the estimation. However, the effect of mutation on small sample properties of the estimator needs to be tested. The quality of an estimator is usually assessed through the computation of its bias and its mean square error (MSE). These measures are suitable when estimates have approximately a normal distribution but not when the estimate is sometimes infinite. In the present case, a negative slope should be interpreted as an infinite estimate of $D\sigma^2$. Therefore we chose to work on the slope values and not on $D\sigma^2$ estimates. The following statistics were estimated over all repetitions: (1) the mean relative bias between the value of the slope and the expected value $1/(4\pi D\sigma^2)$; (2) the standard error on this relative bias; and (3) the mean square error (MSE = Bias^2 + var). The bias and the MSE are relative values, as they are computed from the ratio of the estimate to the value to be estimated, $1/(4\pi D\sigma^2)$. We also computed the proportion of negative slopes found and the probability that the estimate was within a factor of 2 from $1/4\pi D\sigma^2$. Note that the latest measure is strictly equivalent to the probability that the $D\sigma^2$ estimate was within a factor of 2 from the expected $D\sigma^2$ value.

Table 1
Coverage Probability of 95% Confidence Intervals Around the Regression Slope Using an ABC Bootstrap Procedure

	Bootstrap Sample Size		
	7 loci	13 loci	25 loci
Coverage probability	0.842	0.885	0.90
Proportion of intervals below the slope value	0.020	0.030	0.030
Proportion of intervals above the slope value	0.138	0.085	0.070

An accurate estimate of the uncertainty associated with parameter estimates is important to avoid misleading inferences. The nonparametric ABC bootstrap procedure described in DiCiccio and Efron (1996) was adapted to compute 95% confidence intervals around the regression slope. ABC bootstrap is a procedure that generates approximated bootstrap confidence intervals without real resampling. It is useful for estimation methods with high computation time needs. In this procedure, we considered genotypic data at each locus as independent replicates of the genealogical process. Tests of this procedure were performed using the same simulation program described above by calculating probability coverage of the confidence intervals for 1,000 simulated data sets. We choose arbitrarily a dispersal distribution with $\sigma^2 = 4$ [parameters given in equation (1)]. For each repetition, 100 individuals were sampled every two lattice nodes within an area of $(10\sigma \times 10\sigma)$ on a (100×100) lattice. Estimates of a_r and 95% confidence intervals were calculated for 7, 13, or 25 loci evolving under a SMM with a mutation rate equal to 5×10^{-4}.

Results

ABC Bootstrap

Table 1 shows that the non parametric ABC bootstrap procedure gives inaccurate 95% confidence intervals in terms of coverage probability even for large number of loci (e.g., coverage probability is 0.90 instead of 0.95 for 25 loci). The inaccuracy mostly concerns the lower bound of the confidence intervals for the regression slope (i.e., the proportion of intervals above the slope value is 0.07 instead of 0.025 for 25 loci; table 1). This may reflect the asymmetrical shape of the distribution with a long tail for small values (i.e., large $D\sigma^2$, data not shown). The effect of asymmetrical distribution on ABC bootstrap was tested on a simpler statistical model. ABC confidence intervals were computed for the mean of a random sample drawn in a bivariate student distribution with density

$$\Pr(r) = 2\pi r \frac{\Gamma[1+p]}{\pi u \Gamma[p]} (1 + r^2/u)^{-1-p}$$

and parameters (p,u) being (1,1). This distribution is asymmetrical with an infinite kurtosis and an infinite skewness. Even for very large sample sizes (5000

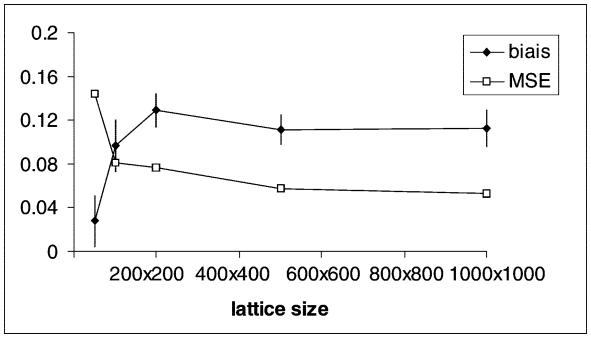

FIG. 2.—Influence of the lattice size on the estimation of the product $1/4\pi D\sigma^2$. NOTE—Only 500 iteration were done for each case. Vertical bars represent standard errors on the bias.

replicates, results not shown), the ABC procedure gives an inaccurate upper bound, resulting in underestimated confidence intervals (results not shown). In the case of the regression slope, the inaccuracy increases for small sample size (e.g., 0.842 instead of 0.95 for seven loci; table 1).

Because of the important computation time needed to construct ABC confidence intervals, this procedure was not used for evaluating the influence of the sampling scale and mutational factors on the estimation of $D\sigma^2$ (see *Models and Methods*).

Influence of the Sampling Scale

Previous simulations with two-allele loci suggested that the regression method would be efficient if one can sample all individuals within an area of about $10\sigma \times 10\sigma$, giving a sample size of $100D\sigma^2$ individuals (Rousset 2000). It is worth noting that if $D\sigma^2$ is greater than say 5, it becomes difficult in practice to sample and genotype all individuals (>500 individuals). Hence, since the number of individuals to sample is necessarily limited, the method should be less efficient when $D\sigma^2$ increases. In practice, biologists collect samples of a reasonably large number of individuals (say 100) within an area larger or smaller than the recommended ($10\sigma \times 10\sigma$) area when $D\sigma^2$ is small or large respectively. In order to assess the effect of such practical "non-scaled sampling," we simulated a distribution of dispersal with $\sigma^2 = 4$ [parameters given in expression (1)] and four different sampling schemes. One hundred individuals were taken: (1) every lattice node within an area of ($5\sigma \times 5\sigma$), for the first sampling scheme; (2) every two lattice nodes within an area of ($10\sigma \times 10\sigma$), for the second one; (3) every five lattice nodes within an area of ($25\sigma \times 25\sigma$) for the third one; and (4) every ten lattice nodes within an area of ($50\sigma \times 50\sigma$) for the last one. For each repetition the parameter estimated is a_r for 13 loci evolving under a SMM with a mutation rate equal to 5×10^{-4}. We considered that a set of 13 loci represents a reasonable number of loci in empirical studies using microsatellites. A two dimensional lattice of (200×200) individuals was considered for the first three sampling schemes and of (500×500) individuals for the last one, to avoid edge effects on the estimations when considering samples larger than half the length of the lattice. Figure 2 shows that lattice size has no major effect on the estimation, except if it is less than ten times the mean dispersal distance (simulation parameters are those used in this paragraph). Unless the lattice size is very small (50*50), the bias and the MSE do not differ notably from those for a very large lattice size (1000*1000).

The sampling scale seems to have only a limited effect on the MSE of the $D\sigma^2$ estimation (table 2). Whatever sampling scale is considered (i.e., smaller or larger than the recommended area) the MSE is low (values between 5% and 12% in the studied cases). In contrast, the sampling scale has a great effect on the bias. A sample taken from an area two times smaller than the recommended area (first column of table 2) gave a large and positive bias (22%). The bias decreases when the sampling area increases and becomes negative when the sampling area is larger than the recommended area, reaching high values (e.g., −21%, fifth column of table 2). However, it is worth noting that even for extreme sampling situations, estimates of $D\sigma^2$ are not very different from the expected value, as shown by the large proportion of estimated values falling within a factor of two from $D\sigma^2$ (>93%).

Influence of the Mutation Model

The following mutation models were considered: (1) the infinite allele model (IAM); (2) the *K*-allele model (KAM) with an arbitrary choice of $K = 10$ possible allelic states; (3) the stepwise mutation model (SMM); (4) the generalized stepwise model (GSM) with variance of the geometric distribution equal to 0.36; and (5) the GSM with constraints on allele size (bounded GSM). In the bounded GSM, the number of possible allelic states was equal to 10 or 20, each allelic state being separated by a single repeat unit.

Simulations were run considering a sample of 100 individuals for 13 loci evolving in a two-dimensional lattice of (100×100) individuals. For each repetition of the simulation process the parameter estimated is a_r. As it is often not easy in practice to sample most individuals from a small area, we considered a sample of (10×10) individuals taken every two nodes from an area of (20×20) nodes in the lattice. By doing so, we approximated the sampling scheme typically used in empirical studies. We also chose a dispersal distribution with a relatively large σ^2 value [i.e., $\sigma^2 = 4$, parameters given in equation (1)]. The logic underlying this choice is that the method may be inaccurate in this case and that it is more relevant to distinguish differences in efficiency when the method does not perform extremely well, than when it performs well, whatever the mutation model.

The mutation rate was first fixed at 5×10^{-4} for all loci for each mutation model. Our results show that the nature of the mutation model has little influence on the estimation of the product $D\sigma^2$ (table 3). Whatever mutation model is considered, the bias is positive and around 10%. Although the precision of the method is maximum under the IAM (MSE of 6%) and minimum under the GSM with strong constraints ($K = 10$, MSE = 0.11), these differences are small. For all mutation models more

Table 2
Influence of Sampling Scale on the Estimation of $1/4\pi D\sigma^2$

	Sampling Scale (Sampling Area)			
	1 (10 × 10)	2 (20 × 20)	5 (50 × 50)	10 (100 × 100)
Bias	0.219	0.130	−0.056	−0.205
(standard error)	(0.0077)	(0.0077)	(0.0072)	(0.0064)
MSE	0.106	0.0763	0.0554	0.082
2× coverage	0.999	0.996	0.967	0.93
Negative slope	0	0	0	0

NOTE—Sampling area is expressed in lattice node unit (see text for details). 2× coverages correspond to the probability that the estimate was within a factor of 2 from $1/4\pi D\sigma^2$.

than 97% of the estimations are within a factor 2 from the expected $D\sigma^2$ value.

For a given mutation rate, level of genetic diversity varies according to the mutation model considered. Because the level of genetic diversity is likely to have an important effect on the estimation of the product $D\sigma^2$, we studied the influence of different mutational models for the same level of diversity. The genetic diversity can be expressed in terms of probability of identity by $(1 - Q_w)$, where Q_w is the probability of identity in state of two genes taken in the same individual. This corresponds to the fraction of heterozygous individuals in the population. The influence of mutation models was thus studied with the same Q_w value for all mutation models. The conclusions are similar to those obtained with a mutation rate fixed at the same value for all mutation models (table 3). For a given value of genetic diversity, the bias and the MSE of $D\sigma^2$ estimates shows little variation among mutational models.

Influence of the Mutation Rate

The influence of the mutation rate (or the genetic diversity) has been studied for the GSM, a mutation model considered as more realistic for microsatellite loci than the SMM, the KAM, or the IAM (e.g., Estoup and Cornuet 1999). All other simulation parameters are those used for evaluating the influence of the mutation model. Our simulations showed that the mutation rate has a substantial effect on the bias and the MSE (fig. 3 and table 4). The MSE is more strongly influenced by the mutation rate than the bias. For "low" genetic diversities (i.e., H = 0.5), the observed bias is positive and never greater than 12%. In contrast, for genetic diversity lower than 0.6, the MSE is greater than 20% and increases relatively rapidly when the genetic diversity decreases. However, even for a genetic diversity lower than the mean genetic diversity observed in most microsatellite studies (e.g., about 0.5), 85% of the estimations are within a factor of two from $D\sigma^2$, but 15 negative slopes were found (table 4).

It is worth mentioning that the observed bias may be of two types: (1) the bias, inherent in the method, that is due to the effect of high mutation rate on the parameter value (we will name it the "parametric bias"); and (2) the bias due to the deviation of the estimates in relation to the parameter value considering a finite sample of individuals and loci (which we will name "small sample bias"). The method is expected to perform poorly for very high mutation rates because distances between some pairs of sampled individuals are then larger than

$$\frac{\sigma}{\sqrt{2\mu}}$$

(Rousset 1997). In such a case, the parametric bias is expected to be negative because the slope of the regression line will be underestimated (for details, see Rousset 1997). In our simulations, we have $\sigma = 2$ and the maximal distance between individuals equals $20\sqrt{2}$ lattice units, which is within

$$\frac{\sigma}{\sqrt{2\mu}}$$

for mutation rates lower than 0.001. However, our results show that for a genetic diversity of 0.8 (corresponding to a mutation rate of c. 0.005 in our model) the bias and the MSE are very low. The low values of the bias and the MSE in this case are likely to result from some compensatory effects between a positive "small sample bias" and a negative "parametric bias." When higher genetic diversity is considered (i.e., H = 0.85 corresponding to mutation rates of c. 0.05 in our model), the bias becomes large and negative and the MSE rapidly increases (table 4). This result is in agreement with the above prediction: for very high mutation rates the "parametric bias" becomes more important than the "small sample bias," so that the global bias observed for high mutation rates is negative.

It is sometimes considered that the large variation between loci of the mutation rate decreases the precision of parameter estimation in population genetics (e.g., Takezaki and Nei 1996; Gonser et al. 2000). To address this question, we considered 13 loci evolving under the GSM with mutation rates drawn for each locus in a gamma distribution of mean 5×10^{-4} (see earlier under *Models and Methods: Mutation Model*), all other simulation parameter values being the same as those used in the previous section. Our simulation results show that variable mutation rates for microsatellite loci have little effect on the estimation of $D\sigma^2$ (table 4). The bias and the MSE values are 11% and 11%, respectively, which does not differ much from the values of 10% and 9% obtained with a fixed mutation rate of 5×10^{-4}. More than 98% of the estimations are within a factor of 2 from $D\sigma^2$ and no negative estimates were found. Finally, our simulation results show that a linear increase in mutation rates with allele length has little effect on the estimation of $D\sigma^2$ (table 4). Strong or weak

Table 3
Influence of Mutational Processes on the Estimation of $1/4\pi D\sigma^2$ with Constant Mutation Rate or Constant Genetic Diversity for All Mutation Models

	Mutation Model											
	Constant Mutation Rate						Constant Genetic Diversity					
	IAM	KAM ($K = 10$)	SMM	GSM	Bounded GSM ($K = 10$)	Bounded GSM ($K = 20$)	IAM	KAM ($K = 10$)	SMM	GSM	Bounded GSM ($K = 10$)	Bounded GSM ($K = 20$)
Genetic diversity	0.787	0.711	0.703	0.772	0.679	0.720	0.68	0.68	0.68	0.68	0.68	0.68
Mutation rate	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0001	0.000218	0.000342	0.00012	0.0005	0.0002
Bias	0.109	0.0919	0.0917	0.104	0.0997	0.128	0.111	0.104	0.118	0.121	0.0997	0.108
(standard error)	(0.0067)	(0.0088)	(0.0093)	(0.00863)	(0.0101)	(0.009)	(0.01)	(0.01)	(0.015)	(0.0120)	(0.0101)	(0.010)
MSE	0.057	0.0853	0.0953	0.0852	0.112	0.098	0.119	0.109	0.119	0.159	0.112	0.121
2× coverage	0.998	0.982	0.975	0.987	0.976	0.984	0.96	0.97	0.96	0.938	0.976	0.962
Negative slope	0	0	0	0	0	0	0.001	0.001	0	0.001	0	0.002

FIG. 3.—Influence of the mutation rate on the estimation of the product $1/4\pi D\sigma^2$. The mutation model is a GSM.

variations give similar results. The bias and the MSE values are about 10%–11% and 8%, respectively, which again does not differ much from the values of 10% and 9% obtained with a fixed mutation rate of 5×10^{-4}. No negative estimates were found, and more than 99% of the estimations are within a factor of 2 from $D\sigma^2$.

Test for a Statistic Taking into Account Allele Size Differences

The behavior of the statistic b_r, an equivalent of a_r based on allele sizes, has been studied under both the SMM (i.e., the mutation model under which this statistic is expected to perform optimally) and the GSM with a mutation rate fixed at 5×10^{-4}. All other simulation parameters values are those used in the two previous sections. Table 5 shows that the method of estimation of $D\sigma^2$ performs poorly when b_r is used. Under both the SMM and GSM, the increase in MSE as well as the number of negative slopes is spectacular. For instance the MSE goes from about 10% when using the classical measure a_r to values greater than 100% when using b_r. In contrast, the bias is only slightly increased compared to estimations using a_r. Although slight, the bias increase appears higher under the GSM than the SMM (+ 9% versus + 4%).

Discussion

A first general conclusion of this study is that the mutation model of the markers has little influence on the efficiency of the method of estimation of $D\sigma^2$ based on individual genotypes and allelic identity. Hence, the allele size homoplasy typically produced under stepwise mutation models (SMM and GSM), and specifically of microsatellite markers (reviewed in Estoup, Jarne, and Cornuet 2002 for different population genetics statistics), is not a feature prejudicial for the method described in this article. Our results dealing with constraints on allele sizes, an evolutionary feature also specific to microsatellite markers and known for substantially increasing size homoplasy, show that even extremely strong constraints (e.g., $K = 10$) have little effect on the estimation of $D\sigma^2$. These results can be interpreted in the context of

Table 4
Influence of the Mutation Rate on the Estimation of the Product $1/4\pi D\sigma^2$

	Mutation Rate					Interloci Variability (*)	Intraloci Variability (**)	
	0.00005	0.00012	0.0005	0.005	0.05		Weak	Strong
Genetic diversity	0.56	0.68	0.77	0.82	0.85	0.77	0.77	0.77
Bias	0.0972	0.121	0.104	0.00946	−0.390	0.114	0.0965	0.111
(standard error)	(0.01609)	(0.0120)	(0.00863)	(0.00616)	(0.0055)	(0.0096)	(0.00846)	(0.0081)
MSE	0.268	0.159	0.0852	0.0380	0.182	0.105	0.0808	0.0778
2× coverage	0.844	0.938	0.987	0.996	0.761	0.983	0.991	0.993
Negative slope	0.015	0.001	0	0	0	0	0	0

NOTE.—The mutation model is a GSM. (*) Mutation rate drawn in a gamma (2, 2.5 10^{-4}) distribution. (**) Variation in mutation rate with allele length is 0.1% and 1% per repeat unit for weak and strong variation, respectively (see text under *Influence of Mutation Rate* for details).

coalescent theory. Values of *F*-statistics, under the assumption of low mutation rate, can be deduced from the comparison between the distributions of coalescence probability for different pairs of genes (e.g., pairs from the same deme and pairs from different demes) (Rousset 1996, 2002). These distributions differ essentially by an "excess" of coalescence probability for the most related genes, this excess being concentrated in a brief period in the recent past. Under isolation by distance, the more distant the demes are, the more the "recent past" is extended to the distant past, permitting more mutations to act and thus to increase the sensitivity to variation in the mutation process. By contrast, sensitivity to range constraints has been observed for statistics that are not related to differences of distribution of coalescence times (e.g., genetic distances, Nauta and Weissing 1996) or for *F*-statistics when the excess probability of coalescence is not concentrated in a recent enough past (large subpopulation sizes and low dispersal rates, Gaggiotti et al. 1999). Because the method of Rousset (2000) focuses on local differentiation and thus on recent evolutionary processes corresponding to a narrow recent past zone, it is no surprise that mutation processes (including allele size constraints) have little influence on the estimation of $D\sigma^2$.

A second major conclusion of this study is that the mutation rate, or the genetic diversity (the latest being largely dependent on the mutation rate), has a strong influence on the estimation of $D\sigma^2$. This is in agreement with previous studies demonstrating that mutation rate is a more important feature than mutation processes for the estimation of demographic parameters through *F*-statistics (reviewed in Rousset 2001*a*; Estoup, Jarne, and Cornuet 2002). Interestingly, the heterozygosities at microsatellite loci are typically between 0.5 and 0.8 (reviewed in Estoup and Angers 1998), a range of values corresponding to the level of genetic diversity that was found to maximize the efficiency of the estimation of $D\sigma^2$. Moreover, the potential effect on the estimation of interlocus and intralocus variability in the mutation rate seems to be weak. Therefore microsatellites are more appropriate to estimate the product $D\sigma^2$ than less polymorphic markers such as allozymes. The importance of the level of variability of the loci used to estimate population parameters has been illustrated by several theoretical and empirical studies. For example, Robertson and Hill (1984) showed that precision in estimates of heterozygote deficiency (F_{is}) increases with the level of variability of the markers. Goudet et al. (1996) also showed that the power of statistical tests of differentiation increases with the number of alleles. In practice, although precise information on mutation rate is difficult to obtain, it is straightforward to calculate a genetic diversity index for a set of markers from which a level of efficiency can be inferred for the estimation of $D\sigma^2$. Our simulations also indicate that future studies should avoid loci with a very high level of genetic diversity (higher than, say, 0.85), because those loci were found to strongly bias negatively the estimations of $D\sigma^2$.

Many studies emphasize that traditional F_{ST} does not make use of the additional information provided by the difference in the number of repeat units at microsatellite loci. However, statistics developed for this purpose often have higher variance than statistics based on allele frequencies (e.g., Gaggiotti et al. 1999). In agreement with this finding, estimates computed using a statistic taking into account allele size differences increases by at least a factor of 10 the MSE compared to a statistic based on identity in state. This result parallels those of Gaggiotti et al. (1999), which showed that in many cases, especially when sample size and number of loci are "small" (i.e., under the conditions of most empirical studies), population structure measures based on allele frequencies alone are more reliable than measures specifically designed for microsatellite loci. Takezaki and Nei (1996) also showed that even for loci evolving under a strict SMM, genetic distances taking into account allele size differences are less efficient for phylogenetic inference than those based on identity in state, especially for short to moderate divergence times. The poor efficiency of this category of statistics appears to be a general feature of studies of evolutionary events, especially those referring to fine geographical and temporal scales.

The effects of the mutation processes and high mutation rates on the estimation of $D\sigma^2$ are expected to be more important at large geographical scales (Rousset 1997). In agreement with this expectation, our results showed that sampling at large distance leads to an underestimation of the regression slope and thus to an overestimation of $D\sigma^2$. Therefore sampling at large distance makes it less likely to detect a pattern of isolation by distance. In contrast, sampling from too small an area leads to an overestimation of the regression slope and thus

Table 5
$D\sigma^2$ Estimation Using a Statistic Taking into Account the Differences in Allele Length (B_r)

	Mutation Model[a]			
	SMM	SMM	GSM	GSM
Parameter estimated	a_r	b_r	a_r	b_r
Bias	0.0917	0.128	0.104	0.19
(standard error)	(0.0093)	(0.036)	(0.00863)	(0.034)
MSE	0.0953	1.13	0.0852	1.25
2× coverage	0.975	0.518	0.987	0.497
Negative slope	0	0.154	0	0.141

NOTE.—Mutation rate is 5.10^{-4}.

[a] SMM: stepwise mutation model; GSM: generalized stepwise mutation model.

to an underestimation of the product $D\sigma^2$. A possible explanation for this overestimation is that the linear relationship between estimates of a_r and the logarithm of the geographical distance is expected to hold less well over very short distances (Rousset 1997). However, using a sample not exactly appropriate to the biological case studied [i.e., a few times larger or smaller than the recommended area of $(10\sigma \times 10\sigma)$] still gives reasonably robust estimations because, in most cases, the estimated $D\sigma^2$ fell within a factor of 2 from the expected $D\sigma^2$ value.

Given our result on bootstrap confidence intervals, we alert biologists using this method on a standard-sized data set (10 loci and 150 individuals, e.g., Sumner et al. 2001) that ABC confidence intervals overestimate the lower bound for the regression slope and thus underestimate the upper bound for $D\sigma^2$. Construction of reliable confidence intervals based on the bootstrap is an ongoing problem for which a satisfactory solution has not yet been found, especially when the number of replications is limited computationally (DiCiccio and Efron 1996). Nevertheless, the ABC bootstrap procedure evaluated here should give an idea of the uncertainty of the $D\sigma^2$ estimate, namely a correct lower bound for $D\sigma^2$ and a minimal value for the upper bound. This procedure will be implemented in the next version of the population genetics package Genepop (Raymond and Rousset 1995).

Conclusion

Three conclusions inferred from our simulation study have important consequences for empirical investigations. First, we recommended using loci with high levels of polymorphism (genetic diversity around 0.7), although loci with too high genetic diversity, e.g., more than 0.85, should be avoided. Because the mutational processes, specifically size homoplasy and allele size constraints, have little influence on $D\sigma^2$ estimations, microsatellite markers seem to be the best choice at the present time. Second, using statistics based on allele size differences at microsatellite loci gives unreliable estimations of $D\sigma^2$ because of the very high variance of those estimations. Third, it is important to restrict the sampling design to a relatively small geographical area in order to work at a local geographical scale; however, it is necessary to sample on a relatively large scale when σ is high. Optimizing the method studied here requires a previous knowledge of σ, and we therefore recommended using a preliminary estimate of σ to allow subsequent design of an appropriate sampling scheme. In the absence of a preliminary estimate of σ, a rough estimate of this parameter deduced from consideration of known dispersal mechanisms should be useful to define the minimal scale of the study (e.g., Leblois et al. 2000). If these aspects are approximately satisfied, the method should give estimates of the product $D\sigma^2$ with low bias and low mean square error. Finally, the ABC bootstrap procedure, as implemented in the package Genepop (Raymond and Rousset 1995), should be useful to estimate a 95% confidence interval on $D\sigma^2$, although the upper bound of this interval is likely to be underestimated.

Acknowledgments

We thank R. Streiff, B. Danforth, and three anonymous reviewers for constructive comments on an earlier version of the manuscript. This work was supported financially by the AIP no. 00202 "biodiversité" from the Institut Français de Biodiversité. This is paper 2003-002 of the Institut des Sciences de l'Evolution.

Literature Cited

Amos, W. 1999. A comparative approach to the study of microsatellite evolution. Pp. 66–79 *in* D. B. Goldstein and C. Schlötterer, eds. Microsatellites: evolution and applications. Oxford University Press, Oxford.

Bahlo, M., and R. C. Griffiths. 2000. Inference from GeneTree in a subdivided population. Theor. Pop. Biol. **57**:79–95.

Barton, N. H., F. Depaulis, and A. M. Etheridge. 2002. Neutral evolution in spatially continuous populations. Theor. Popul. Biol. **61**:31–48.

Brohede, J., C. Primmer, A. Møller, and H. Ellegren. 2002. Heterogeneity in the rate and pattern of germline mutation at individual microsatellite loci. Nucleic Acids Res. **30**:1997–2003.

Clark, J. S., M. Silman, R. Kern, E. Macklin, and J. HilleRis-Lambers. 1999. Seed dispersal near and far: patterns across temperate and tropical forests. Ecology **80**:1475–1494.

Crawford, T. J. 1984. The estimation of neighborhood parameters for plant populations. Heredity **52**:273–283.

Crow, J. F., and M. Kimura, 1970. An introduction to population genetics theory. Harper & Row, New York.

Dib, C., S. Faure, C. Fizames et al. 1996. A comprehensive genetic map of the human genome based on 5,264 microsatellites. Nature **380**:152–154.

DiCiccio, T. J., and B. Efron. 1996. Bootstrap confidence intervals (with discussion). Stat. Sci. **11**:189–228.

Ellegren, H. 2000. Heterogeneous mutation processes in human microsatellite DNA sequences. Nat. Genet. **24**:400–402.

Endler, J. A. 1977. Geographical variation, speciation, and clines. Princeton University Press, Princeton, N.J.

Estoup, A., and B. Angers. 1998. Microsatellites and minisatellites for molecular ecology: theoretical and empirical considerations. Pp. 55–86 *in* G. Carvalho, ed. Advances in molecular ecology. NATO ASI series. IOS Press, Amsterdam.

Estoup, A., and J.-M. Cornuet. 1999. Microsatellite evolution: inferences from population data. Pp. 49–65 *in* D. B. Goldstein and C. Schlötterer, eds. Microsatellites: evolution and applications. Oxford University Press, Oxford.

Estoup, A., P. Jarne, and J.-M. Cornuet. 2002. Homoplasy at microsatellite loci and its consequences for population genetics analysis. Mol. Ecol. **11**:1591–1604.

Estoup, A., I. J. Wilson, C. Sullivan, J.-M. Cornuet, and C. Moritz. 2001. Inferring population history from microsatellite and enzyme data in serially introduced cane toads, *Bufo marinus*. Genetics **159**:1671–1687.

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

Gaggiotti, O. E., O. Lange, K. Rassmann, and C. Gliddon. 1999. A comparison of two methods for estimating average levels of gene flow using microsatellites data. Mol. Ecol. **8**:1513–1520.

Goldstein, D. B., A. R. Linares, L. L. Cavalli-Sforza, and M. W. Feldman. 1995. Genetic absolute dating based on microsatellites and the origin of modern humans. Proc. Natl. Acad. Sci. USA **92**:6723–6727.

Gonser, R., P. Donnelly, G. Nicholson, and A. Di Rienzo. 2000. Microsatellite mutations and inferences about human demography. Genetics **154**:1793–1807.

Goudet, J., M. Raymond, T. de Meeüs, and F. Rousset. 1996. Testing differentiation in diploid populations. Genetics **144**:1931–1938.

Hastings, A., and S. Harrison. 1994. Metapopulation dynamics and genetics. Annu. Rev. Ecol. Syst. **25**:167–188.

Huang, Q.-Y., F.-H. Xu, H. Shen, H.-Y. Deng, Y.-J. Liu, Y.-Z. Liu, J.-L. Li, R. R. Becker, and H.-W. Deng. 2002. Mutation patterns at dinucleotide microsatellite loci in humans. Am. J. Hum. Genet. **70**:625–634.

Hudson, R. R. 1990. Gene genealogies and the coalescent process. Pp. 1–44 *in* D. Futuyama and J. Antonovics, eds. Oxford surveys in evolutionary biology. Oxford University Press, Oxford.

Kimura, M., and J. F. Crow. 1964. The number of alleles that can be maintained in a finite population. Genetics **49**:725–738.

Kingman, J. F. C. 1982*a*. The coalescent. Stochast. Proc. Appl. **13**:235–248.

———. 1982*b*. On the genealogy of large populations. J. Appl. Prob. **19A**:27–43.

Koenig, W. D., D. Van Vuren, and P. N. Hooge. 1996. Detectability, philopatry, and the distribution of dispersal distances in vertebrates. Trends Ecol. Evol. **11**:514–517.

Kot, M., M. A. Lewis, and P. van den Driessche. 1996. Dispersal data and the spread of invading organisms. Ecology **77**:2027–2042.

Leblois, R., F. Rousset, D. Tikel, C. Moritz, and A. Estoup. 2000. Absence of evidence for isolation by distance in expanding cane toad (*Bufo marinus*) population: an individual-based analysis of microsatellite genotypes. Mol. Ecol. **9**:1905–1909.

Malécot, G. 1950. Quelques schémas probabilistes sur la variabilité des populations naturelles. Ann. Univ. Lyon A **13**:37–60.

———. 1967. Identical loci and relationship. Pp. 317–332 *in* L. M. Lecam and J. Neyman, eds. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 4. California University Press, Berkeley.

———. 1975. Heterozygoty and relationship in regularly subdivided populations. Theor. Popul. Biol. **8**:212–241.

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

Michalakis, Y., and L. Excoffier. 1996. A generic estimation of population subdivision using distances between alleles with special interest to microsatellite loci. Genetics **142**: 1061–1064.

Nath, H. B., and R. C. Griffiths. 1996. Estimation in an island model using simulation. Theor. Pop. Biol. **50**:227–253.

Nauta, M. J., and F. J. Weissing. 1996. Constraints on allele size at microsatellite loci: implications for genetic differentiation. Genetics **143**:1021–1032.

Nordborg, M. 2001. Coalescent theory. Pp. 179–208 *in* D.A. Balding, M. Bishop and C. Cannings, eds. Handbook of statistical genetics. John Wiley & Sons, Chichester, U.K.

Ohta, T., and M. Kimura. 1973. A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genet. Res. **22**:201–204.

Pope, L. C., A. Estoup, and C. Moritz. 2000. Phylogeography and population structure of an ecotonal marsupial, *Bettongia tropica*, determined using mtDNA and microsatellites. Mol. Ecol. **9**:2041–2053.

Portnoy, S., and M. F. Willson. 1993. Seed dispersal curves: behavior of the tail of the distribution. Evol. Ecol. **7**:25–44.

Pritchard, J. K., M. T. Seielstad, A. Perez-Lezaun, and M. W. Feldman. 1999. Population growth of human Y chromosome microsatellites. Mol. Biol. Evol. **16**:1791–1798.

Raymond, M., and F. Rousset. 1995. GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. J. Hered. **86**:248–249.

Reich, D. E., and D. B. Goldstein. 1998. Genetic evidence for a paleolithic human population expansion in Africa. Proc. Natl. Acad. Sci. USA **95**:8119–8123.

Robertson, A., and W. G. Hill. 1984. Deviations from Hardy-Weinberg proportions: sampling variances and use in estimation of inbreeding coefficients. Genetics **107**:703–718.

Rousset, F. 1996. Equilibrium values of measures of population subdivision for stepwise mutation processes. Genetics **142**:1357–1362.

———. 1997. Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. Genetics **145**:1219–1228.

———. 2000. Genetic differentiation between individuals. J. Evol. Biol. **13**:58–62.

———. 2001*a*. Genetic approaches to the estimation of dispersal rates. Pp. 18–28 *in* J. Clobert, E. Danchin, A. A. Dhondt, and J. D. Nichols, eds. Dispersal: individual, population and community. Oxford University Press, Oxford.

———. 2001*b*. Inferences from spatial population genetics. Pp. 239–265 *in* D. A. Balding, M. Bishop, and C. Cannings, eds. Handbook of statistical genetics. John Wiley & Sons, Chichester, U.K.

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

Schlötterer, C. 2000. Evolutionary dynamics of microsatellite DNA. Chromosoma **109**:365–371.

Slatkin, M. 1993. Isolation by distance in equilibrium and non-equilibrium populations. Evolution **47**:264–279.

———. 1994. Gene flow and population structure. Pp. 3–17 *in* L. A. Real, ed. Ecological genetics. Princeton University Press, Princeton, N.J.

———. 1995. A measure of population subdivision based on microsatellite allele frequencies. Genetics **139**:457–462.

Spong, G., and S. Creel. 2001. Deriving dispersal distances from genetic data. Proc. R. Soc. Lond. Ser. B **268**:2571–2574.

Stumpf, M. P. H., and D. B. Goldstein. 2001. Genealogical and evolutionary inference with the human Y chromosome. Science **291**:1738–1742.

Sumner, J., F. Rousset, A. Estoup, and C. Moritz. 2001. "Neighborhood" size, dispersal and density estimates in the prickly forest skink (*Gnypetoscincus queenslandiae*) using individual genetic and demographic methods. Mol. Ecol. **10**:1917–1927.

Tajima, F. 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics **105**:437–460.

Takezaki, N., and M. Nei. 1996. Genetic distances and reconstruction of phylogenetic trees from microsatellites DNA. Genetics **144**:389–399.
Weber, J. L., and C. Wong. 1993. Mutation of human short tandem repeats. Hum. Mol. Genet. **2**:1123–1128.
Wilson, I. J., and D. J. Balding. 1998. Genealogical inference from microsatellite data. Genetics **150**:499–510.
Wright, S. 1943. Isolation by distance. Genetics **28**:114–138.
———. 1946. Isolation by distance under diverse systems of mating. Genetics **31**:39–59.

Pierre Capy, Associate Editor

Accepted October 11, 2002