# Estimation of Effective Population Size and Migration Rate From One- and Two-Locus Identity Measures

**Renaud Vitalis*,†,‡ and Denis Couvet§**

*Laboratoire Génétique et Environnement, Institut des Sciences de l'Évolution de Montpellier, Université Montpellier II, 34095 Montpellier Cedex 05, France, †Station Biologique de la Tour du Valat, Le Sambuc, 13200 Arles, France, ‡Laboratoire Génome, Populations et Interactions, Université Montpellier II, 34095 Montpellier Cedex 05, France and §CRBPO-Museum National d'Histoire Naturelle, 75005 Paris, France

Manuscript received May 10, 2000
Accepted for publication November 9, 2000

ABSTRACT

Standard methods for inferring demographic parameters from genetic data are based mainly on one-locus theory. However, the association of genes at different loci (*e.g.*, two-locus identity disequilibrium) may also contain some information about demographic parameters of populations. In this article, we define one- and two-locus parameters of population structure as functions of one- and two-locus probabilities for the identity in state of genes. Since these parameters are known functions of demographic parameters in an infinite island model, we develop moment-based estimators of effective population size and immigration rate from one- and two-locus parameters. We evaluate this method through simulation. Although variance and bias may be quite large, increasing the number of loci on which the estimates are derived improves the method. We simulate an infinite allele model and a *K* allele model of mutation. Bias and variance are smaller with increasing numbers of alleles per locus. This is, to our knowledge, the first attempt of a joint estimation of local effective population size and immigration rate.

IN finite populations, genes undergo a random sampling process, known as genetic drift. As a consequence of this process, genetic variation is continuously lost. A powerful way to illustrate this point is that if we could trace the genealogy of genes, going backward in time, we would observe a continuous decrease in the number of "ancestor" genes. For neutral genes, the dynamic of this process depends mostly on the effective population size. Indeed, the expected time elapsed since two genes diverged from their common ancestor (the coalescence time) increases with the effective population size. The effective population size, noted $N_e$, has been defined as the size of an ideal population, in which a given genetic parameter (*e.g.*, the rate of inbreeding) takes the same expected value as in the population under scrutiny (Wright 1931). This suggests that there may be many different effective sizes depending on which parameter is chosen. Indeed, the inbreeding effective size has been defined from the expected rate of increase in homozygosity, and the variance effective size is derived from the expected rate of increase of variance in allele frequency per generation (Crow and Denniston 1988). Ewens (1979) also defined the eigenvalue effective size as the first nonunit eigenvalue of the transition matrix for allelic changes in a population. In the particular case of the Wright-Fisher model, inbreeding and variance effective sizes take the same value (Crow and Kimura 1970, chap. 8) as does the eigenvalue effective size (Ewens 1982). This model assumes a finite monoecious population of constant size, nonoverlapping generations, random mating, and equal expected contribution of individuals to the next generation (Fisher 1930; Wright 1931).

The conditions under which the different definitions of $N_e$ coincide can be understood by considering the relationship between the rate of coalescence and the effective population size. The asymptotic rate at which two genes coalesce (the probability of coalescence) in a single ideal monoecious diploid population of $N$ individuals is $1/(2N)$. But consider a structured population with different classes of individuals. Those classes may represent groups of individuals differing by their sex, age, stage, social rank, or geographical position. In contrast with unstructured populations, lines of descent may now coalesce at different rates within classes and among classes.

There is a strict relationship between probabilities of identity by descent and coalescence times (Malécot 1975; Slatkin 1991). Indeed, the leading eigenvalue of the transition matrix describing the increase in identity by descent for pairs of genes gives the long-term rate of coalescence of lineages. It is also equal to the leading nonunit eigenvalue of the transition matrix that gives the change in the distribution of allelic states (Whitlock and Barton 1997). This is true for an arbitrarily structured population. Therefore, the long-term rate of coalescence defines an asymptotic effective size that gives the asymptotic rate of increase in inbreeding,

*Corresponding author:* Renaud Vitalis, Laboratoire de Génétique et Environnement, C.C. 065, Institut des Sciences de l'Évolution de Montpellier, Université de Montpellier II, Place Eugène Bataillon, 34095 Montpellier Cedex 05, France.
E-mail: vitalis@isem.univ-montp2.fr

gene frequency variance, and the expected rate of change in the distribution of allelic states (Whitlock and Barton 1997).

Mating systems, variance in the reproductive success of individuals, changes in population size through time, skewed sex ratios, and overlapping generations are some factors expected to make the effective size different from the census size (Caballero 1994). The effective population size also affects the efficiency of natural selection in maintaining advantageous mutations or promoting the spread of new ones. A small effective size also reduces the ability of natural selection to purge slightly deleterious alleles. Slightly deleterious mutations are more likely to be fixed, decreasing the mean fitness of the population and, thus, further reducing the effective population size (Lynch and Gabriel 1990; Gabriel and Bürger 1994). This runaway process, described as the mutational meltdown, eventually leads to extinction (Lande 1994; van Noordwijk 1994; Lynch *et al.* 1995a,b). But despite its importance in predicting population health or inferring past demography, estimation of effective size in natural populations is still a difficult task (Waples 1989; Schwartz *et al.* 1999).

Since the discrepancy between the census and the effective size of a population depends mainly on the parameters of the mating system and relative reproductive success of individuals, effective population size can be calculated from the direct evaluation of these parameters (Nunney and Elam 1994). However, the demographic data needed to calculate the effective population size are difficult to obtain practically in natural situations (Crow and Denniston 1988). Moreover, single-season assessments of these parameters are known to overestimate long-term effective population size, since interannual fluctuations in population size are not taken into account (Vucetich *et al.* 1997). As an alternative to direct evaluation of effective population size from ecological data, indirect estimates of effective size from genetical data have attracted much attention. So far, indirect estimates have been based on the temporal change in allele frequencies (Nei and Tajima 1981; Pollak 1983; Waples 1989; Williamson and Slatkin 1999), the excess of heterozygotes in progeny (Pudovkin *et al.* 1996; Luikart and Cornuet 1999), and linkage disequilibrium (Hill 1981; Waples 1991; Bartley *et al.* 1992).

It is worth noting that all of these approaches assume a single isolated population. However, in natural situations, it is more likely that populations of the same species are connected, to some extent, by gene flow (Slatkin 1987). Indeed, from an evolutionary and ecological perspective, the only way a species can persist in fragmented and changing habitats is to disperse (Hanski and Gilpin 1997). Moreover, patterns of dispersal are of primary importance in preventing or favoring local adaptation (Maynard Smith and Hoekstra 1980). There have been several attempts to estimate gene flow from genetic data (Slatkin and Barton 1989). However, all the approaches based on $F_{ST}$ or the rare allele method of Slatkin (1985) cannot untangle the effect of drift from the migration pattern (Slatkin and Barton 1989). Indeed, gene flow is estimated as $N_e m$, the product of effective population size and immigration rate, and is referred to as the "effective number of migrants" per generation. Even more powerful methods, based on coalescent theory, have the same drawback (Slatkin and Maddison 1989; Beerli and Felsenstein 1999). Tufto *et al.* (1996) developed a maximum-likelihood method to estimate patterns of migration from allele frequency distribution. Their method provides estimates for short and long range migration rates, but with the effective population size treated as a known parameter.

Here, we propose a method-of-moments approach to infer effective population size on the one hand and immigration rate on the other hand, based on estimates of functions of one- and two-locus identity probabilities. One-locus parameters are functions of first and second moments of allele frequencies. Two-locus identity probabilities are functions of higher-order moments, up to the fourth (Ohta 1982a,b). One-locus identity probabilities are influenced primarily by genetic drift and local immigration. Two-locus identity probabilities are also influenced by drift and migration, through the formation of gametic disequilibria (Weir and Cockerham 1969; Cockerham and Weir 1977; Ohta 1982a,b; Tachida and Cockerham 1986). Thus, the joint analysis of one- and two-locus identity probabilities provides more information on the parameters of interest than the analysis of only one-locus parameters.

In a companion article, we derived the expected values for two-locus probabilities for identity in state in an island model with partial selfing (Vitalis and Couvet 2001). We defined a two-locus parameter, which we call the "within-subpopulation identity disequilibrium," as the excess of two-locus identity probabilities over the product of single-locus probabilities among individuals within subpopulations. Here, we first recall the definitions of one- and two-locus identity probabilities and derive the expectation for one-locus parameters. We then define appropriate statistics for estimating one- and two-locus parameters, whose expectations depend on the population parameters of interest (effective size $N_e$, migration rate $m$) and not on some nuisance parameters, such as the mutation rate or the model of mutation. Then, we present a simple method-of-moments joint estimation of effective size and migration rate. We further explore the reliability of $N_e$ and $m$ estimators by means of stochastic simulations.

## THEORY

**One-locus population structure:** The expectation of any descriptive statistic of population genetic structure

can ultimately be expressed as a function of some probabilities of identity in state for pairs of genes taken at different hierarchical levels (within individuals, among individuals within a population, among populations).

Let us consider an infinite island model of population structure (Wright 1940). This model assumes that the whole population (or the species) is subdivided into an infinite number of subpopulations that exchange genes. We may relax the usual assumption of equal subpopulation sizes and immigration rates and then define as many sets of parameters as there are subpopulations. A focal subpopulation $i$ has $N_{ei}$ diploid individuals and receives $m_i$ migrant genes per generation. We define $Q_{0,i}$ as the probability that two homologous genes randomly sampled in one individual from subpopulation $i$ are identical in state (IIS), $Q_{1,i}$ as the IIS probability for two genes randomly sampled in subpopulation $i$, and $Q_2$ may also be defined as the IIS probability for two genes randomly sampled in the pool of immigrants.

Let us first assume an infinite allele model (IAM). In this model, mutation always creates a new allelic state in the population. Therefore, genes that are IIS are also identical by descent (IBD), *i.e.*, two exact copies (without mutation) of the same ancestral gene (Malécot 1948). Let $a_i = (1 - m_i)^2$ be the frequency of pairs of genes that come from the same subpopulation in the previous generation. Each generation, some offspring are produced by selfing. We define $s$ as the conditional probability that two homologous genes of one individual were produced by the same individual, given that they are copies of genes from one subpopulation in the previous generation. Mutations arise at rate $\mu$, and $\nu = (1 - \mu)^2$. The recursion equations for IBD probabilities in the IAM are given by

$$Q_{0,i}^{t+1} = \nu\left[a_i\left(s\frac{1 + Q_{0,i}^t}{2} + (1 - s)Q_{1,i}^t\right) + (1 - a_i)Q_2^t\right]$$

$$Q_{1,i}^{t+1} = \nu\left[a_i\left(\frac{1}{N_{ei}}\frac{1 + Q_{0,i}^t}{2} + \left(1 - \frac{1}{N_{ei}}\right)Q_{1,i}^t\right) + (1 - a_i)Q_2^t\right]$$

$$Q_2^{t+1} = \nu Q_2^t. \quad (1)$$

A useful parameter to consider is

$$F_i = \frac{Q_{1,i} - Q_2}{1 - Q_2}. \quad (2)$$

The weighted sum of $F_i$ over subpopulations is the intraclass correlation for the probability of identity in state of genes within subpopulations relative to the whole population, which is $F_{ST}$ (Cockerham and Weir 1987; Rousset 1996). At equilibrium,

$$F_i = \frac{\nu a_i}{\nu a_i(2 - \nu a_i) + N_{ei}(1 - \nu a_i)(2 - \nu a_i s)}. \quad (3)$$

For small mutation rates, this parameter is inversely related to the effective number of migrants per generation, $N_{ei}m_i$,

$$F_i \approx \frac{1}{1 + 4N_{ei}m_i}. \quad (4)$$

A measure of this parameter can therefore be used to make some inferences about the product of the local effective population size and immigration rate.

We also derived the recursion equations for IIS probabilities in the $K$ allele model (KAM). In this model, there are a finite number ($K$) of possible allelic states at a locus. Since the unconditional probability of mutation is $\mu$, each gene mutates toward a particular allele with probability $\mu/(K - 1)$. Thus, genes that were IIS in the previous generation are still IIS in the current generation with probability $\nu' = (1 - \mu)^2 + \mu^2/(K - 1)$. Also, genes that were different in state in the previous generation can become IIS in the current generation with probability $\omega = (1 - \nu')/(K - 1)$. Therefore, recursion equations for the IIS probabilities in the KAM are given by

$$Q_{0,i}^{t+1} = a_i\left(s\frac{\nu' + Q_{0,i}'^t}{2} + (1 - s)Q_{1,i}'^t\right) + (1 - a_i)Q_2'^t$$

$$Q_{1,i}^{t+1} = a_i\left(\frac{1}{N_{ei}}\frac{\nu' + Q_{0,i}'^t}{2} + \left(1 - \frac{1}{N_{ei}}\right)Q_{1,i}'^t\right) + (1 - a_i)Q_2'^t$$

$$Q_2^{t+1} = Q_2'^t, \quad (5)$$

with $Q_h'^t = \nu' Q_h^t + \omega(1 - Q_h^t)$ defined as the conditional IIS probabilities for pairs of genes taken in the $h$th state of the hierarchy ($h = 0, 1, 2$), after mutation, given the IIS probabilities $Q_h^t$ before mutation (see Crow and Aoki 1984; Rousset 1996, for similar developments). The range of $Q_{1,i}$ is strongly dependent on the total number of alleles at a locus. In particular, when there are $K$ possible allelic states at a locus, the theoretical range of $Q_{1,i}$ is bounded below by $1/K$ (see Appendix 10 in Crow and Kimura 1970, p. 515). On the other hand, at equilibrium,

$$F_i = \frac{\alpha a_i}{\alpha a_i(2 - \alpha a_i) + N_{ei}(1 - \alpha a_i)(2 - \alpha a_i s)}, \quad (6)$$

with $\alpha = (\nu' K - 1)/(K - 1)$. Equation 6 is of the same form as Equation 3, with $\alpha$ replacing $\nu$. Therefore, for $\mu \ll m$, $F_i$ has the same expectation in the IAM and the KAM. This result suggests that $F_i$ is an appropriate function of IIS probabilities, since this parameter is nearly independent of the number of allelic states (Table 1).

**Two-locus population structure:** The one-locus theory for IIS probabilities can be extended to the two-locus case (Cockerham 1984; Goodnight 1987, 1988). In a random mating population, three two-locus IIS probabilities need to be defined (see Whitlock *et al.* 1993) for pairs, triplets, and quadruplets of sampled haplotypes (we call a haplotype a set of two genes taken at two distinct loci, which is inherited from a single parent, after recombination). However, in any case of departure

**TABLE 1**

**Expected one-locus identity probabilities and two-locus identity disequilibrium, with special reference to the effect of the mutation model**

| $\mu$ | $K$ | $Q_1$ | $F$ | $\eta_S$ | $\eta'_S$ |
|---|---|---|---|---|---|
| $10^{-6}$ | 2 | 0.597586 | 0.195172 | 0.000599 | 0.002398 |
| | 5 | 0.356142 | 0.195177 | 0.001536 | 0.002399 |
| | 10 | 0.275661 | 0.195179 | 0.001944 | 0.002400 |
| $10^{-4}$ | $\infty$ | 0.194397 | 0.194397 | 0.002390 | 0.002390 |
| $10^{-5}$ | | 0.195108 | 0.195108 | 0.002399 | 0.002399 |
| $10^{-6}$ | | 0.195179 | 0.195179 | 0.002400 | 0.002400 |

Parameters of the model are $N_e = 50$, $N_e m = 1$ for a sample size of $n = 50$ diploid individuals. Random mating is assumed in the population.

from random mating, more coefficients must be defined, since IIS probabilities defined for pairs, triplets, or quadruplets of haplotypes may have different expectations whether these haplotypes are sampled in two, three, or four individuals. For this purpose, we defined a total of 10 IIS two-locus probabilities (see Vitalis and Couvet 2001, for further details). Figure 1 depicts the definition of two-locus IIS probabilities.

For example, for two-individual IIS probabilities, $\phi$ is defined as the probability that two haplotypes sampled in two distinct individuals are IIS at the two loci considered. $\gamma$ is defined as the probability that, when a single haplotype of one individual and the two haplotypes of a second individual are sampled, both pairs of homologous genes taken among the two individuals are IIS. And $\delta$ is the probability that, when both haplotypes of each individual are sampled, one pair (across individuals) is IIS at the first locus, and the other (distinct) pair is IIS at the second locus. See Figure 1 and Vitalis and Couvet (2001) for the definition of other IIS probabilities.

In a diploid population, the gametic phase is usually not known. Therefore, we define the compound IIS two-locus probability for two pairs of genes, each taken at two distinct loci, among individuals in a population as

$$\Phi = \frac{\phi + 2\gamma + \delta}{4}. \quad (7)$$

The expected gametic disequilibrium can be expressed by the within-subpopulation identity disequilibrium ($\eta_{S,ij}$), across loci $i$ and $j$. Identity disequilibrium is equivalent to the covariance for a pair of one-locus identity probabilities in a random pair of individuals,

$$\eta_{S,ij} = \Phi_{ij} - \delta_{4ij}, \quad (8)$$

where $\delta_{4ij}$ is the two-locus probability of identity by descent among loci $i$ and $j$, when all genes are sampled from distinct haplotypes. $\eta_S$ is equivalent to the excess

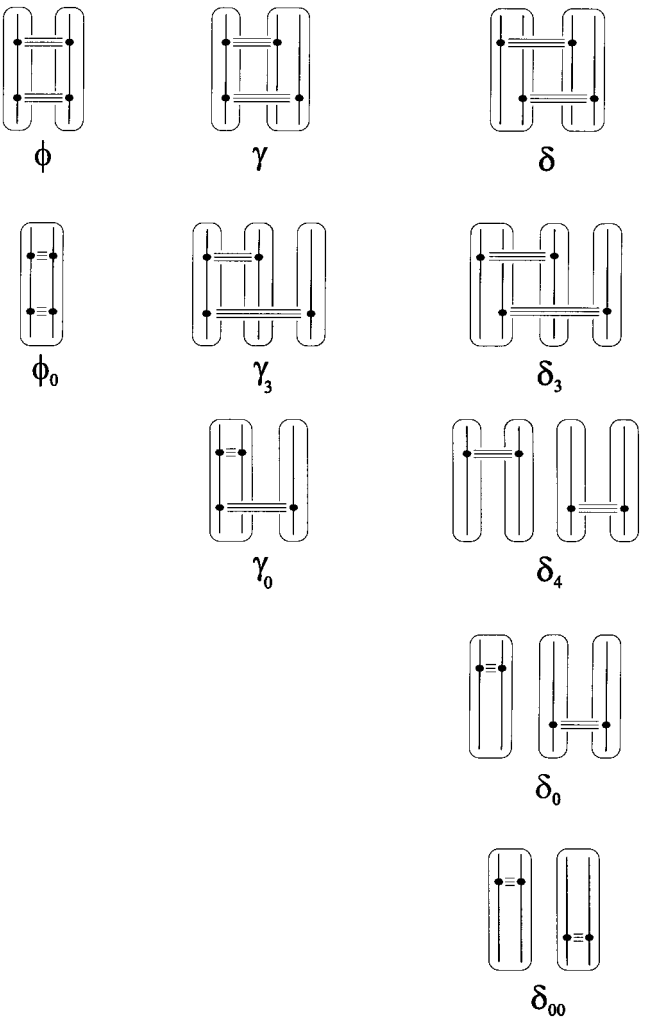


Figure 1.—Definition of two-locus probabilities for the probability of identity in state (IIS). Vertical lines show sampled haplotypes, on which upper and lower positions of solid circles represent two loci. Each diploid individual is represented by a box. Horizontal lines (≡) between pairs of homologous genes stand for identity in state. In the infinite allele model, these coefficients define the corresponding probabilities for identity by descent (IBD). Only the sampled haplotypes are shown.

probability of simultaneous identity over that expected from random combination of the identity at two loci (Ohta 1980). It is also equivalent to the covariance of nonidentity at two loci within populations (Avery and Hill 1979). We derived the recursion equations for all these two-locus probabilities in the IAM (IBD probabilities) and in the KAM (IIS probabilities; Vitalis and Couvet 2001). The parameter $\eta_S$ is a monotonic decreasing function of $N_e$, for a given value of $F$ (Figure 2). Therefore, there is a single pair of $N_e$ and $m$ values that provides a given pair of $\eta_S$ and $F$ values.

It follows also from this graph that, besides any consideration about the variance of the estimates, large effective sizes would be difficult to estimate, since the amplitude of the variation of $\eta_S$ with effective population size

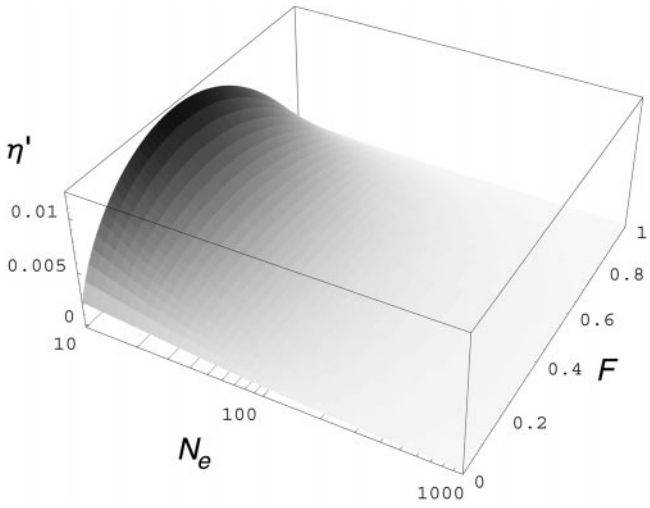


FIGURE 2.—Expected identity disequilibrium $\eta'_S$ as a function of effective population size and one-locus identity probability $F$, in an infinite allele model with $\mu = 10^{-6}$. The effective number of immigrant individuals per generation ($N_e m$) was fixed to 1 and random mating was assumed in the population. Note the logarithmic scale on the $x$-axis.

decreases as the effective size increases. Moreover, the amplitude of variation depends on $F$, suggesting that the efficiency of estimation should vary with the population genetic diversity. The joint analysis of one- and two-locus identity probabilities for unlinked loci should thus permit the estimation of effective population size, provided this parameter is not of too high an order. Obtaining reliable estimates for large population sizes would require tightly linked loci (see also HILL and WEIR 1994).

## ESTIMATION

**One-locus statistics:** For any given allele $u$, we define the indicator variable $x_{ijku}$ to describe the state of the $k$th gene, with $k = (1, 2)$, of the $j$th individual in the $i$th subpopulation. $x_{ijku} = 1$ if the allelic state is $u$, $x_{ijku} = 0$ otherwise. Let $P_i^u$ be the frequency of allele $u$ in the subpopulation $i$. Then, $P_i^u = \mathscr{E}(x_{ijku}|\mathbf{P})$, with $\mathscr{E}(\ |\mathbf{P})$ denoting the expectation, conditional on the array $\mathbf{P}$ of all the allele frequencies. Considering the second moments of the random variable $x_{ijku}$, it follows that $\mathscr{E}(x_{ijku}^2|\mathbf{P}) = P_i^u$ and $\mathscr{E}(x_{ijku}x_{ij'k'u}|\mathbf{P}) = (P_i^u)^2$, with $j' \neq j$ and $k' \neq k$. Summing over all alleles gives the probability that two genes randomly taken from distinct individuals are IIS,

$$Q_{1,i} = \mathscr{E}\left[\sum_{u=1}^{K}(P_i^u)^2\right], \tag{9}$$

where $\mathscr{E}$ denotes now the expectation over the distribution of allele frequencies $\mathbf{P}$. We also define $P_i^{uu}$ as the frequency of homozygotes for allele $u$ in the subpopulation $i$. Then, $\mathscr{E}(x_{ijku}x_{ijk'u}|\mathbf{P}) = (P_i^{uu})^2$. The IIS probability for two genes randomly taken in the whole population can be defined as

$$Q_2 = \mathscr{E}\left[\sum_{u=1}^{K}(x_{ijku}x_{i'j'k'u})\right] \tag{10}$$

with $i' \neq i$. An unbiased estimator of the frequency of allele $u$ among $n$ sampled individuals is given by $\hat{P}_i^u = \sum_{j=1}^{n}\sum_{k=1}^{2}x_{ijku}/(2n)$. Expanding the square of this expression and taking expectation gives $\mathscr{E}[(\hat{P}_i^u)^2|\mathbf{P}] = [P_i^u(1 + 2(n-1)P_i^u) + P_i^{uu}]/(2n)$. Therefore,

$$\hat{Q}_{1,i} = \sum_{u=1}^{K}[\hat{P}_i^u(2n\hat{P}_i^u - 1) - \hat{P}_i^{uu}]/[2(n-1)], \tag{11}$$

where $\hat{P}_i^{uu} = \sum_{j=1}^{n}\sum_{k'\neq k}^{2}x_{ijku}x_{ijk'u}$ is an unbiased estimator of the frequency of homozygotes for allele $u$. An estimator for the IIS probability for genes taken in different subpopulations is given by

$$\hat{Q}_2 = \sum_{u=1}^{K}\sum_{i'\neq i}^{d}\hat{P}_i^u\hat{P}_{i'}^u/[d(d-1)] \tag{12}$$

for $d$ sampled subpopulations. Finally, approximating the expectation of a ratio by the ratio of expectations, an estimator of $F_i$ can be given as

$$\hat{F}_i = \sum_u[(\hat{P}_i^u(2n\hat{P}_i^u - 1) - \hat{P}_i^{uu})/[2(n-1)] - \sum_{i'\neq i}\hat{P}_i^u\hat{P}_{i'}^u/[d(d-1)]]/(1 - \sum_u\sum_{i'\neq i}\hat{P}_i^u\hat{P}_{i'}^u/[d(d-1)]). \tag{13}$$

To combine the information over loci, we define a multilocus estimator as the ratio of the sum of locus-specific numerators over the sum of locus-specific denominators (REYNOLDS *et al.* 1983; WEIR and COCKERHAM 1984).

**Two-locus statistics:** Now, $x_{iju}$ is the indicator variable that describes the state of gene $j$ in individual $i$ at a first locus and $y_{ijv}$ is the indicator variable that describes the state of gene $j$ in individual $i$ at a second, distinct, locus. For the sake of clarity, all subpopulation indices are dropped in this section. $x_{iju} = 1$ if the allelic state at the first locus is $u$, $x_{iju} = 0$ otherwise; and $y_{ijv} = 1$ if the allelic state at the second locus is $v$, $y_{ijv} = 0$ otherwise. Let $P_v^u$ be the frequency of two-locus haplotypes bearing alleles $u$ and $v$ (alleles in phase). Then $P_v^u = \mathscr{E}(x_{iju}y_{ijv}|\mathbf{P})$, with $\mathscr{E}(\ |\mathbf{P})$ denoting the expectation, conditional on the array $\mathbf{P}$ of all the haplotype frequencies. We also define $P_{\cdot v}^{u\cdot} = \mathscr{E}(x_{iju}y_{ij'v}|\mathbf{P})$ as the frequency of pairs of alleles $u$ and $v$ taken from the same individual but from different haplotypes (alleles in repulsion). As in the one-locus case, summing over all possible pairs of alleles and then taking expectations over the distribution of haplotype frequencies $\mathbf{P}$ give the following IIS probabilities:

$$\begin{aligned}\phi &= \mathscr{E}\left[\sum_{u,v}(P_v^u)^2\right]\\ \gamma &= \mathscr{E}\left[\sum_{u,v}(P_v^uP_{\cdot v}^{u\cdot})\right]\\ \delta &= \mathscr{E}\left[\sum_{u,v}(P_{\cdot v}^{u\cdot}P_{\cdot v}^{u\cdot})^2\right].\end{aligned} \tag{14}$$

However, since the gametic phase is usually unknown, $P_v^u$ and $P_{\cdot v}^{u\cdot}$ are not measurable in practice from genotypic

data. We rather evaluate their mean, noted $P_{\bar{v}}^{\bar{u}}$, as $P_{\bar{v}}^{\bar{u}} = (P_v^u + P_{\cdot v}^{u\cdot})/2$. Therefore, the IIS probability for two haplotypes among individuals within a population is defined as

$$\Phi = \mathscr{E}\left[\sum_{u,v}(P_{\bar{v}}^{\bar{u}})^2\right]. \tag{15}$$

An unbiased estimate of $P_{\bar{v}}^{\bar{u}}$ is given by $\hat{P}_{\bar{v}}^{\bar{u}} = \sum_{i=1}^{n}\sum_{j,j'}^{2} \times x_{iju}y_{ij'v}/(4n)$. Expanding the square of this expression and then taking expectation, conditional on the array **P** of all the haplotype frequencies, gives

$$\mathscr{E}\left[(\hat{P}_{\bar{v}}^{\bar{u}})^2|\mathbf{P}\right] = [P_{\bar{v}}^{\bar{u}}[1 + 4(n-1)P_{\bar{v}}^{\bar{u}}] + P_{vv}^{uu} + P_{vv}^{u\cdot} + P_{v\cdot}^{uu}]/(4n), \tag{16}$$

where $P_{vv}^{uu}$ is the frequency of double homozygotes for alleles $u$ and $v$, $P_{v\cdot}^{uu}$ is the frequency of individuals that are homozygotes for allele $u$ at the first locus and that carry one copy of allele $v$ at the second locus, and where $P_{vv}^{u\cdot}$ is the frequency of individuals that are homozygotes for allele $v$ at the second locus and that carry one copy of allele $u$ at the first locus. Therefore, an unbiased estimator for $\Phi$ is

$$\hat{\Phi} = \sum_{u,v}[\hat{P}_{\bar{v}}^{\bar{u}}(4n\hat{P}_{\bar{v}}^{\bar{u}} - 1) - \hat{P}_{vv}^{uu} - \hat{P}_{v\cdot}^{uu} - \hat{P}_{vv}^{u\cdot}]/[4(n-1)], \tag{17}$$

where $\hat{P}_{vv}^{uu}$ is the observed frequency of double homozygotes for alleles $u$ and $v$, and where $\hat{P}_{v\cdot}^{uu}$ (respectively $\hat{P}_{vv}^{u\cdot}$) is the observed frequency of homozygotes for allele $u$ (respectively $v$) that carry one copy of allele $v$ (respectively $u$) at the second locus. An estimator of the identity disequilibrium among loci $i$ and $j$ is given by

$$\hat{\eta}_{S,ij} = \hat{\Phi}_{ij} - \hat{Q}_{1i}\hat{Q}_{1j}. \tag{18}$$

Indeed, the expectation of this statistic is

$$\mathscr{E}(\hat{\eta}_{S,ij}) = \Phi_{ij}\left[1 - \frac{2}{n(n-1)}\right] - \frac{4(n-2)}{n(n-1)}\Gamma_{ij} - \frac{(n-2)(n-3)}{n(n-1)}\delta_{4ij} \tag{19}$$

with $\Gamma_{ij} = (\gamma_{3ij} + \delta_{3ij})/2$. For large samples sizes,

$$\mathscr{E}(\hat{\eta}_{S,ij}) \approx \Phi_{ij} - \delta_{4ij}. \tag{20}$$

Thus, provided the sample size is not too small, $\hat{\eta}_{S,ij} = \hat{\Phi}_{ij} - \hat{Q}_{1i}\hat{Q}_{1j}$ is an unbiased estimator of the identity disequilibrium among loci $i$ and $j$.

However, $\eta_S$ depends on the underlying mutation model (Table 1). There has been some debate in the literature about the dependence on some measures of gametic disequilibrium of the underlying allellic frequencies (Hedrick 1987; Lewontin 1988). At the origin of this dispute is the fact that Hedrick (1987) considered allelic frequencies as fixed parameters. An alternative approach is to consider allelic frequencies as random variables, whose distribution depends on some parameters of the population model. In this sense, we expect that a good measure of disequilibrium shall satisfy the following criteria: (i) The parameter depends on the population parameters of interest, and not on any other parameter; (ii) the expectation of an estimator of this quantity, taken over replicates of the stochastic process of drift, depends only on the parameters of interest (unbiased statistic); and (iii) the distribution of this unbiased estimator depends on the parameters of interest, and not on any other parameter. This allows the measures to be compared across loci or populations, as well as to be pooled over loci.

Various parameters of gametic disequilibrium among loci $i$ and $j$ have been standardized with the product of the Hardy-Weinberg heterozygosities or nonidentities (Hedrick 1987; Ohta 1980),

$$\eta'_{S,ij} = \frac{\Phi_{ij} - \delta_{4ij}}{(1 - Q_{2i})(1 - Q_{2j})}. \tag{21}$$

An estimator for the standardized identity disequilibrium $\eta'_{S,ij}$ among loci $i$ and $j$ is

$$\hat{\eta}'_{S,ij} = \frac{\hat{\Phi}_{ij} - \hat{Q}_{1i}\hat{Q}_{1j}}{(1 - \hat{Q}_{2i})(1 - \hat{Q}_{2j})}. \tag{22}$$

Whereas the absolute measure of identity disequilibrium $\eta_S$ depends on the mutation rate and the model of mutation, the standardized measure $\eta'_S$ is insensitive to the mutational process (Table 1), even if the mating system departs from panmixia (not shown). The combination of identity disequilibrium measures over pairs of loci is achieved by taking the ratio of averaged numerators and denominators over pairs of loci and over samples (Ohta 1980).

Estimators for $F$ and $\eta'_S$ have been evaluated through the simulation of a population in an infinite island model. Several mutation models (from the two-allele model to the infinite allele model) were compared. For five alleles and more, the mean values for the parameters are close to their expectations and do not depend on the mutation model (Table 2). This agrees with the criteria given above. The effect of sample size is more pronounced for identity disequilibrium measures than for one-locus parameter measures. Indeed, increasing the sample size decreases the variance among $\eta'_S$ estimates.

**Estimation of $N_e$ by the method-of-moments:** For a focal population within an infinite island model, we have two statistics $\hat{F}$ and $\hat{\eta}'_S$, whose expectations are known functions of the population parameters of interest, namely the effective population size $N_e$ and the immigration rate $m$. Their expectations do not depend on any nuisance parameters such as the mutation model, but do depend on some other parameters, such as the reproductive system or the extent of (physical) linkage of markers. However, these latter parameters may be estimated from independent data. Therefore, we assume in the following that the selfing rate, as well as the recombination rate among loci, is known.

We solve numerically a system of two simultaneous

**TABLE 2**

**Estimated properties for one- and two-locus parameters**

| | | | | $\hat{F}$ | | | $\hat{\eta}'_S$ | |
|---|---|---|---|---|---|---|---|---|
| $K$ | $N_e$ | $n$ | $F$ | Mean | σ | $\eta'_S$ | Mean | σ |
| 2 | 20 | 25 | 0.187917 | 0.187614 | 0.086208 | 0.005229 | 0.005133 | 0.007491 |
| | 20 | 50 | | 0.186695 | 0.080968 | 0.005362 | 0.005379 | 0.005641 |
| | 50 | 25 | 0.195172 | 0.196953 | 0.087050 | 0.002338 | 0.002209 | 0.006024 |
| | 50 | 50 | | 0.196924 | 0.081832 | 0.002398 | 0.002552 | 0.004265 |
| 5 | 20 | 25 | 0.187919 | 0.187542 | 0.058231 | 0.005230 | 0.005152 | 0.003507 |
| | 20 | 50 | | 0.190018 | 0.055673 | 0.005363 | 0.005352 | 0.002881 |
| | 50 | 25 | 0.195177 | 0.194750 | 0.056116 | 0.002339 | 0.002394 | 0.002734 |
| | 50 | 50 | | 0.192533 | 0.050428 | 0.002399 | 0.002437 | 0.001846 |
| 10 | 20 | 25 | 0.187919 | 0.190234 | 0.054219 | 0.005230 | 0.005332 | 0.003074 |
| | 20 | 50 | | 0.189385 | 0.050953 | 0.005363 | 0.005355 | 0.002304 |
| | 50 | 25 | 0.195179 | 0.193779 | 0.046168 | 0.002339 | 0.002337 | 0.001872 |
| | 50 | 50 | | 0.196065 | 0.047252 | 0.002400 | 0.002505 | 0.001540 |
| ∞ | 20 | 25 | 0.187920 | 0.186700 | 0.046215 | 0.005230 | 0.005111 | 0.002249 |
| | 20 | 50 | | 0.189787 | 0.043126 | 0.005363 | 0.005515 | 0.002018 |
| | 50 | 25 | 0.195179 | 0.194349 | 0.040393 | 0.002339 | 0.002265 | 0.001428 |
| | 50 | 50 | | 0.197306 | 0.039202 | 0.002400 | 0.002423 | 0.001133 |

Arithmetic means as well as standard deviations (σ) of $\hat{F}$ and $\hat{\eta}'_S$ are given for various simulated effective population sizes and mutation models. Estimates are based on 1000 measures (see text for details) from a simulated population of size $N_e$, receiving a proportion $1/N_e$ of migrant haplotypes each generation. A total of 50 individuals were sampled without replacement. Random mating was assumed in the population. Eight loci were simulated. The mutation rate was $10^{-6}$. $K$ indicates the number of alleles simulated per locus.

equations, with two unknowns, as follows. For a wide range of $N_e$ values starting from $N_e = 2$, Equation 6 is solved for $m$, with $F = \hat{F}$. Then, for each pair of $N_e$ and $m$ values, the expected value of $\hat{\eta}'_S$ is calculated as

$$\mathcal{E}(\hat{\eta}'_{S,ij}) = [\Phi_{ij}[1 - 2/[n(n-1)]] - \Gamma_{ij}[4(n-2)]/[n(n-1)] - \delta_{4ij}[(n-2)(n-3)]/[n(n-1)]]/[(1-Q_{2i})(1-Q_{2i})] \quad (23)$$

(see Equations 19 and 22) from the recursive equations given in Vitalis and Couvet (2001). The solutions for $N_e$ and $m$ are then obtained for the best fit between this expected value and $\hat{\eta}'_S$ over a wide range of $N_e$ values (up to 10 times the true $N_e$).

## ASSESSING THE METHOD-OF-MOMENTS ESTIMATOR

**Simulation procedure in an infinite island model:** We evaluated the method of inference through simulations. We focused our analysis on a focal population receiving migrants from an infinitely large number of populations. In practice, a single local population consisting of $N_e$ diploid individuals was simulated. Gametic dispersal occurred prior to reproduction. Under the IAM, at every locus, immigrant individuals carried alleles that were absent from the local population. Therefore, the probability to draw IIS genes from two different populations ($Q_2$) was considered to be zero in the IAM and $Q_2 = 1/K$ in the KAM. In both cases, $Q_2$ was taken as a fixed parameter when calculating statistics for the population parameters. Local random mating was assumed in all the results presented thereafter. For each set of simulations, the initial population was formed with $2N_e$ different alleles at each locus. For each set of parameters, the results from 10 independent simulations were pooled. In each simulation replicate, the first measure was realized after 1000 generations, and then every 100 generations, to avoid temporal correlation across samples, for 100 times. Fifty individuals were sampled for each measure. Indeed, since the identity disequilibrium may be very small in many circumstances, reasonably large sample sizes are required. However, this sample size is representative of those found in the literature (see Williamson and Slatkin 1999, and references therein).

Estimation efficiency was also assessed by examining the percentage of successful inferences as compared with unsuccessful ones. Indeed, among the 1000 identity measures obtained for a set of simulations, there were some cases in which reliable $N_e$ estimates could not be inferred. Overall, we distinguished three cases.

*Case 1:* For a given $F$ value, the $\hat{\eta}'_S$ estimate overrides its possible range. This is interpreted as a very small effective population size ($N_e \leq 2$).

*Case 2:* The estimate for $\hat{\eta}'_S$ is negative. This is interpreted as an infinite effective population size. We also excluded estimates that were >10 times the true $N_e$ in this latter category.

*Case 3:* The estimation of $N_e$ is reliable [lying in the interval $(2, 10 \times N_e)$].

**Simulation results:** Concerning the estimation of effective population size, a first source of bias is expected

**TABLE 3**

**Results from simulations for an infinite allele model**

| | | Estimated $\hat{N}_e$ | | | | Estimated $\hat{m}$ | | | Inference efficiency | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $L$ | $N_e$ | $H$ | c.v. | 95% C.I. | $m$ | Mean | c.v. | 95% C.I. | 1 | 2 | 3 |
| 4 | 20 | 17.75 | 0.63 | 8.10–51.88 | 0.05 | 0.0553 | 0.51 | 0.0194–0.1070 | 0.0 | 0.2 | 99.8 |
| | 50 | 43.32 | 0.88 | 19.21–192.77 | 0.02 | 0.0232 | 0.62 | 0.0052–0.0487 | 0.0 | 5.3 | 94.7 |
| | 100 | 70.02 | 1.07 | 26.91–465.71 | 0.01 | 0.0141 | 0.70 | 0.0025–0.0337 | 0.0 | 17.3 | 82.3 |
| 8 | 20 | 18.75 | 0.45 | 10.18–39.06 | 0.05 | 0.0537 | 0.40 | 0.0235–0.0933 | 0.0 | 0.0 | 100.0 |
| | 50 | 47.74 | 0.61 | 24.92–131.15 | 0.02 | 0.0212 | 0.46 | 0.0071–0.0384 | 0.0 | 0.3 | 99.7 |
| | 100 | 91.70 | 0.94 | 41.95–426.41 | 0.01 | 0.0110 | 0.60 | 0.0023–0.0229 | 0.0 | 5.5 | 94.5 |
| 12 | 20 | 18.86 | 0.37 | 9.94–36.64 | 0.05 | 0.0530 | 0.39 | 0.0251–0.0910 | 0.0 | 0.0 | 100.0 |
| | 50 | 48.45 | 0.48 | 26.06–107.73 | 0.02 | 0.0209 | 0.42 | 0.0085–0.0366 | 0.0 | 0.0 | 100.0 |
| | 100 | 98.84 | 0.80 | 47.53–332.14 | 0.01 | 0.0102 | 0.54 | 0.0028–0.0205 | 0.0 | 2.9 | 97.1 |
| 16 | 20 | 18.81 | 0.36 | 10.73–34.68 | 0.05 | 0.0533 | 0.38 | 0.0260–0.0888 | 0.0 | 0.0 | 100.0 |
| | 50 | 48.64 | 0.46 | 26.86–105.44 | 0.02 | 0.0210 | 0.41 | 0.0088–0.0366 | 0.0 | 0.0 | 100.0 |
| | 100 | 97.49 | 0.76 | 50.83–292.93 | 0.01 | 0.0105 | 0.49 | 0.0033–0.0205 | 0.0 | 0.8 | 99.2 |

Results are based on 1000 measures (see text) from a simulated population of size $N_e$ receiving a proportion $m$ of migrant individuals per generation. $n$ is the number of diploid individuals that were sampled without replacement. Harmonic means ($H$) for $N_e$ and arithmetic means for $m$ are given for various simulated effective population sizes and various numbers of loci. Coefficients of variation (c.v.), the ratio of standard deviation over the mean, are also indicated. A 95% confidence interval (C.I.) is defined from the 5th and 95th percentiles. An assessment of inference efficiency is given with special reference to three distinct cases in the process of inference. Case 1 is the percentage of too large $\eta_S$ values for a fit to be reliable, which is expected for a very small effective population size. Case 2 is the percentage of either negative $\eta_S$ estimates, which give an infinite effective size, or $\hat{N}_e$ estimates >10 times the true $N_e$. Case 3 is the percentage of successful inferences from which results are given.

to result from the following. One- and two-locus identity parameters are inversely related to effective population size. As a consequence, even if estimates of parameters are unbiased, effective population size estimates may be biased, since the expectation of an inverse function is not the inverse of the expectation of the function, as discussed in Cockerham and Weir (1993). This is why we present harmonic rather than arithmetic means of $\hat{N}_e$ estimates.

In the IAM, harmonic means for $\hat{N}_e$ may be underestimated (43.32 for $N_e = 50$ and 70.02 for $N_e = 100$) when 4 loci are pooled (Table 3). However, for a greater number of loci, estimates are in very close agreement with true values (Figure 3, Table 3). Indeed, increasing the number of loci reduces the confidence interval, as given by the interval between the 5th and 95th percentiles of the overall distribution of estimates. Increasing the number of loci also increases the percentage of successful inferences (Table 3). Estimations of immigration rates were also in close agreement with the truth. Estimates generally have a high variance, as indicated by the coefficients of variation. However, the variance was substantially reduced when at least 8 loci were sampled. The estimation was very efficient in most cases. In all cases, we note that estimating effective size <2 never occurred (Case 1). Conversely, for 8 loci and less, estimating infinite effective size (Case 2) occurred only when the true effective size was large. For 12 loci and more, the percentage of successful estimation (Case 3) was >95% (Table 3).

In the KAM, harmonic means for $N_e$ are biased downward for 8 loci and less, in a more significant way than in the infinite allele model (Figure 3, Table 4). This point is also indicated by a lower percentage of successful inferences (Table 4). The variance was smaller in a 10-allele rather than in a 5-allele model. Again, increasing the number of loci reduces both the variance and the confidence interval for the estimates. The estimation was efficient (>85% of successful estimates) for 12 loci and more, or for 8 loci with at least 10 alleles. As in the IAM, unsuccessful estimates (Case 1) occurred only when the true effective size was small, and too large estimates (Case 2) occurred only when the true effective size was large. Again, estimates of immigration rates were close to the truth, although they were slightly overestimated in all cases.

We also evaluated the method for various migration rates in an infinite allele model with 8 and 16 independent loci (Table 5, Figure 4). Bias and variance were reduced with a higher effective number of migrants per generation, $N_e m$. Indeed, the dispersion around the mean was large for low migration rate. In this model, lower immigration rates imply also lower levels of gene diversity within subpopulations, and therefore a poorer estimation efficiency.

We obtained a confidence region from the joint distri-

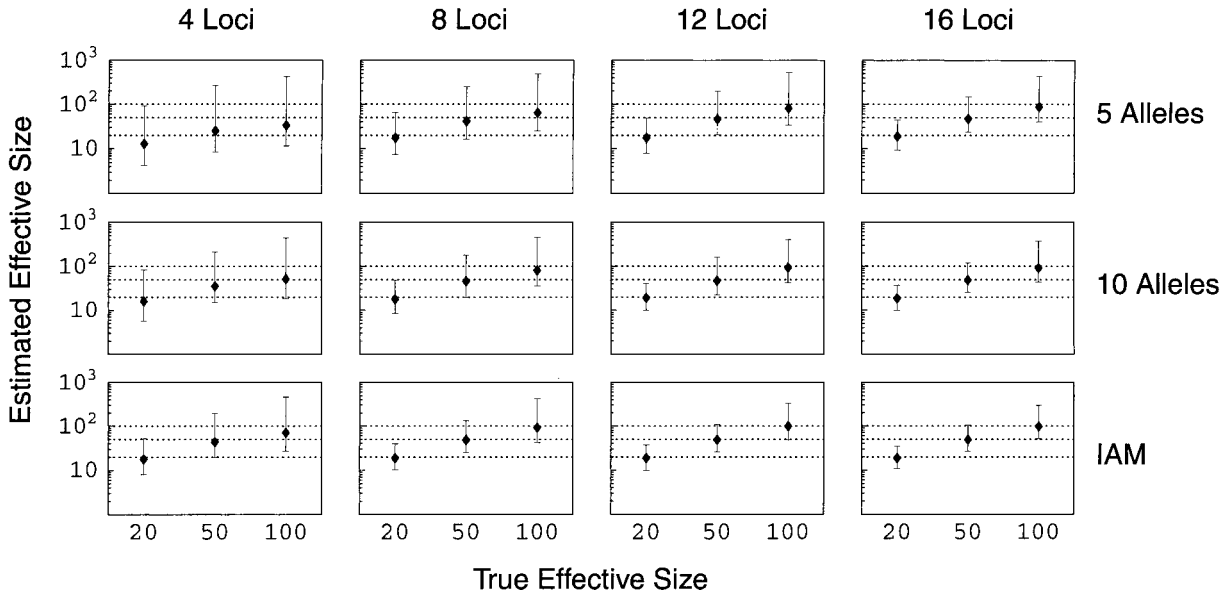


FIGURE 3.—Harmonic means and distribution of $N_e$ estimates for a range of parameters. Three true effective sizes were simulated (20, 50, and 100, as indicated by horizontal dashed lines) with $N_e m = 1$. Simulations were performed for various numbers of unlinked loci, as well as for various mutation models, ranging from 5- and 10-allele models to the infinite allele model, with $\mu = 10^{-6}$. For each set of parameters, 1000 measures of one- and two-locus parameters were performed. The confidence limit gives the 5th and 95th percentiles of the distribution of all realized inferences. Note the logarithmic scale on the *y*-axis. See Tables 3 and 4 for a summary of these sets of simulations.

bution of $N_e$ and $m$ estimates in the following way: All simulations were run to generate a large number of observations, each of which consisted of a pair of $N_e$ and $m$ estimates. These observations were binned to a two-dimensional array of size $100 \times 100$. The bins did not overlap, had the same width, and were evenly distributed in both dimensions. All bin counts were standardized by the total number of counts in the two-dimensional array, and the discrete probability distribution was derived. Then, the cells were sorted in order of decreasing probability. Finally, starting from the cells with the highest associated probabilities, cells were sequentially added to the confidence region until the cumulative probability of the whole set of cells obtained was less than or equal to the preliminary fixed $q$-value. From this procedure, we obtained for each simulation a region within which a fraction $q$ of the data lay. This confidence region was not constrained to be continuous.

## ROBUSTNESS TO MODEL ASSUMPTIONS

**Finite number of demes:** The reliance on the infinite island model might be seen as a serious drawback of our method. However, developing the theory in a finite island model would necessitate as many as 28 distinct two-locus identity coefficients (instead of 10 here) and would be thus far more complicated. In contrast, we assumed that the number of demes had a small effect on our estimators. Therefore, we chose to test our method on data sets generated from a more realistic population model. For this purpose, we simulated a finite island model. Twenty subpopulations were used, each containing the same number of reproducing individuals and receiving the same number of immigrant haplotypes per generation. As in the infinite island model, a single subpopulation was repeatedly sampled.

In contrast to the infinite island model, migration *per se* cannot maintain polymorphism within a subpopulation in a finite island model. Therefore, in our simulations, mutations will arise at a sufficient rate to maintain polymorphism. We chose the mutation rate so that the product $dN_e\mu$ is at least equal to 1. With smaller $dN_e\mu$ values, not enough variation was maintained at equilibrium. In particular, we used two mutation rates, $\mu = 10^{-3}$ and $\mu = 5 \times 10^{-3}$, giving $dN_e\mu = 1$ and $dN_e\mu = 5$ (with $N_e = 50$ and $d = 20$). These two mutation rates remain in the range of realistic values, as given by ESTOUP and ANGERS (1998) for microsatellite markers.

As one might expect, the results depend on the level of genetic variation (Table 5). With $\mu = 10^{-3}$, the bias and variance of the $N_e$ and $m$ estimates are always larger than those obtained with the infinite island model. With $\mu = 5 \times 10^{-3}$, we obtained results that were in very close agreement with those obtained with the infinite island model. With only eight loci, bias and coefficient of variation can even be smaller with the simulations based on the finite island model. In all cases, however, the bias for $N_e$ estimates was positive, although it was always negative when an infinite island model was simulated.

Figure 5 shows the joint distributions of $N_e$ and $m$ estimates over different numbers of loci and different

**TABLE 4**

**Results from simulations for a *K* allele model**

| | | | Estimated $\hat{N}_e$ | | | | Estimated $\hat{m}$ | | | Inference efficiency | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *L* | *K* | $N_e$ | *H* | c.v. | 95% C.I. | *m* | Mean | c.v. | 95% C.I. | 1 | 2 | 3 |
| 4 | 5 | 20 | 12.92 | 1.02 | 4.23–91.29 | 0.05 | 0.0743 | 0.81 | 0.0117–0.1899 | 1.2 | 9.5 | 89.3 |
| | | 50 | 25.18 | 1.23 | 8.37–268.07 | 0.02 | 0.0384 | 0.93 | 0.0048–0.1081 | 0.1 | 26.2 | 73.7 |
| | | 100 | 33.44 | 1.27 | 11.58–423.45 | 0.01 | 0.0294 | 1.06 | 0.0026–0.0835 | 0.0 | 36.1 | 63.9 |
| | 10 | 20 | 16.03 | 0.92 | 5.63–81.84 | 0.05 | 0.0603 | 0.72 | 0.0120–0.1404 | 0.9 | 2.7 | 96.4 |
| | | 50 | 35.21 | 1.04 | 14.70–210.67 | 0.02 | 0.0286 | 0.78 | 0.0048–0.0657 | 0.0 | 16.5 | 83.5 |
| | | 100 | 51.26 | 1.23 | 18.31–440.86 | 0.01 | 0.0191 | 0.82 | 0.0024–0.0488 | 0.0 | 28.2 | 71.8 |
| 8 | 5 | 20 | 17.49 | 0.77 | 7.41–65.3 | 0.05 | 0.0577 | 0.63 | 0.0160–0.1294 | 0.2 | 0.6 | 99.2 |
| | | 50 | 41.36 | 1.03 | 16.27–248.39 | 0.02 | 0.0248 | 0.74 | 0.0042–0.0610 | 0.0 | 9.7 | 90.3 |
| | | 100 | 64.02 | 1.14 | 25.25–489.45 | 0.01 | 0.0158 | 0.75 | 0.0023–0.0393 | 0.0 | 22.3 | 77.7 |
| | 10 | 20 | 17.80 | 0.57 | 8.34–47.65 | 0.05 | 0.0553 | 0.51 | 0.0197–0.1133 | 0.0 | 0.0 | 100.0 |
| | | 50 | 45.57 | 0.80 | 19.60–179.6 | 0.02 | 0.0222 | 0.59 | 0.0056–0.0466 | 0.0 | 2.3 | 97.7 |
| | | 100 | 80.23 | 1.01 | 36.02–460.98 | 0.01 | 0.0126 | 0.64 | 0.0022–0.0276 | 0.0 | 14.3 | 85.7 |
| 12 | 5 | 20 | 17.53 | 0.63 | 7.94–49.46 | 0.05 | 0.0570 | 0.56 | 0.0197–0.1173 | 0.0 | 0.0 | 100.0 |
| | | 50 | 46.00 | 0.89 | 20.59–200.37 | 0.02 | 0.0218 | 0.62 | 0.0048–0.0449 | 0.0 | 3.1 | 96.9 |
| | | 100 | 81.25 | 1.05 | 34.29–525.36 | 0.01 | 0.0127 | 0.72 | 0.0021–0.0304 | 0.0 | 14.5 | 85.5 |
| | 10 | 20 | 19.22 | 0.44 | 9.89–41.09 | 0.05 | 0.0538 | 0.45 | 0.0239–0.1001 | 0.0 | 0.0 | 100.0 |
| | | 50 | 47.08 | 0.79 | 22.74–164.06 | 0.02 | 0.0214 | 0.52 | 0.0061–0.0424 | 0.0 | 0.9 | 99.1 |
| | | 100 | 92.90 | 0.90 | 42.03–400.13 | 0.01 | 0.0108 | 0.62 | 0.0025–0.0235 | 0.0 | 7.8 | 92.2 |
| 16 | 5 | 20 | 18.87 | 0.48 | 9.26–44.62 | 0.05 | 0.0534 | 0.47 | 0.0239–0.1004 | 0.0 | 0.0 | 100.0 |
| | | 50 | 46.83 | 0.75 | 23.65–148.73 | 0.02 | 0.0221 | 0.53 | 0.0067–0.0457 | 0.0 | 1.5 | 98.5 |
| | | 100 | 87.19 | 1.00 | 39.74–430.81 | 0.01 | 0.0117 | 0.65 | 0.0021–0.0262 | 0.0 | 11.0 | 89.0 |
| | 10 | 20 | 18.62 | 0.41 | 10.09–37.61 | 0.05 | 0.0546 | 0.42 | 0.0248–0.0967 | 0.0 | 0.0 | 100.0 |
| | | 50 | 48.60 | 0.52 | 25.83–117.28 | 0.02 | 0.0213 | 0.47 | 0.0077–0.0392 | 0.0 | 0.0 | 100.0 |
| | | 100 | 91.16 | 0.88 | 44.05–370.74 | 0.01 | 0.0110 | 0.56 | 0.0026–0.0221 | 0.0 | 4.2 | 95.8 |

See Table 3 legend for details. *K* is the number of alleles simulated per locus.

**TABLE 5**

**Results from simulations for various population models**

| | Estimated $\hat{N}_e$ | | | | Estimated $\hat{m}$ | | | Inference efficiency | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $N_e m$ | *H* | c.v. | 95% C.I. | *m* | Mean | c.v. | 95% C.I. | 1 | 2 | 3 |
| Simulations based on an infinite island model: $\mu = 10^{-6}$ | | | | | | | | | | |
| 0.5 | 42.70 | 0.95 | 16.59–222.61 | 0.01 | 0.0117 | 0.68 | 0.0022–0.0273 | 0.0 | 5.4 | 94.6 |
| 1.0 | 47.79 | 0.68 | 24.16–133.76 | 0.02 | 0.0211 | 0.48 | 0.0069–0.0392 | 0.0 | 0.3 | 99.7 |
| 2.0 | 48.45 | 0.31 | 30.66–82.54 | 0.04 | 0.0420 | 0.32 | 0.0226–0.0669 | 0.0 | 0.0 | 100.0 |
| Simulations based on a finite island model: $\mu = 10^{-3}$ | | | | | | | | | | |
| 0.5 | 38.87 | 0.98 | 14.62–251.29 | 0.01 | 0.0166 | 1.21 | 0.0023–0.0446 | 0.2 | 13.7 | 86.1 |
| 1.0 | 38.95 | 1.01 | 14.87–226.16 | 0.02 | 0.0334 | 1.25 | 0.0044–0.0934 | 0.4 | 9.3 | 90.3 |
| 2.0 | 36.95 | 0.99 | 15.33–228.84 | 0.04 | 0.0633 | 1.12 | 0.0081–0.1913 | 3.0 | 10.0 | 87.0 |
| Simulations based on a finite island model: $\mu = 5 \times 10^{-3}$ | | | | | | | | | | |
| 0.5 | 56.72 | 0.77 | 27.00–219.13 | 0.01 | 0.0141 | 0.59 | 0.0032–0.0296 | 0.0 | 5.3 | 94.7 |
| 1.0 | 53.49 | 0.64 | 27.91–152.62 | 0.02 | 0.0258 | 0.51 | 0.0082–0.0500 | 0.0 | 0.7 | 99.3 |
| 2.0 | 50.32 | 0.44 | 30.43–97.78 | 0.04 | 0.0502 | 0.48 | 0.0219–0.0906 | 0.0 | 0.2 | 99.8 |

See Table 3 legend for details. For all sets of parameters, eight loci were scored among 50 sampled individuals. An infinite island model was assumed.

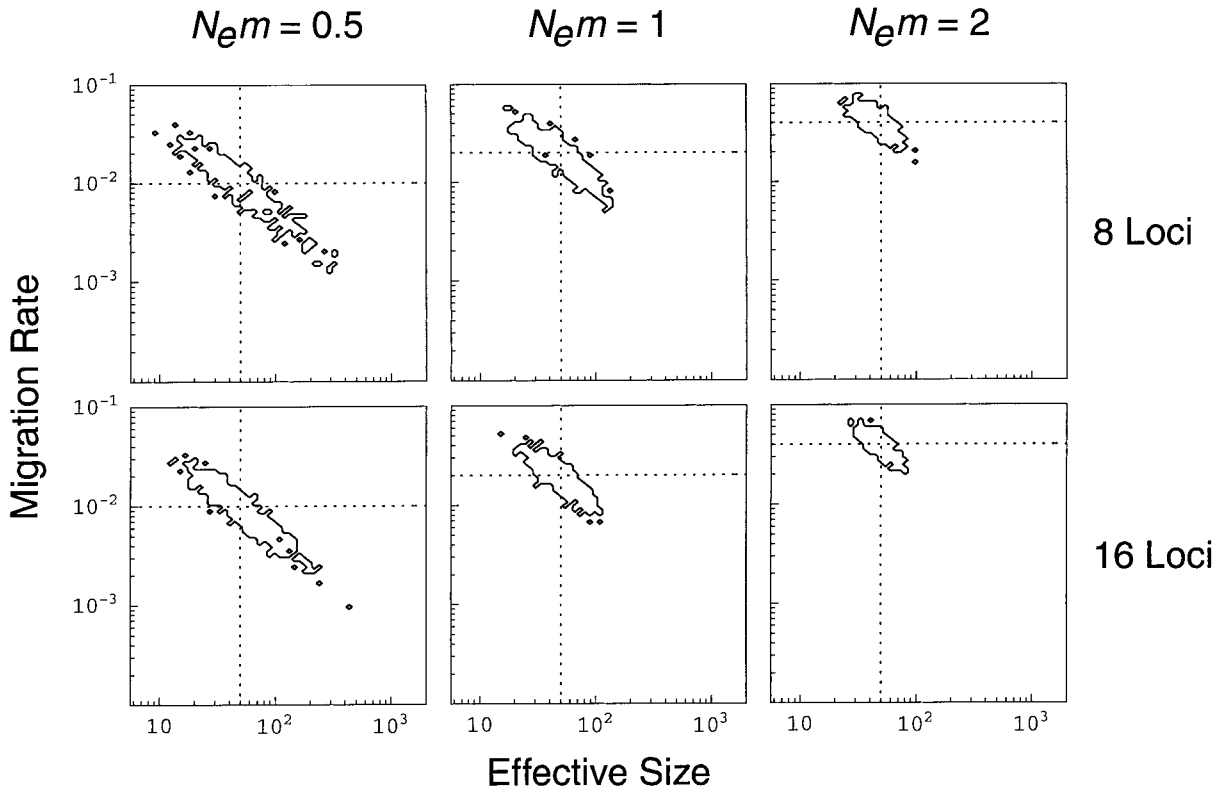


FIGURE 4.—Joint distribution of $N_e$ and $m$ estimates for various effective numbers of migrant individuals per generation and various numbers of loci. In all cases, the true local population size was fixed to 50. Local random mating was assumed, and an infinite allele model of mutation was used ($\mu = 10^{-6}$). For each set of parameters, 1000 estimates of one- and two-locus parameters were obtained (see text for details). The regions plotted are the 95% confidence regions for the sampling distribution of the estimators and were obtained directly from two-dimensional histograms of $N_e$ and $m$ estimates, as explained in the text. Dotted lines show the true values for the parameters. Note the logarithmic scale in both dimensions. See Table 5 for a summary of this set of simulations.

effective numbers of migrants per generation, for two different mutation rates. As in the infinite island model, we obtained better estimates with larger numbers of migrants per generation ($N_em$). With $\mu = 10^{-3}$, even with 16 loci, the distributions of pairwise $N_e$ and $m$ estimates are always broader than the distributions obtained with data generated from infinite island model simulations (to compare, see Figure 4). With $\mu = 5 \times 10^{-3}$ (Table 5, Figure 5B), the bias and variance were very close to the values obtained in the infinite island model. Bias and variance were also greatly reduced when 16 loci were scored (Figure 5). Increasing the sample size up to $n = 100$ also increases the estimation efficiency: more successful inferences are made, all of which give slightly less biased estimates with less variance.

**Departure from equilibrium:** Among the methods that aim at inferring population parameters, many rely on the hypothesis of equilibrium between mutation, migration, and drift. Our approach is no exception. To test whether our method was sensitive to recent departures from migration drift equilibrium, we conducted the same simulations as before with an infinite island model of population structure. But this time, individuals were sampled 10, 20, 30, 40, or 50 generations after the initial state. This process was repeated 1000 times.

Surprisingly, after only 50 generations, bias and variance of joint $N_e$ and $m$ estimates were as small as after 1000 generations (Figure 6). With eight loci sampled among 50 individuals, $\hat{N}_e = 52.33$ [coefficient of variation (c.v.) = 0.60] and $\hat{m} = 0.0216$ (c.v. = 0.45). Estimates were more biased and the variance of the distribution was larger for $m$ than for $N_e$ estimates. Moreover, after 30–40 generations, $N_e$ estimates showed bias and variance as low as after 1000 generations. In all cases, the joint inference was successful in >99% of cases.

## DISCUSSION

Effective population size is directly related to the asymptotic rate of coalescence for neutral genes. Thus, this population parameter determines the rate at which neutral genetic variation is lost from the population. Therefore, any attempt to provide a reliable estimate of effective population size should deserve careful evaluation. Our method is, to our knowledge, the first attempt for a joint estimation of (local) effective population size and immigration rate.

An advantage of using this method-of-moments to estimate $N_e$ is that it requires only a single sample. Our estimates of effective population size are in general

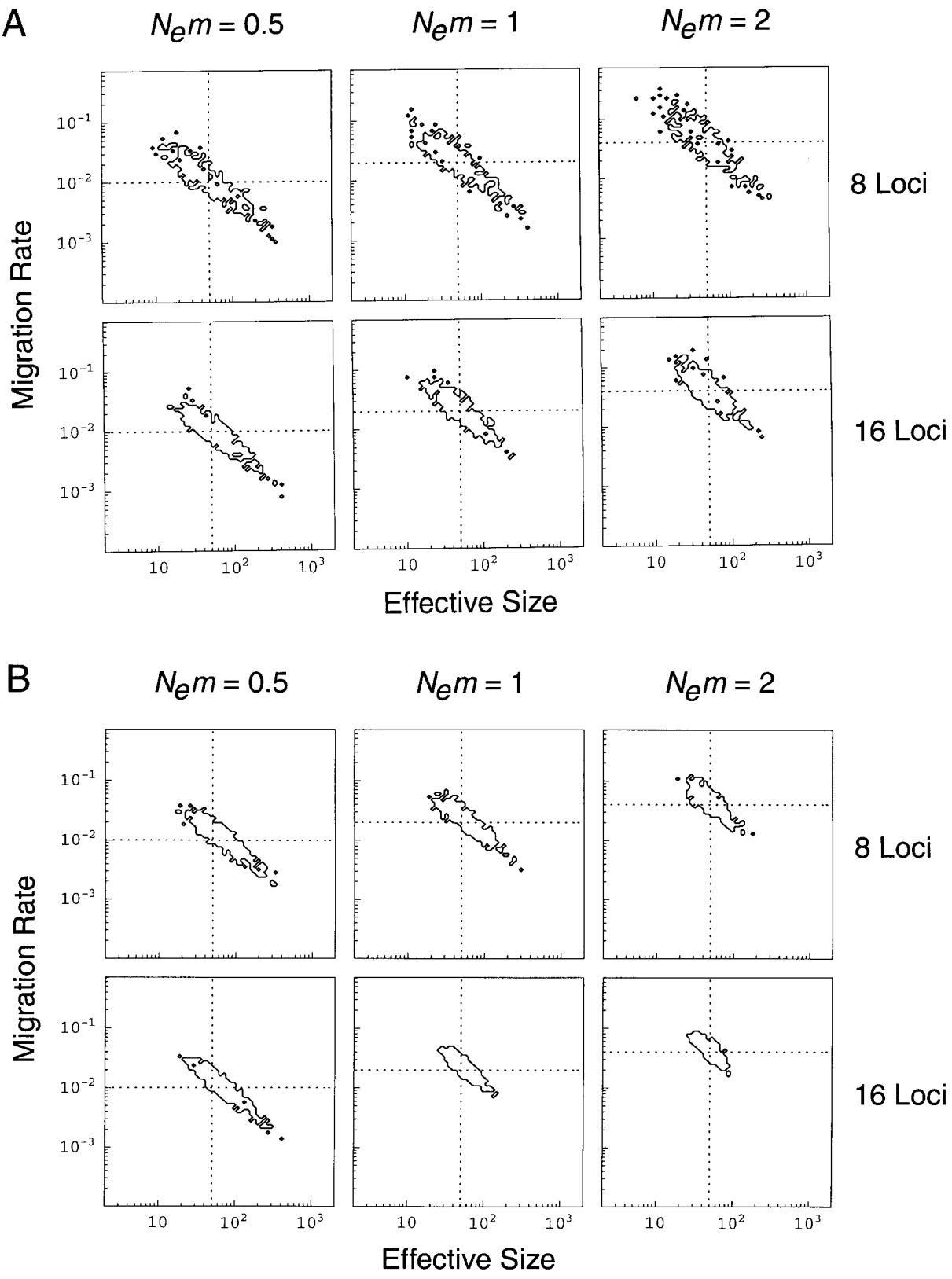


FIGURE 5.—Joint distribution of $N_e$ and $m$ estimates for different values of $N_em$ and different numbers of loci. A finite island model was simulated with 20 demes of equal effective population sizes $N_e = 50$. The migration rate was set to $m = 0.02$. Each time, 50 diploid individuals within a single subpopulation were sampled. See Figure 4 for additional details. (A) $\mu = 10^{-3}$. (B) $\mu = 5 \times 10^{-3}$. Note the logarithmic scale on *x*- and *y*-axes. See Table 5 for a summary of this set of simulations.

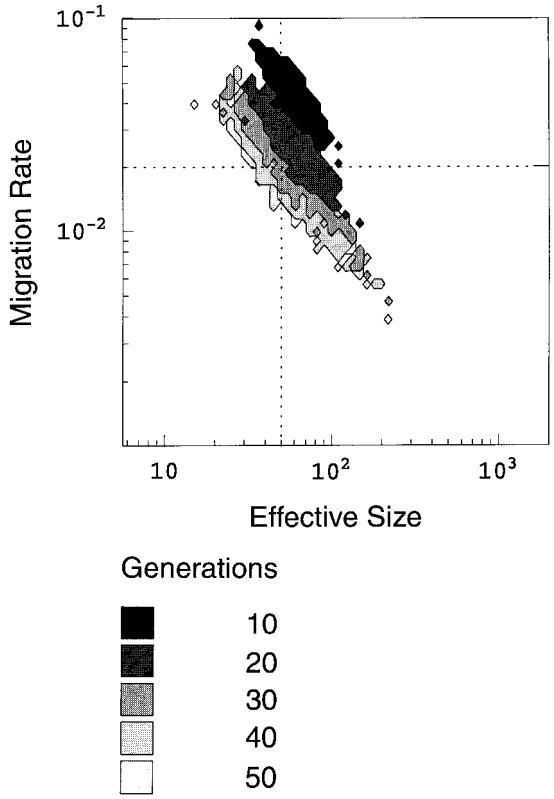


FIGURE 6.—Joint distribution of $N_e$ and $m$ estimates in non-equilibrium situations. An infinite island model was simulated with $N_e = 50$ and $m = 0.02$. Five data sets were obtained as follows. Starting from an initial state where all genes in the population are distinct, 50 individuals were sampled after 10, 20, 30, 40, or 50 generations. In each case, the process was repeated 1000 times. The 95% confidence regions for each data set are shown. An infinite allele model of mutation was used ($\mu = 10^{-6}$). Dotted lines show the true values for the parameters. Note the logarithmic scale on both axes.

slightly biased (Figure 3). The dispersion around the mean, as shown by the coefficient of variation as well as by the 95% confidence interval, is large and increases with the true population size (Figure 3, Table 3). Bias and variance are also decreased when the local immigration rate is increased (Figure 4). In this latter situation, the genetic diversity within the population is increased, as well as the gametic disequilibrium, making the estimation procedure more efficient.

However, we show that the bias and variance of effective population size estimates can be substantially reduced when $\hat{F}$ and $\hat{\eta}'_S$ are estimated over 8 loci or more in the infinite allele model, or 12 loci or more in $K$ allele model (Figure 3, Table 4). Increasing the number of allelic states has also been shown to improve the estimation. We thus recommend using a large number of highly polymorphic loci for this method to be reliable. With the advances of molecular techniques in the last decade, this recommendation (using at least 8 highly polymorphic loci) is not unrealistic. Indeed, it is now common practice to work with at least 8–10 highly polymorphic loci, such as microsatellite markers. In all the results that we presented, the sample size was representative of those used in empirical studies found in the literature (WILLIAMSON and SLATKIN 1999). Of course, the joint estimates may also be improved with larger sample sizes. As we have shown, the reliance of our method on the infinite island model of population structure is not tremendous (Table 5, Figure 5). Although the mutation rates we used may seem to be high, they fall in the range of realistic values for microsatellite markers (ESTOUP and ANGERS 1998). When using unlinked loci, our method seems to give reasonably robust estimates even when recent changes in population structure or size have arisen (Figure 6). This may not be true, however, if one intends to use closely linked loci (to estimate larger effective sizes, for example).

We did not directly compare the results obtained with this method to other tentative estimations. The main reason is that alternative methods do not always estimate the same quantities. Another reason is that the number of required samples may be different. Variance of effective population size has been tentatively estimated from temporal changes in allele frequency (NEI and TAJIMA 1981; POLLAK 1983; WAPLES 1989). These estimates also exhibited high variance (WAPLES 1989). WILLIAMSON and SLATKIN (1999) recently proposed a maximum-likelihood method to estimate effective population size from temporal change in allele frequencies. This improvement of the method, although having smaller bias and lower variance, still exhibited a large variance. However, WILLIAMSON and SLATKIN (1999) stated that the maximum-likelihood estimate of population size from temporal change in allele frequency should be difficult to compute exactly when there are more than two alleles per locus. Maximum-likelihood methods for highly polymorphic loci could be implemented using Monte Carlo Markov chain computations.

PUDOVKIN *et al.* (1996) provided an estimate of the effective number of breeders in a population from the observed excess of heterozygotes in the progeny relative to Hardy-Weinberg expected proportions in the base population. Any heterozygote excess relative to the expected distribution for random mating populations results from a difference in allele frequencies among different gamete pools. Indeed, in finite diploid populations, a difference in the allele frequencies between male and female uniting gametes is expected from the sampling error due to the finite size of male and female gamete pools and from the difference in allele frequencies among male and female parents (WANG 1996). LUIKART and CORNUET (1999) further evaluated the accuracy and precision of this method. This procedure provides reliable estimates for very small numbers of breeders, but is sensitive to the mating system. Moreover, this method does not hold for consanguineous reproductive systems.

Effective population size has also been evaluated from

the measure of the variance of the correlation of allele frequencies between pairs of loci (Hill 1981; Waples 1991; Bartley *et al.* 1992). In an infinite isolated random mating population at mutation-drift equilibrium, the correlation of (neutral) allele frequencies at a pair of loci is zero. In a finite population, however, genetic drift causes the allele frequencies at independent loci to be correlated. Moreover, the magnitude of this correlation depends on the effective population size (Weir and Hill 1980; Hill 1981; Waples 1991). This method gave discouraging results and has not been evaluated through simulations.

Finally, none of these methods allow for migration or population subdivision. They assume single isolated populations that do not receive any migrant individuals from elsewhere. If this hypothesis were false, estimates of effective population size would be biased upward. Moreover, all previous attempts to estimate gene flow could not untangle the effect of local drift from immigration.

In conclusion, we address the issue of obtaining a confidence interval in practical cases. Resampling methods have their drawbacks. For example, the bootstrap does not seem to be adapted to correctly estimating second moments statistics. Estimates of one- and two-locus identity probabilities, which are second and fourth moments of allelic frequency, are obtained through the comparison of pairs of haplotypes in a sample. Random sampling with replacement within a sample increases the probability of sampling the same individuals twice. Therefore, bootstrapping over individuals indubitably overestimates the measures of identity among pairs of individuals. So, we suggest obtaining a confidence interval by means of stochastic simulations. Since we suppose the population to be at migration-drift equilibrium, we can simulate the stochastic process of dispersal and reproduction with estimated values of $N_e$ and $m$ used as input parameters. The distribution of new estimates could then be used to build a confidence interval.

We thank K. Dawson, G. Luikart, O. Hardy, I. Olivieri, and C. Garza for helpful comments on a previous draft of this manuscript. We thank M. Slatkin and two anonymous reviewers for constructive criticism of this manuscript. R.V. acknowledges financial support from the Training and Mobility of Researchers programme "FRAGLAND" of the European Commission, coordinated by I. Hanski, from contract number BIO4-CT96-1189 of the Commission of the European Communities (DG XII), and from the Fondation Sansouire. This is contribution 2000-099 of the Institut des Sciences de l'Évolution de Montpellier.

## LITERATURE CITED

Avery, P. J., and W. G. Hill, 1979 Variance in quantitative traits due to linked dominant genes and variance in heterozygosity in small populations. Genetics **91:** 817–844.

Bartley, D., M. Bagley, G. Gall and B. Bentley, 1992 Use of linkage disequilibrium data to estimate effective size of hatchery and natural fish populations. Conserv. Biol. **6:** 365–375.

Beerli, P., and J. Felsenstein, 1999 Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. Genetics **152:** 763–773.

Caballero, A., 1994 Developments in the prediction of effective population size. Heredity **73:** 657–679.

Cockerham, C. C., 1984 Additive by additive variance with inbreeding and linkage. Genetics **108:** 487–500.

Cockerham, C. C., and B. S. Weir, 1977 Digenic descent measures for finite populations. Genet. Res. **30:** 121–147.

Cockerham, C. C., and B. S. Weir, 1987 Correlations, descent measures: drift with migration and mutation. Proc. Natl. Acad. Sci. USA **84:** 8512–8514.

Cockerham, C. C., and B. S. Weir, 1993 Estimation of gene flow from *F*-statistics. Evolution **47:** 855–863.

Crow, J. F., and K. Aoki, 1984 Group selection for a polygenic behavioural trait: estimating the degree of population subdivision. Proc. Natl. Acad. Sci. USA **81:** 6073–6077.

Crow, J. F., and C. Denniston, 1988 Inbreeding and variance effective population numbers. Evolution **42:** 482–495.

Crow, J. F., and M. Kimura, 1970 *An Introduction to Population Genetics Theory*. Burgess Publishing, Minneapolis, MN.

Estoup, A., and B. Angers, 1998 Microsatellites and minisatellites for molecular ecology: theoretical and empirical considerations, pp. 55–86 in *Advances in Molecular Ecology*, edited by G. R. Carvalho. IOS Press, Amsterdam.

Ewens, W. J., 1979 *Mathematical Population Genetics*. Springer-Verlag, Berlin.

Ewens, W. J., 1982 On the concept of effective population size. Theor. Popul. Biol. **21:** 373–378.

Fisher, R. A., 1930 *The Genetical Theory of Natural Selection*. Clarendon Press, London.

Gabriel, W., and R. Bürger, 1994 Extinction risk by mutational meltdown: synergistic effects between population regulation and genetic drift, pp. 69–84 in *Conservation Genetics*, edited by V. Loeschcke, J. Tomiuk and S. K. Jain. Birkhäuser Verlag, Basel.

Goodnight, C. J., 1987 On the effect of founder events on epistatic genetic variance. Evolution **41:** 80–91.

Goodnight, C. J., 1988 Epistasis and the effect of founder events on the additive genetic variance. Evolution **42:** 441–454.

Hanski, I., and M. E. Gilpin, 1997 *Metapopulation Biology: Ecology, Genetics, and Evolution*. Academic Press, San Diego.

Hedrick, P. W., 1987 Gametic disequilibrium: proceed with caution. Genetics **117:** 331–341.

Hill, W. G., 1981 Estimation of effective population size from data on linkage disequilibrium. Genetics **38:** 200–216.

Hill, W. G., and B. S. Weir, 1994 Maximum-likelihood estimation of gene location by linkage disequilibrium. Am. J. Hum. Genet. **54:** 705–714.

Lande, R., and B. S. Weir, 1994 Risk of population extinction from fixation of new deleterious mutations. Evolution **48:** 1460–1469.

Lewontin, R. C., 1988 On measures of gametic disequilibrium. Genetics **120:** 849–852.

Luikart, G., and J. M. Cornuet, 1999 Estimating the effective number of breeders from heterozygote excess in progeny. Genetics **151:** 1211–1216.

Lynch, M., and W. Gabriel, 1990 Mutation load and the survival of small populations. Evolution **44:** 1725–1737.

Lynch, M., J. Conery and R. Bürger, 1995a Mutation accumulation and the extinction of small populations. Am. Nat. **146:** 489–518.

Lynch, M., J. Conery and R. Bürger, 1995b Mutational meltdowns in sexual populations. Evolution **49:** 1067–1080.

Malécot, G., 1948 *Les Mathématiques de l'Hérédité*. Masson, Paris.

Malécot, G., 1975 Heterozygosity and relationship in regularly subdivided populations. Theor. Popul. Biol. **24:** 268–294.

Maynard Smith, J., and R. Hoekstra, 1980 Polymorphism in a varied environment: How robust are the models? Genet. Res. **35:** 45–57.

Nei, M., and F. Tajima, 1981 Genetic drift and estimation of effective population size. Genetics **98:** 625–640.

Nunney, L., and D. R. Elam, 1994 Estimating effective size of conserved populations. Conserv. Biol. **8:** 175–184.

Ohta, T., 1980 Linkage disequilibrium between amino acids sites in immunoglobulin genes and other multigene families. Genet. Res. **36:** 181–197.

Ohta, T., 1982a Linkage disequilibrium due to random genetic drift in finite subdivided populations. Proc. Natl. Acad. Sci. USA **79:** 1940–1944.

Ohta, T., 1982b Linkage disequilibrium with the island model. Genetics **101:** 139–155.

Pollak, E., 1983 A new method for estimating the effective population size from allele frequency changes. Genetics **104:** 531–548.

Pudovkin, A. I., D. V. Zaykin and D. Hedgecock, 1996 On the potential for estimating the effective number of breeders from heterozygote-excess in progeny. Genetics **144:** 383–387.

Reynolds, J., B. S. Weir and C. C. Cockerham, 1983 Estimation of the coancestry coefficient: basis for a short term genetic distance. Genetics **105:** 767–779.

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

Schwartz, M. K., D. A. Tallmon and G. Luikart, 1999 Using genetics to estimate the size of wild populations: many methods, much potential, uncertain utility. Anim. Conserv. **2:** 321–323.

Slatkin, M., 1985 Rare alleles as indicators of gene flow. Evolution **39:** 53–65.

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

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

Slatkin, M., and N. H. Barton, 1989 A comparison of three indirect methods for estimating average levels of gene flow. Evolution **43:** 1349–1368.

Slatkin, M., and W. P. Maddison, 1989 A cladistic measure of gene flow inferred from the phylogeny of alleles. Genetics **123:** 603–613.

Tachida, H., and C. C. Cockerham, 1986 Analysis of linkage disequilibrium in an island model. Theor. Popul. Biol. **29:** 161–197.

Tufto, J., S. Engen and K. Hindar, 1996 Inferring patterns of migration from gene frequencies under equilibrium conditions. Genetics **144:** 1911–1921.

van Noordwijk, A. J., 1994 The interaction of inbreeding depression and environmental stochasticity in the risk of extinction of small populations, pp. 131–146 in *Conservation Genetics*, edited by V. Loeschcke, J. Tomiuk and S. K. Jain. Birkhäuser Verlag, Basel.

Vitalis, R., and D. Couvet, 2001 Two-locus identity probabilities and identity disequilibrium in a partially selfing subdivided population. Genet. Res. (in press).

Vucetich, J. A., T. A. Waite and L. Nunney, 1997 Fluctuating population size and the ratio of effective to census population size. Evolution **51:** 2017–2021.

Wang, J., 1996 Deviation from Hardy-Weinberg proportions in finite populations. Genet. Res. **68:** 249–257.

Waples, R. S., 1989 A generalized approach for estimating effective population size from temporal changes in allele frequency. Genetics **121:** 379–391.

Waples, R. S., 1991 Genetic methods for estimating the effective size of cetacean populations, pp. 279–300 in *Genetic Ecology of Whales and Dolphins*, Special Issue 13, edited by A. R. Hoelzel. International Whale Commission, London.

Weir, B. S., and C. C. Cockerham, 1969 Group inbreeding with two linked loci. Genetics **63:** 711–742.

Weir, B. S., and C. C. Cockerham, 1984 Estimating *F*-statistics for the analysis of population structure. Evolution **38:** 1358–1370.

Weir, B. S., and W. G. Hill, 1980 Effect of mating structure on variation in linkage disequilibrium. Genetics **93:** 477–488.

Whitlock, M. C., and N. H. Barton, 1997 The effective size of a subdivided population. Genetics **146:** 427–441.

Whitlock, M. C., P. C. Phillips and M. J. Wade, 1993 Gene interaction affects the additive genetic variance in subdivided populations with migration and extinction. Evolution **47:** 1758–1769.

Williamson, E. G., and M. Slatkin, 1999 Using maximum likelihood to estimate size from temporal changes in allele frequencies. Genetics **152:** 755–761.

Wright, S., 1931 Evolution in mendelian populations. Genetics **16:** 97–159.

Wright, S., 1940 Breeding structure of population in relation to speciation. Am. Nat. **74:** 232–248.

Communicating editor: M. Slatkin