Likelihood and Approximate Likelihood Analyses of Genetic Structure in a Linear Habitat: Performance and Robustness to Model Mis-Specification

Francois Rousset, Raphaël Leblois

▶ **To cite this version:**

Francois Rousset, Raphaël Leblois. Likelihood and Approximate Likelihood Analyses of Genetic Structure in a Linear Habitat: Performance and Robustness to Model Mis-Specification. Molecular Biology and Evolution, 2007, 24 (12), pp.2730-2745. 10.1093/molbev/msm206 . halsde-00321541

HAL Id: halsde-00321541
https://hal.science/halsde-00321541v1

Submitted on 16 Jun 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

Likelihood and Approximate Likelihood Analyses of Genetic Structure in a Linear Habitat: Performance and Robustness to Model Mis-Specification

François Rousset and Raphaël Leblois†*

*Université, Montpellier 2, CNRS, Institut des Sciences de l'Évolution, France; and †Unité Origine, Structure et Évolution de la Biodiversité, Museum National d'Histoire Naturelle, Paris, France

We evaluate the performance of maximum likelihood (ML) analysis of allele frequency data in a linear array of populations. The parameters are a mutation rate and either the dispersal rate in a stepping stone model or a dispersal rate and a scale parameter in a geometric dispersal model. An approximate procedure known as maximum product of approximate conditional (PAC) likelihood is found to perform as well as ML. Mis-specification biases may occur because the importance sampling algorithm is formally defined in term of mutation and migration rates scaled by the total size of the population, and this size may differ widely in the statistical model and in reality. As could be expected, ML generally performs well when the statistical model is correctly specified. Otherwise, mutation rate estimates are much closer to mutation probability scaled by number of demes in the statistical model than scaled by number of demes in reality when mutation probability is high and dispersal is most limited. This mis-specification bias actually has practical benefits. However, opposite results are found in opposite conditions. Migration rate estimates show roughly similar trends, but they may not always be easily interpreted as low-bias estimates of dispersal rate under any scaling. Estimation of the dispersal scale parameter is also affected by mis-specification of the number of demes, and the different biases compensate each other in such a way that good estimation of the so-called neighborhood size (or more precisely the product of population density and mean-squared parent–offspring dispersal distance) is achieved. Results congruent with these findings are found in an application to a damselfly data set.

Despite lasting efforts, estimating dispersal rates from genetic data remain a challenging problem. Many uncertainties remain about the various complicating factors that may invalidate inferences. It is not clear how many parameters can be estimated accurately and whether the results will be robust to various factors such as the mode of evolution of the markers, ancestral history of the species, and populations unaccounted for in the statistical model (e.g., Slatkin 1994; Arbogast et al. 2002; Rousset 2007 for reviews).

Nevertheless, in spatially subdivided populations, some statistical patterns depend mainly on the recent history of the population. This makes it possible to develop statistical methods that specifically exploit these patterns, and therefore could be robust to various uncontrolled factors (e.g., Slatkin 1993, 1994). For example, previous works on moment-based methods (i.e., methods based on Wright's F_{ST} and similar measures) have shown that reliable estimation of some dispersal parameters is possible under isolation by distance because such estimation may be based on genetic patterns independent of unsampled populations, of mutation models, and robust to past demographic fluctuations (Slatkin 1993; Rousset 1997; Leblois et al. 2004).

On the other hand, moment methods may throw out too much of the information in the data. Much recent efforts have been focused on developing maximum likelihood (ML) methods (Beerli and Felsenstein 1999, 2001; Bahlo and Griffiths 2000; Stephens and Donnelly 2000; Beerli 2004; de Iorio and Griffiths 2004b; de Iorio et al. 2005), which in principle use more information in the data than moment methods. Although ML methods could allow to estimate more parameters and to estimate them more accurately, the same general robustness concerns arise as for any other method. The effect of populations that are connected by dispersal to the sampled ones, but that are not accounted for in the statistical model, has received some attention (Beerli 2004; Slatkin 2005). Beerli (2004) investigated the effect of a third population on the estimation of dispersal between 2 populations, for sequence data (100,000 bp per individual sampled). Slatkin (2005) considered predicting the magnitude of these effects from a simple algebraic argument based on expected coalescence times of pairs of genes. This argument predicts a bias when there is some estimator bias in Beerli's simulations, but the predicted bias is an overestimate of the observed bias. Slatkin also notes that the method cannot be applied to all possible dispersal patterns and sampling designs, in particular in a linear array as will be considered below.

So far, performance of ML methods has been analyzed in models with up to 4 demes (Beerli and Felsenstein 2001; Beerli 2006). The ultimate aim of the present work is the application and assessment of ML methods in much larger networks of subpopulations. The type of data considered here are allelic counts at high mutation rate loci such as many microsatellites.

Estimation is expected to be less precise when the number of parameters increases, and this effect is already apparent in a 4-demes model (Beerli 2006). We therefore focus on dispersal models with few parameters, such as the stepping stone/isolation by distance models on a homogeneous lattice, rather than the whole migration matrix approach implemented, for example, in Migrate (Beerli and Felsenstein 1999, 2001). We have implemented the algorithm of de Iorio and Griffiths (2004a, 2004b) to handle the case of localized dispersal (isolation by distance) in a linear habitat. This scenario has been chosen because it is a relatively simple starting point for a larger simulation project, yet it is realistic enough to have allowed reasonably accurate statistical analyses on real data sets. Although the isolation by distance models neglect spatial heterogeneities, these do not appear to be a major concern in a number

Key words: dispersal, maximum likelihood, coalescence, isolation by distance, microsatellites.

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

Mol. Biol. Evol. 24(12):2730–2745. 2007
doi:10.1093/molbev/msm206
Advance Access publication September 24, 2007

© The Author 2007. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. All rights reserved.
For permissions, please e-mail: journals.permissions@oxfordjournals.org

Downloaded from https://academic.oup.com/mbe/article/24/12/2730/978152 by Bibliothèque Universitaire de médecine - Nîmes user on 16 June 2021

of applications, for example, allowing good estimation of "neighborhood size" by nonlikelihood methods (Rousset 1997, 2000; Sumner et al. 2001; Fenster et al. 2003; Watts et al. 2007), and a similar result will be achieved here using the data of Watts et al. (2007). We attempt to analyze samples from large sets of subpopulations, not only because of the problem of unaccounted populations but because, as shown in these references, many natural populations may be described as a large network of small subpopulations connected by a large amount of dispersal, even up to the point where no subpopulations are distinguished from individuals or mating pairs ("continuous" populations). Moreover, the same works confirm the theoretical expectation that such conditions are favorable to the reliable estimation of dispersal rates.

We will see that although ML estimation under models of 10–15 populations is easy based on the algorithms of de Iorio and Griffiths, it becomes progressively more difficult as the number of subpopulations increases, and analysis of an average data set would require weeks on most desk computers when more than 40 subpopulations are considered. Hence, after a check of the method and assessment of numerical factors that may affect the precision of the estimates in simple conditions, we will consider the effects of unaccounted subpopulations on the analyses and a fast approximation to ML.

Under a nearest neighbor stepping stone model, unaccounted populations will be found to have little effect, but if dispersal distance follows a geometric distribution, stronger mis-specification effects will be obtained. Irrespective of mis-specification, the shape of the dispersal distribution will appear difficult to estimate. We will also test less thoroughly the effect of some other deviations from the model, which will appear to have less impact on performance.

A fast heuristic approximation, product of approximate conditional (PAC) likelihood (Li and Stephens 2003; Cornuet and Beaumont 2007), will be shown to yield results very close to those based on likelihood itself and will allow a more thorough investigation of possible causes of poor performance as well as of a wider range of parameter values, in particular higher dispersal among smaller demes, and lower mutation rates.

Methods

Design of Simulation Study

Population Models

As a first approximation, we may consider many species as collections of clusters of subpopulations (or of "demes") with abundant dispersal within each cluster and relatively much less dispersal among clusters. We consider the analysis of one such cluster. Typical values for the biological scenarios envisioned here would be deme size ≈1–100, dispersal probability ≈0.5, and therefore expected number of immigrants ≈0.5–50 per deme. In order to maintain enough genetic variability within the total population, we must also consider a large array of demes and/or large deme size and small migration rates.

These different requirements somehow conflict with each other and with constraints on computation times. Thus, we first consider large haploid deme size (400), small dispersal probability (0.01), and high mutation probability (10^{-3} per gene copy per generation), so that we can check performance in a small network of populations; then we will take benefit of the fast PAC likelihood method and will increase lattice size, reduce deme size, increase dispersal, and decrease mutation probability. In all cases the probability of dispersal to signed distance $k \neq 0$ can be described as

$$\frac{m}{2}(1-g)g^{|k-1|}, \tag{1}$$

for given g. g is thus a shape parameter which describes dispersal distances. The stepping stone model is the limit case $g \to 0$.

Some effects of dispersal on population processes, such as cline shape, are well quantified by the axial mean-squared parent–offspring distance, σ^2 (e.g., (Barton and Gale 1993). In the geometric dispersal model,

$$\sigma^2 = \frac{m(1+g)}{(1-g)^2}. \tag{2}$$

The product $D\sigma^2$, where D is population density, also determines spatial variation in the probability of identity of genes (isolation by distance: Sawyer 1977 for the most accurate results). Although different views have been held about the reasonable magnitude of σ^2 and $D\sigma^2$, these parameters can be low in natural populations. In such organisms, as *Dipodomys* rodents (Rousset 2000; Winters and Waser 2003), humans in the rainforest (Wood et al. 1985; Rousset 1997), *Chamaecrista fasciculata* (Fabaceae; (Fenster et al. 2003), American marten (Broquet et al. 2006), and *Gnypetoscincus queenslandiae* skinks (Sumner et al. 2001), concurrent genetic and demographic estimates of $D\sigma^2$ were $2.5 \leq \cdot \leq 40$, and such is σ^2 when measured in unit of interindividual distance (such that $D = 1$). In the latter units, σ^2 will be $7.5 \leq \cdot \leq 840$ in our simulations. Note that in the linear habitats considered in this work, $D\sigma^2$ values cannot be compared unless they are measured in the same spatial unit because density scales as distance^{-1} hence $D\sigma^2$ scales as distance. In this work, it will be reported as $N\sigma^2$, that is, in units of array step (except for the actual data analysis).

The assumed mutation probability is 10^{-3} or 10^{-4}, which is not unrealistic for microsatellite markers (reviewed in Ellegren 2000; see also e.g., Vigouroux et al. 2002; Gusmão et al. 2005). At high mutation rate loci, the allelic type of rare immigrants from distant populations should be uncorrelated to that of resident individuals, so the mutation events can also represent immigration from distant clusters of populations (Kimura and Weiss 1964). Only the 10^{-3} mutation probability will be considered in the smallest populations simulated as it is required to maintain substantial variation. Both mutations rates will be considered in larger populations, where this 10-fold variation in mutation probability will have notable consequences for the interpretation of the results.

The K-allele mutation model will be assumed in the data-generating simulations, except for a few cases where

Downloaded from https://academic.oup.com/mbe/article/24/12/2730/978152 by Bibliothèque Universitaire de médecine - Nîmes user on 16 June 2021

Table 1
Notation

Population (data-generating) model	
N	Deme size (gene copies or haploid individuals)
σ^2	Mean-squared parent–offspring distance
n_d	Number of demes
$v_{\alpha\beta}$	Dispersal probability between demes α and β in de Iorio and Griffiths (2004b)
m	Total dispersal probability (this work)
Additional parameters of statistical model	
n_m	Number of demes in statistical model
N_T	Total haploid size in statistical model (N in de Iorio and Griffiths, 2004b)
N_e	Effective size in statistical model (haploid equivalent)
$\theta \equiv 2N_T\mu$	Scaled mutation rate
$m_{\alpha\beta} \equiv 2N_T v_{\alpha\beta}$	Scaled dispersal rate in between demes α and β in de Iorio and Griffiths (2004b)
Numerical parameters	
n_t	Number of ancestral trees (IS algorithm) or sequences (PAC likelihood algorithm)
n_p	Number of parameter points in which likelihood or other statistics are computed

a 10-allele–bounded stepwise mutation model (SMM) will be considered in order to test the robustness of the analyses when the marker mutational process deviates from the one assumed in the statistical model.

Statistical Models

In its most general form the statistical model considered here allows estimation of 3 parameters: a mutation rate, a migration rate (scaled probability of immigration), and the g parameter describing the geometric distribution of dispersal distances. We also investigated the performance of the estimator obtained by plugging the ML estimates of Nm and g in the parametric expression for $N\sigma^2$ in terms of these parameters (eq. 2).

Most simulations assume localized dispersal on a linear array of populations, with absorbing boundaries, much as in the population models under which samples are simulated. However, in practice this either assumes that the positions of the sampled demes in the linear array are known or this forces the user to make assumptions about this position. Hence, estimation under a circular lattice model will also be considered.

Sampling Design

We assume that samples are taken as follows: in the 4-demes model, in each deme; in the 100-demes models, at positions $50, 52, \ldots, 50 + 2(n_s - 1)$ where n_s is the number of demes sampled; in the 1,000-demes models, at positions $500, 502, \ldots, 500 + 2(n_s - 1)$ for $n_s = 4$ or 10 and at positions $500, \ldots, 519$ for $n_s = 20$.

The Algorithms and their Implementation

Notation

Some notation is summarized in table 1. Note that de Iorio and Griffiths (2004a, 2004b) denote N, the total size of the population, which we here denote N_T.

Computation of the Likelihood

The detailed features of the algorithm have been described in de Iorio and Griffiths (2004a, 2004b) and are not repeated here, although some guidance is given. In this section, their notation is followed, unless indicated otherwise. Their algorithm computes likelihood under the structured coalescent models described by Notohara (1990) and Herbots (1997), which are limit processes, for large deme size and low migration rates, of the classical migration matrix models (e.g., Nagylaki 1983; Rousset 2004, p. 54 sqq). In these algorithms, one considers an absorbing Markov chain over the state $\mathbf{n}$ of the set of ancestral lineages of a sample of genes from the time of sampling up to the most recent common ancestor. $\mathbf{n}$ is characterized by the allelic type and the geographic position of the lineages. A sample can be represented as the sequential addition from 0 to n genes, where the probability that any additional gene is of a given type will depend on the state of genes already present. The likelihood can thus be written in terms of any given sequence of gene states s_l (allelic type and geographic position) leading to the observed sample, as

$$\binom{n}{\mathbf{n}} \prod_{l=1}^{l=n} \pi(s_l|\mathbf{n}_{l-1}), \tag{3}$$

where $\pi(s_l|\mathbf{n}_{l-1})$ is the probability that an additional sampled gene s_l is of a given type, given the configuration $\mathbf{n}_{l-1}$ already generated by the sequence, and $\binom{n}{\mathbf{n}}$ is the multinomial coefficient in terms of the allelic counts $\mathbf{n}$ (de Iorio et al. 2005). de Iorio and Griffiths define an importance sampling (IS) algorithm considering the successive events (mutation, migration or coalescence) that may affect the ancestral lineages of the sample. $\pi(\cdot)$ terms can be defined for migration and mutation events from the above ones (e.g., the π for a migration event leading from some configuration "$\mathbf{n}$ + migrating gene in deme 1" to "$\mathbf{n}$ + migrating gene in deme 2" is defined from the ratio of the $\pi(\cdot|\mathbf{n})$ for addition of the migrating allele to deme 1 and of the $\pi(\cdot|\mathbf{n})$ for its addition to deme 2). The IS algorithm is defined in terms of approximations $\hat{\pi}$ to the π's. If $\hat{\pi}=\pi$, one iteration of this algorithm (i.e., one ancestral history) is enough to compute the likelihood. In general, the π's cannot be computed exactly so that $\hat{\pi} \neq \pi$. In this case, the IS algorithm may still allow consistent estimation of the likelihood, and fewer iterations

Downloaded from https://academic.oup.com/mbe/article/24/12/2730/978152 by Bibliothèque Universitaire de médecine - Nîmes user on 16 June 2021

of this algorithm should be needed the closer the $\hat{\pi}$'s are to the π's. Poor choice of $\hat{\pi}$ may result in inefficient estimation of the likelihood (requiring too many iterations of IS for practical applications) or even in inconsistent estimation (Stephens and Donnelly 2000). To overcome these limitations, de Iorio and Griffiths proposed to use the following $\hat{\pi}$. Denote $\hat{\pi}(j|\alpha, \mathbf{n})$ the coefficients considered when a lineage of allelic type j in deme α is affected by some event. They are obtained as solutions of linear equations of the form:

$$[n_\alpha q_\alpha^{-1} + m_\alpha + \theta]\hat{\pi}(j|\alpha, \mathbf{n}) = n_{\alpha j} q_\alpha^{-1} + \theta \sum_i P_{ij} \hat{\pi}(i|\alpha, \mathbf{n}) + \sum_{\beta \neq \alpha} m_{\alpha\beta} \hat{\pi}(j|\beta, \mathbf{n}) \quad (4)$$

(de Iorio and Griffiths 2004b, eq. 2.11) for each j and α. Here q_α is the deme size relative to the total population size, and (P_{ij}) is a matrix of relative forward mutation rates. In the present work, we assume a 1-dimensional lattice, with n_d demes of equal size, so that $q_\alpha = 1/n_d$.

Solving a system of Z linear equations typically requires approximately $O(Z^3)$ computations (e.g., Press et al. 1988; Golub and van Loan 1996). The $\hat{\pi}(j|\alpha, \mathbf{n})$ are the solutions of a system of Kn_m equations of the form (4) and most of the computation time is spent solving such systems of equations. This is the limiting step in considering scenarios with large numbers of alleles or of demes. Ways of dealing with a large number of alleles are discussed below. The increase in computation time with the number of demes actually scales higher than n_m^3 because the number of events in the history of a sample increases as n_m increases. Iterative methods of solution of linear systems of equations can speed up the computations with a negligible loss of accuracy when compared with the direct solvers. A preconditioned conjugate gradient method (e.g., Golub and van Loan 1996) was found useful for $n_m \geq 60$ in this study.

Computation time can be substantially reduced if the above system of equations can be broken down in disjunct subsystems of equations. This occurs in particular in the symmetric K-allele model (KAM) that was the only model assumed in the estimation procedure in this work. In the KAM, $P_{ij} = 1/K$ (for $i = j$ included if we follow de Iorio and Griffiths' convention). $\sum_i \hat{\pi}(i|\alpha, \mathbf{n}) = 1$, so that the mutation term on the right-hand side of equation (4) simplifies: for each allele type j, the recursion (4) can be written as

$$[n_\alpha q_\alpha^{-1} + m_\alpha + \theta]\hat{\pi}(j|\alpha, \mathbf{n}) - \sum_{\beta \neq \alpha} m_{\alpha\beta} \hat{\pi}(j|\beta, \mathbf{n}) = n_{\alpha j} q_\alpha^{-1} + \theta/K. \quad (5)$$

Hence, for each allele j, the $\hat{\pi}(j|\alpha, \mathbf{n})$ are obtained as a solution of the system of n_d linear equations for $\alpha = 1, \ldots, n_d$. The system of Kn_d equations separates in K disjunct systems of n_d equations, only one of which is solved once a given allelic type has been chosen. Systems of n_d linear equations also arise for more complex mutation models if the mutation and genealogical processes are independent (as is usually assumed for neutral genetic variation): for each right eigenvector $\mathbf{r}_k \equiv (r_{kj})$ of (P_{ij}), one has to deduce equations for $\sum_j r_{kj} \hat{\pi}(j|\alpha, \mathbf{n})$ from equation (4). For each eigenvector $\mathbf{r}_k$, there are n_d such equations. Solutions of such a system can then be back transformed to obtain solutions of equation (4). This procedure is illustrated for an unbounded SMM, where it amounts to Fourier analysis, in de Iorio et al. (2005), but it increases computation time in comparison to the KAM and was not considered here. Instead, the performance of KAM-based estimation on data following a bounded SMM will be presented.

Another potential solution to reduce the computation time is bridge sampling (Meng and Wong 1996; Fearnhead and Donnelly 2001), in which the proposal distributions (hence the $\hat{\pi}$'s) are not computed independently for each parameter point, but only for a few driving values. There could be a trade-off between the cost of computing the $\hat{\pi}$'s for each point and the potential loss in efficiency of likelihood estimation when suboptimal proposal distributions are used, and the efficiency of de Iorio and Griffiths' proposal distribution has initially drawn us away from methods such as bridge sampling. However, this could prove useful in later applications.

The PAC Likelihood Heuristics

Cornuet and Beaumont (2007) proposed to use de Iorio and Griffiths's $\hat{\pi}$ directly as a substitute to π in equation (3) and to average over different sequences of genes leading to the sample (see also RoyChoudhury and Stephens 2007). This follows a similar suggestion by Li and Stephens (2003) who described this procedure as PAC likelihood and as maximum PAC likelihood, the procedure of maximizing this product with respect to parameters.

There is no general result showing that the PAC likelihood algorithm consistently estimates the likelihood. It clearly does so when $\hat{\pi} = \pi$, in which case simulation is not necessary. π is known in particular when the stationary joint distribution of allele frequencies in different demes is known. This occurs in the n-coalescent with parent-independent mutation (Stephens and Donnelly 2000; de Iorio and Griffiths 2004a) and can be extended to the island model with the same mutation model and a large number of islands. On the other hand, the distribution is unknown for stepwise mutation and/or for isolation by distance. Nevertheless, simulation results of Cornuet and Beaumont and RoyChoudhury and Stephens for stepwise mutation suggest that PAC likelihood may be used as a practical substitute to likelihood, and it is much faster to compute because the number of systems of linear equations to be solved for each sequence is the sample size, whereas in the IS algorithm, the number of linear systems will be increased beyond this in proportion to the number of mutation and migration events in an ancestral history.

As will be seen, the PAC likelihood statistic is not a consistent estimator of the likelihood. On the other hand, the variance of estimation of the PAC likelihood for a given number of iterations of the PAC likelihood algorithm is lower than the variance of estimation of likelihood from the same number of iterations of the IS algorithm, as was already observed by Cornuet and Beaumont and RoyChoudhury and Stephens. This reduced variance more

Downloaded from https://academic.oup.com/mbe/article/24/12/2730/978152 by Bibliothèque Universitaire de médecine - Nîmes user on 16 June 2021

than compensates for the small bias in estimating likelihood. So, even if maximizing likelihood is better, in terms of mean square error (MSE), than maximizing the expectation of PAC likelihood, the estimation of demographic parameters by maximum PAC likelihood may appear better than ML estimation when both the likelihood and the PAC likelihood are estimated with some error. We will indeed find that maximum PAC likelihood estimation is at least as good, if not slightly better than ML estimation by the IS algorithm.

Likelihood Surface and ML Estimation

The likelihood in any given parameter point is estimated with some error rather than computed. This prevents the straightforward application of most algorithms for finding the maximum of a function. A convenient way to address this problem is to interpolate the likelihood surface from the estimated points by predicting it under some probabilistic model for the shape of the surface. If the surface is assumed to be the realization of a Gaussian process, this prediction can be achieved by techniques known as kriging (e.g., Cressie 1993). Both prediction uncertainty and prediction bias, when the Gaussian assumption does not hold, are expected, but in most cases this appears to be a minor source of inaccuracy, as can be tested by increasing the density of points on which interpolation is based. We use kriging as in de Iorio et al. (2005; see also Sacks et al. 1989; Welch et al. 1992): for each point in parameter space, the likelihood is estimated by simulations of n_t trees. This is repeated at n_p points in parameter space. Kriging fits a surface to the estimated likelihood values in the different parameter points. This is an estimate of the likelihood surface, of which the maximum may be sought by any of the usual algorithms. We used the package fields (Fields Development Team 2006) in the R statistical environment (R Development Core Team 2004) for the kriging computations. The Nelder–Mead algorithm as implemented in the R optim function was used to find the maximum of the estimated likelihood surface.

As in de Iorio et al. (2005), points are selected by Latin hypercube sampling (a form of stratified random sampling). The number of points is adjusted as function of time constraints and efficient use of hypercube sampling. The range of parameter space explored has to be provided by the user. In general, one should first explore a wide parameter space then focus the search around the first estimate obtained. Here, preliminary work (not shown) helped select parameter ranges within which all estimates would be found, and only results for these ranges are presented because they give the relevant information about the performance of ML estimation per se. Alternatively, a more automated iterative procedure has been used where a wide parameter range is used in the first iteration and the parameter range used in the later iterations is narrowed around the previous estimate obtained for each sample. In particular, a 2-steps procedure has been used recurrently, where estimates were first deduced from 512 points; 512 additional points were sampled from an approximately 10-fold smaller parameter space around the first estimates for each sample, and final estimates are deduced from the 1024 points thus obtained.

Computer Implementation

The C++ program used in all data analyses will be distributed as a free software, MIGRAINE, available through URL http://kimura.univ-montp2.fr/~rousset/Migraine.htm. It writes the required R code and can call R interactively to perform the above iterative procedure. It has been run on PCs under Windows and Linux, a Sun workstation, SGI Origin 3800 and IBM Power4 parallel computers of the CINES (www.cines.fr), and several Linux PC clusters. Some representative computation times are given in table legends.

Programs Checks

The likelihood estimation procedure was checked against standard formulas for probability of identity of pairs of genes (e.g., Maruyama 1970a; Malécot 1975) adapted to the KAM (e.g., Crow and Aoki 1984; Rousset 2004) and taken in the limit $N \to \infty$ for $N\mu$ and Nm fixed as in the coalescent algorithm. The simulation program generating samples has been previously described (Leblois et al. 2003, 2004) and has been checked as described in these papers.

Comparison of Performance of Different Implementations

There are 2 sources of inaccuracy of estimates. One is the inaccuracy of the ML estimate relative to the parameter value. The other is the inaccuracy of the numerical method in locating the ML estimate, which may be due to considering not enough replicate trees per point in the IS computation or not enough points. It is possible to evaluate the inaccuracy due to the numerical method by comparing independent runs on the same data (de Iorio et al. 2005). However, it would have been too time consuming to do so in all cases. Rather, the impact of numerical settings on performance will be checked in several cases.

Distributions of estimators (or distributions of differences in cases of paired simulations) were compared primarily through the estimation of differences in MSE or relative MSE, and further by estimation of differences in bias and variance. Maximum differences consistent with the data were deduced from 95% confidence intervals (CIs) for effects on MSE, bias, and variance constructed by the "bootstrap corrected and accelerated" (BC_a) method of DiCiccio and Efron (1996). However, because it would be inconvenient to report all CIs, synthetic bounds on maximum absolute effect size and/or P values derived from the confidence curves are reported when they bring the main information together with estimates reported in the tables.

Results

Numbers in brackets refer to the numbered cases in the different tables. For the parameters Nm, $N\mu$, and $N\sigma^2$, we present relative bias and relative root MSE ($\sqrt{\text{MSE}}$) as this may be more important than absolute bias and MSE in practice. These relative error measures cannot apply for g (in particular in the nearest neighbor stepping stone model, $g = 0$), for which bias and MSE are directly computed.

Downloaded from https://academic.oup.com/mbe/article/24/12/2730/978152 by Bibliothèque Universitaire de médecine - Nimes user on 16 June 2021

Table 2
Performance of Estimation in a Nearest Neighbor Stepping Stone Model

	$N\mu$ relative bias (relative $\sqrt{\text{MSE}}$)	Nm relative bias (relative $\sqrt{\text{MSE}}$)
Linear array of 4 demes of 400 individuals		
$K = 4$		
[1]	0.04 (0.28)	0.11 (0.36)
[2] ($n_p = 5{,}000$)	0.04 (0.31)	0.11 (0.33)
[3] ($n_p = 5{,}000$, 10 loci)	−0.01 (0.22)	0.09 (0.25)
[4] ($n_p = 1{,}000$, 50 loci)	0.11 (0.19)	−0.03 (0.12)
[5] PAC	−0.002 (0.28)	0.09 (0.33)
$K = 10$		
[6]	0.52 (0.60)	−0.04 (0.25)
[7] ($n_t = 300$)	0.53 (0.61)	−0.08 (0.23)
[8] (50 loci)	0.47 (0.48)	−0.11 (0.13)
Linear array of 100 demes of 400 individuals		
$K = 4$		
[9] (n_p, n_t) = (512, 10)	0.33 (0.72)	0.10 (0.53)
[10] (n_p, n_t) = (512, 30)	0.27 (0.67)	0.12 (0.51)
[11][a] (n_p, n_t) = (5,000, 30)	0.25 (0.63)	0.16 (0.53)
[12] $n_m = 100$	0.30 (0.50)	−0.04 (0.32)
[13] PAC(n_p, n_t) = (512, 10)	0.26 (0.73)	0.06 (0.48)
[14] PAC(n_p, n_t) = (512, 30)	0.29 (0.75)	0.08 (0.49)
[15] PAC(n_p, n_t) = (512, 300)	0.28 (0.74)	0.07 (0.49)
[16] PAC(n_p, n_t) = (5,000, 30)	0.25 (0.69)	0.10 (0.51)
[17] PAC, $n_m = 100$	0.19 (0.42)	−0.12 (0.24)
$K = 10$		
[18]	0.40 (0.51)	−0.005 (0.29)
[19] Bounded SMM	−0.15 (0.25)	−0.06 (0.29)

NOTE.—Sixty samples were analyzed except 30 for case [4] and 120 for cases [9] and [18].

[a] Likelihood computations for case [11] took ≈84 min per sample on 2.66 GHz processors, whereas the otherwise identical PAC likelihood analysis took less than 7 min per sample.

Stepping Stone Dispersal

Numerical results are presented in table 2. For all cases in this table, $\mu = 0.001$ and $m = 0.01$ ($N\mu = 0.4$, $Nm = 4$). The following values apply to all cases unless noted otherwise: sample sizes were 5 loci, 4 demes sampled, and 60 genes sampled per deme; sampled ranges of parameter values were $2N\mu \in [0.125, 5]$ and $2Nm \in [0.125, 20]$; $n_m = 4$, $n_p = 512$, and $n_t = 10$.

The precision would be excellent for most practical purposes. For 5 different analyses of the same data sets [1]–[5] (4-allele model), the analysis with the largest number of loci (50, case [4]) stands out as the one with the lowest MSE for both estimates (as could be expected) as well as the lowest Nm bias, but also the highest $N\mu$ bias. For an identical total computation effort, reducing the number of loci and increasing the number of points analyzed is less efficient (cases [3] vs. [4], all $P < 0.039$ except for $N\mu$ MSE). Similar observations are made for a population of 100 demes. For the 10-allele model, MSEs and variances are likewise reduced after a 10-fold increase in the number of loci (case [8] vs. [6]); all $P < 0.006$).

For a population of 4 demes, a stronger bias and MSE in $N\mu$ estimation is observed for data generated under the 10-allele model (case [6]) than under the 4-allele model. For 100 demes, estimation is markedly improved in the 10-alleles model relative to the 4-alleles one (cases [18] vs. [9]). The latter observation is more in keeping with the frequent observation that highly polymorphic markers allow more powerful inferences, including about structured populations (power of tests of differentiation, Goudet et al. 1996; estimators of F_{ST}, Raufaste and Bonhomme 2000; assignment, Estoup et al. 1998 and Waples and Gaggiotti 2006; $D\sigma^2$ estimation, Leblois et al. 2003). However, the estimation biases are reduced by less than 15% (CI bound on bias reduction; $P = 0.023$ for Nm bias, 0.24 for $N\mu$). One reason for persistent $N\mu$ biases with increased sample size (most notably when the number of loci is increased) is that the observed number of alleles k in a 1-locus sample is often lower than K, so the program analyzes the data under a k-alleles model rather than under the correct 4- or 10-alleles model. This also readily explains the comparatively poor performance in the 4-demes, 10-alleles cases as some alleles are more likely to be absent in smaller populations. There is no obvious way to avoid the resulting biases, unless external information is provided by the user. Thus, this must be taken as an inherent bias of the method. Further, this will be less of a problem in later applications with larger total population sizes (so that k approaches K), so no attempt was done to correct for this problem in the analyses.

Beyond the number of loci, the number of parameter points considered may also set a limit to the precision that can be reached whatever the number of loci is. However, increasing the number of points from 512 to 5,000 (case [2] vs. [1]) reduces all relative measures of performance by at most 4.5% (upper CI bound, significant only for $N\mu$ MSE and variance). Finally, the performance of maximum PAC likelihood is at least as good as the ML analysis (case [5] vs. [1], all relative effects in favor of PAC likelihood and <0.065; $P = 0.003$ for $N\mu$ bias and >0.24 otherwise). Another test of a numerical factor (the number of trees sampled by the IS algorithm, case [7] vs. [6]) shows weak effect (for MSEs, at most a 6% reduction for Nm, $P > 0.22$, although this hides a bias-variance trade-off, with maximum absolute bound 7.8% and $P = 0.001$ on Nm bias).

Thus, the performance of estimation appears limited more by sample size than by numerical aspects of the algorithms and not worsened by the use of the PAC likelihood approximation. These 2 observations will recur in the sequel.

Estimation of Scaled Parameters under Mis-Specification

We now focus on the effect of unaccounted populations. We assume that unsampled demes in between the sampled ones are known and properly accounted for. Otherwise, serious mis-specification effects would occur, but these should be relatively easy to anticipate and/or avoid. By contrast, we will consider the less trivial effect of unaccounted populations outside the spatial range of sampled populations. Indeed, we will consider the effect of populations that hardly exchange any migrant directly with the sampled populations.

First, we should make clear which parameters are to be estimated when some demes are unaccounted. The coalescent algorithm is based on approximations in terms of scaled parameters, $N_T m$ and $N_T \mu$ in a stepping stone model, and is expected to perform well (in the sense of asymptotic

Downloaded from https://academic.oup.com/mbe/article/24/12/2730/978152 by Bibliothèque Universitaire de médecine - Nîmes user on 16 June 2021

efficiency, at least) when the statistical model and the true population structure match each other, that is, when $n_d = n_m$. But it is not obvious how it will perform when these do not match. Consider, for example, that 4 demes have been sampled out of a large population of $n_d = 100$ demes, and that only $n_m = 10$ demes are considered in the statistical model, so that likelihood is computed for different values of $Nn_m m$. Will ML estimates of $Nn_m m$ be close to $Nn_m m$, close to $N_T m = Nn_d m$, or show a more erratic behavior? In the latter case, the analysis could be useless. In the second case, it would not be possible to infer the dispersal probability m (if N is known) or the number of immigrants Nm, unless there is additional information about n_d. Indeed, n_d is often little more than a convenient abstraction as total population sizes fluctuate over the time span of coalescence of gene lineages. The inferences will therefore be most informative in the first case, when estimates approach $Nn_m m$, and likewise for $Nn_m\mu$ so that estimators of Nm and $N\mu$ can be deduced (and given moderate demographic information, N can be taken out).

This conclusion is a bit caricatural. One could argue that the most informative scaling for mutation and for migration differ from each other. However, for the parameter values of cases [11] and [18], the simulations indeed show that estimates of $Nn_d m$ and $Nn_d\mu$ approach $Nn_m m$ and $Nn_m\mu$. To make this clear, estimates of $Nn_d m$ and $Nn_d\mu$ will be both divided by the (necessarily known) n_m and the resulting values will be compared with Nm and $N\mu$ values in order to assess performance. When we take these values as the estimands, the mis-specification bias appears low (as long as $\mu = 10^{-3}$, as later simulations will emphasize). This turns a substantial mis-specification bias into a benefit of the method.

Under this interpretation, some mis-specification effects remain apparent by comparison with the correctly specified analysis (cases [12] vs. [9]; with both reduced variance and bias of Nm estimates and reduced variance of $N\mu$ estimates, all $P < 0.032$). The correctly specified PAC likelihood analysis also yields clearly better results than all incorrectly specified ML and PAC likelihood analyses (cases [17] vs. [9]–[16]). PAC likelihood estimates of Nm are less biased than ML ones ($P < 0.043$ in all 4 comparisons [9] vs. [13], [10] vs. [14], [11] vs. [16], and [17] vs. [12]), except for Nm in the latter case. However, effect sizes are <10% overall. Numerical settings again appear to be a comparatively minor source of error in estimation, the most notable effect being a reduction in variance of $N\mu$ estimates when a higher number of points is computed (CI for this reduction is 1.1–26.5% for case [9] vs. [11], less clear cut for case [13] vs. [16]). No further attempt was made to further sort out the diverse effects of model mis-specification, small sample size, and PAC likelihood approximation and their interactions as they all appear small.

Comparison of cases [18] and [9] also show, as in previous 10-alleles/4-alleles comparisons, a lower MSE and variance of estimators with 10 alleles than with 4 (all $P < 0.017$), yet the $N\mu$ bias is not reduced (CI for relative effect −0.07 to 0.19), which was explained as an effect of mis-specification of the number of alleles in the mutation model. Because this mis-specification should be less of a concern in sample's larger populations, and given our focus on microsatellite data, all further simulations in this paper are for $K = 10$ alleles.

Table 3
Performance of Estimation for Geometric Dispersal in a Linear Array of 100 Demes

	n_m	$N\mu$ relative bias (relative $\sqrt{\text{MSE}}$)	Nm relative bias (relative $\sqrt{\text{MSE}}$)	g bias ($\sqrt{\text{MSE}}$)
Samples of 5 loci				
$g = 0$				
[20]	10	0.42 (0.54)	−0.17 (0.30)	0.12 (0.14)
[21] PAC	—	0.26 (0.39)	−0.21 (0.28)	0.16 (0.21)
$g = 0.2$				
[22]	—	0.52 (0.65)	0.04 (0.34)	−0.01 (0.14)
[23] PAC	—	0.39 (0.49)	0.03 (0.34)	−0.02 (0.16)
[24]	25	0.30 (0.46)	0.08 (0.38)	−0.02 (0.18)
[25] PAC	40	0.04 (0.25)	0.01 (0.31)	−0.02 (0.18)
[26] PAC	100	0.05 (0.24)	0.001 (0.30)	−0.01 (0.18)
$g = 0.5$				
[27]	10	1.01 (1.13)	0.23 (0.40)	−0.15 (0.29)
[28] PAC	—	0.86 (0.97)	0.16 (0.33)	−0.13 (0.27)
[29]	16	0.77 (0.90)	0.21 (0.39)	−0.21 (0.32)
[30] PAC	—	0.68 (0.85)	0.13 (0.31)	−0.14 (0.27)
[31][a]	25	0.65 (0.84)	0.20 (0.42)	−0.18 (0.30)
[32] PAC	—	0.48 (0.68)	0.15 (0.35)	−0.16 (0.29)
[33] PAC	40	0.20 (0.61)	0.12 (0.25)	−0.12 (0.25)
[34] PAC	100	0.20 (0.61)	0.09 (0.27)	−0.12 (0.26)
$g = 0.5$, 20 loci				
[35] PAC	10	0.76 (0.79)	0.07 (0.17)	−0.07 (0.19)
[36] PAC	100	−0.02 (0.31)	0.07 (0.20)	−0.04 (0.20)

NOTE.—Thirty multilocus samples of 4 sampled demes and 60 genes sampled per deme were analyzed, except 60 samples for cases [20] and [27]–[32].

[a] Case [31] required about 24 h 30 min per sample on 2.66 GHz CPUs.

Effect of Mutation Model

We have considered a KAM for fast computation, but of course this may not be realistic. Because estimating parameters under a more general mutation model appears unpractical, we evaluated the impact of a bounded stepwise mutation process on the estimation procedure (case [19]). The total population is also mis-specified as in the KAM (case [18]). Compared with the latter, there is some reduction in MSE of $N\mu$ estimates due to an ≈55% change in mean value (this is highly significant due to the low variance of estimates, $P < 5.10^{-4}$). The dispersal estimates are robust to mis-specification of the mutation model. Similar results will be obtained for other data generated under the SMM.

Geometric Dispersal

So far only the nearest neighbor stepping stone model has been considered, and mis-specification of the number of demes seemed to have little effect. We now consider whether we can estimate the parameters of a more general dispersal distribution. For these analyses, we assume that dispersal follows a geometric dispersal model described by equation (1). The performance of estimators of Nm, $N\mu$, and g when $g = 0$, 0.2, and 0.5 is presented in table 3. For all cases in this table, $N = 400$ haploid individuals

Downloaded from https://academic.oup.com/mbe/article/24/12/2730/978152 by Bibliothèque Universitaire de médecine - Nîmes user on 16 June 2021

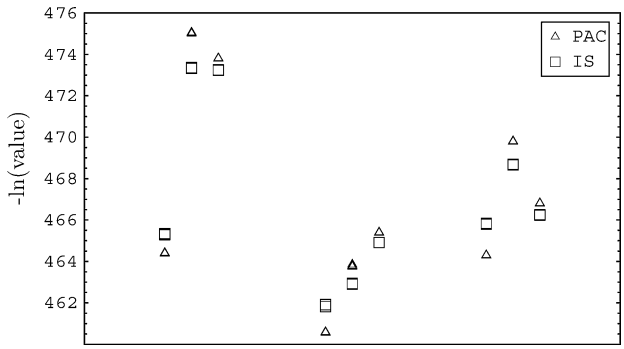

FIG. 1.—Differences between PAC likelihood and IS estimation. Each of the 9 columns shows 3 independent estimates of the PAC likelihood (△) and 3 estimates of the likelihood (□) for one given parameter point and one given sample. For each of them, the 3 replicates are barely distinguishable because the variance of estimation is very low. The 9 columns represent groups of 3 parameters points for each of 3 samples. Samples and PAC likelihood analyses are as in case [34] but with $n_t = 1{,}000$. IS analysis are for the same statistical model and same n_t.

per deme; samples were generated assuming a 10-alleles model; $n_t = 30$ and $n_p = 2197$ or 5,000; and sampled ranges were $2N\mu \in [0.125, 5]$, $2Nm \in [2.5, 20]$, and $g \in [0, 0.8]$.

The estimation of the parameter g appears very poor when $n_m = 10$. For $g = 0$, the estimator is biased upward (case [20]). A bias is expected for a MLE as the parameter value is at the boundary of the feasible parameter range, but in the present case this bias is high. For $g = 0.5$, the estimator of g is biased downward (case [27]). The weak bias observed for $g = 0.2$ may be the midpoint between the positive bias for lower values of g and the negative bias for higher value of g. For $g = 0$, there is also a slightly higher absolute bias of Nm estimates (CI 0.093–0.255) relative to analyses of the same data under the stepping stone model (cases [18] vs. [20]).

These biases may result both from model mis-specification, small sample size, and inaccurate estimation of likelihood. Effects of model mis-specification can be evidenced only by comparison with analyses assuming the true number of demes (100), and such analyses can be done routinely only with PAC likelihood. However, the value of the likelihood statistic under the "true" model was compared with the PAC likelihood in a few points, and there are demonstrable differences (fig. 1). Yet, maximum PAC likelihood performance appears at least as good as ML performance (cases [20]–[32] in table 3) as the differences are mostly in the direction of lower MSE by maximum PAC likelihood. For maximum PAC likelihood when $n_m = 100$ and $g = 0.5$ (case [34]), Nm is relatively well estimated, $N\mu$ estimates remain biased upward, and g estimation is not precise enough to be worth considering in practice. Thus, there is little information about g in the data. Accordingly, we will increase sample size (in particular, the number of demes sampled) in later simulations. With the present sampling design, performance is as good with 40 as with 100 demes, but there is evidence of mis-specification on Nm and $N\mu$ estimation as the bias and MSE of their estimators decrease with n_m increasing from 10 to 40.

Not only the number of demes but also the position of samples relative to the total habitat may be mis-specified.

Table 4
Edge Effects

	n_m	$N\mu$ relative bias (relative $\sqrt{\text{MSE}}$)	Nm relative bias (relative $\sqrt{\text{MSE}}$)	g bias ($\sqrt{\text{MSE}}$)
Position effect, $g = 0.5$, 5 loci				
[37] PAC[a]	25	0.20 (0.59)	0.11 (0.25)	−0.12 (0.25)
Analyses under circular array model				
$g = 0$, 5 loci				
[38]	10	0.32 (0.43)	−0.17 (0.29)	0.06 (0.08)
$g = 0.2$, 5 loci				
[39]	10	0.44 (0.59)	−0.06 (0.34)	−0.07 (0.14)
[40]	25	0.21 (0.35)	0.14 (0.41)	−0.07 (0.14)
[41] PAC	40	0.04 (0.25)	0.01 (0.31)	−0.02 (0.18)
[42] PAC	100	0.04 (0.25)	0.01 (0.31)	−0.01 (0.19)
$g = 0.5$, 5 loci				
[43]	10	0.86 (0.96)	0.17 (0.32)	−0.28 (0.35)
[44][b] $n_t = 120$	10	0.91 (1.04)	0.17 (0.31)	−0.31 (0.36)
[45]	25	0.53 (0.76)	0.19 (0.31)	−0.22 (0.29)
[46] PAC	25	0.22 (0.59)	0.10 (0.25)	−0.13 (0.25)
[47] PAC	40	0.20 (0.61)	0.12 (0.25)	−0.12 (0.25)
$g = 0.5$, 20 loci				
[48][b]	10	0.84 (0.87)	0.11 (0.25)	−0.28 (0.31)
[49] PAC	100	0.02 (0.29)	0.06 (0.19)	−0.06 (0.17)

NOTE.—Samples and simulation conditions as in Table 3.

[a] Samples set in positions 10, 12, 14, 16 of the array versus 3, 5, 7, 9 in other analyses with $n_m = 25$.

[b] The analysis of each sample from cases [44] and [48] (5,000 points) takes 12 CPU hours on 2.66 GHz processors.

Estimation performance may differ whether samples are set close to the assumed edge of the habitat or in its center as shown in one example (case [37] vs. [32], differing in particular through the $N\mu$ mean, CI on relative effect 0.16–0.39). Analyses under a circular array model were investigated as a practical alternative to having to choose the position of samples on a linear lattice (table 4 and fig. 2). Overall, the performance depends somewhat on whether a circular or linear array is assumed, but this does not affect the previous conclusions (including the consistently smaller bias of maximum PAC likelihood estimates relative to MLEs across simulation conditions, and the improvement in $N\mu$ estimation when n_m is increased). As could be expected, the highest discrepancies between linear and circular analyses are observed for the lowest n_m and highest g value and should be generally negligible relative to other causes of error. As before, strong biases may be observed for the lowest n_m values, and increasing the number of replicate ancestral trees ([44] vs. [43]) or the number of loci (case [48]) has little effect, confirming that the biases are mostly due to mis-specification. The differences between circular and linear models are very small for $n_m \geq 40$ (with most CI widths for effects on means narrower than 0.01 in the PAC likelihood analyses; see fig. 2 legend).

Larger Samples

The previous results show that unaccounted populations become important when there is some "long-distance" dispersal ($g > 0$; keeping in mind that $g = 0.5$ implies only limited long-distance dispersal, compared with many biological studies). Further, even with a correctly specified model, there is less information about g in the data

Downloaded from https://academic.oup.com/mbe/article/24/12/2730/978152 by Bibliothèque Universitaire de médecine - Nîmes user on 16 June 2021

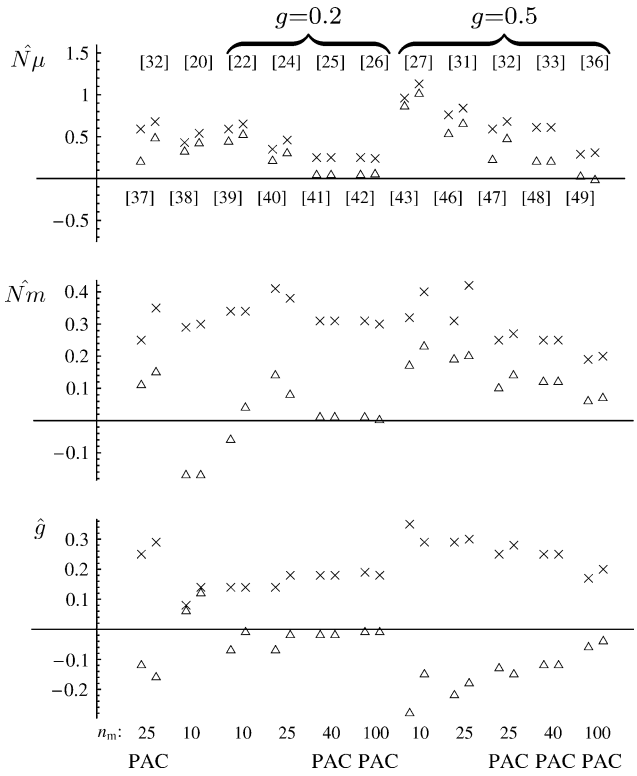

FIG. 2.—Interaction of edge effects with n_m. This figure compares selected results from tables 3 and 4. (Relative) bias (△) and $\sqrt{\text{MSE}}$ (×) are shown for the 3 estimated parameters. Simulation conditions were identical in the paired linear and circular analyses, including the sequence of the random number generator (for cases [20] and [32], this may involve only a subset of all replicates considered in table 3). Hence, with PAC likelihood, the same sequences of $\hat{\pi}$'s were computed, the only difference being the equations that the $\hat{\pi}$'s solve; whereas in the ML analysis, a $\hat{\pi}$ value at some step affects the further sequence of $\hat{\pi}$ values computed. This results in a much smaller variance of differences between paired PAC likelihood analyses compared with paired ML analyses (paired PAC likelihood analyses may remain statistically different even when the differences are not visible).

than about the migration rate. Overall, Nm was relatively well estimated in most cases, estimation of mutation rate was affected by mis-specification, and estimation of g was affected both by mis-specification and by lack of power.

To increase power, both larger number of demes sampled and of loci were considered (table 5). As previously, there is no evidence of mis-specification with 40 demes in the statistical model. The performance of maximum PAC likelihood (with 5 loci, case [51]) or of ML estimation (with 10 loci, case [50]) is excellent. Similar results are obtained with $n_t = 10$ or 100 sequences (case [52] vs. case [51], all $P > 0.059$, all CI bounds on effects < 0.074) for PAC likelihood computation, confirming that a low n_t is enough. Good performance is confirmed in the case $g = 0.2$ (case [56]). However, if samples come from a 1,000-demes array, a slight bias reappears for $N\mu$ estimates (case [53]), whereas a more substantial bias appears if the true mutation rate is reduced to 10^{-4} (case [54]). The latter phenomenon will be investigated more thoroughly below. Finally, analysis of data generated under a bounded SMM shows an ≈50% reduction in mutation rate estimates but no notable effect on dispersal estimation (case [55]).

Table 5
Performance of Estimation for Geometric Dispersal

	$N\mu$ relative bias (relative $\sqrt{\text{MSE}}$)	Nm relative bias (relative $\sqrt{\text{MSE}}$)	g	Bias ($\sqrt{\text{MSE}}$)
[50][a] IS	0.04 (0.25)	0.08 (0.17)	0.5	−0.05 (0.14)
[51] PAC	0.02 (0.29)	0.01 (0.13)	—	−0.02 (0.15)
[52] PAC, $n_t = 100$	0.01 (0.24)	0.03 (0.13)	—	−0.04 (0.14)
[53] $n_d = 1{,}000$, PAC	0.17 (0.33)	−0.01 (0.16)	—	−0.05 (0.15)
[54] PAC, $\mu = 10^{-4}$	0.77 (0.99)	0.03 (0.18)	—	−0.09 (0.13)
[55] SMM, PAC	−0.54 (0.57)	0.11 (0.25)	—	0.006 (0.14)
[56] PAC	−0.08 (0.17)	0.10 (0.3)	0.2	0.04 (0.12)

NOTE.—For each sample analyzed, 60 genes were sampled at each of 5 loci (10 in case [50]) in each of 10 demes out of a linear array of 100 demes (except 1,000 for case [53]). The mutation probability was 10^{-3} (except 10^{-4} for case [54]). $n_m = 40$ and $n_t = 10$ except 100 for case [52]. Other simulations settings were as in table 3. 2197 points were analyzed, except in case [50].

[a] In case [50], 2 steps of 512 points were computed as described in the text.

Smaller Demes with Higher Dispersal

Our aim is to test the performance of the algorithms in scenarios more representative of spatial structure at small spatial scale. In this section, we consider samples from a population of 1,000 demes, with fewer individuals per deme and higher dispersal rate. Twenty demes are sampled, so the number of demes in the statistical model is always larger than 20, and then ML analyses based on the IS algorithm become extremely time consuming. Hence, only maximum PAC likelihood is considered in all but one simulation.

The results are presented in table 6 and fig. 3. For $g = 0.5$ and $n_m = 60$, there are strong biases, in particular for $N\mu$. Mutation and migration rates are overestimated, whereas g is slightly underestimated. As before, simulation conditions were varied to understand these biases. For $n_m = 60$, increasing the number of loci yields reductions of variance of estimators (all >50% for both case [58] vs. [62] and case [61] vs. [60]) but no significant, or even consistent, improvement in biases. Varying the number of points (cases [57] vs. [59], [58] vs. [61], and [60] vs. [62]) or of replicate sequences (cases [57] vs. [58] and [59] vs. [61]) has no detectable effect, except for significant but still small effects (a few percents at most on biases) for cases [60] vs. [62]. Only increasing n_m to 200 demes does result in improved performance, with reduction of $N\mu$ bias (CI on relative reduction 0.17–0.40) and of g bias (CI on reduction 0.007–0.06; $N\sigma^2$ bias is likewise reduced). Improvement in bias of the same parameters for an identical increase in n_m is also apparent for a higher level of dispersal ($g = 0.75$, case [68] vs. [71]; $N\mu$ relative reduction 1.05–1.35, g reduction 0.002–0.07), although all biases are more moderate and less affected by n_m for lower level of dispersal ($g = 0.2$, case [66] vs. [65], all $P > 0.59$ for biases). Thus, the mis-specification problems previously encountered are met again, but at higher n_m values, when the total dispersal rate is increased.

Whether poor performance is due in part to the PAC likelihood heuristics can only be assessed by comparison with the ML analysis. One comparison was conducted for $n_m = 60$ and $g = 0.75$ (cases [67] vs. [68]), and both analyses yield very similar results, except that the PAC likelihood estimates of $N\mu$ are slightly less biased (0.03–0.15 relative reduction; MSE is reduced too). For both IS and

Downloaded from https://academic.oup.com/mbe/article/24/12/2730/978152 by Bibliothèque Universitaire de médecine - Nîmes user on 16 June 2021

Table 6
Performance of Estimation under High Dispersal

	n_m	$N\mu$ relative bias (relative $\sqrt{\text{MSE}}$)	Nm relative bias (relative $\sqrt{\text{MSE}}$)	g bias ($\sqrt{\text{MSE}}$)	$N\sigma^2$ relative bias (relative $\sqrt{\text{MSE}}$)
Samples from an array of 1,000 demes of 40 haploid individuals, with $m = 0.25$ ($N\mu = 0.04$, $Nm = 10$)					
$g = 0.5$, $N\sigma^2 = 60$					
[57]	60	0.58 (0.74)	0.91 (1.05)	−0.19 (0.23)	−0.12 (0.29)
[58] ($n_t = 30$)	—	0.58 (0.76)	0.88 (1.01)	−0.19 (0.21)	−0.10 (0.29)
[59] ($n_p = 2197$)	—	0.61 (0.79)	0.92 (1.05)	−0.18 (0.21)	−0.08 (0.27)
[60] ($n_t = 30$, $n_p = 2{,}197$, 20 loci)	—	0.56 (0.60)	0.86 (0.90)	−0.18 (0.19)	−0.09 (0.17)
[61] ($n_t = 30$, $n_p = 2{,}197$)	—	0.56 (0.74)	0.94 (1.10)	0.18 (0.22)	−0.08 (0.28)
[62] ($n_t = 30$, 20 loci)	—	0.53 (0.55)	0.95 (1.00)	−0.20 (0.21)	−0.12 (0.03)
[63]	200	0.27 (0.69)	0.83 (0.99)	−0.16 (0.19)	−0.05 (0.28)
[64]	40	1.26 (1.34)	0.96 (1.08)	−0.19 (0.22)	−0.06 (0.26)
$g = 0.2$, $N\sigma^2 = 18.75$					
[65]	200	0.38 (0.52)	0.41 (0.52)	−0.07 (0.08)	0.08 (0.22)
[66]	60	0.35 (0.44)	0.39 (0.45)	−0.08 (0.08)	0.09 (0.21)
$g = 0.75$, $N\sigma^2 = 280$					
[67] IS	—	1.51 (1.64)	0.77 (0.86)	−0.04 (0.13)	2.10 (4.01)
[68]	—	1.42 (1.55)	0.74 (0.84)	−0.03 (0.13)	2.45 (4.30)
[69][a] IS	—	1.57 (1.70)	0.75 (0.86)	−0.07 (0.12)	0.59 (2.63)
[70][b]	—	1.47 (1.60)	0.70 (0.78)	−0.07 (0.11)	0.53 (2.58)
[71]	200	0.27 (0.61)	0.54 (0.59)	−0.03 (0.07)	0.46 (1.05)
[72] SMM	—	−0.17 (0.44)	0.78 (0.86)	−0.08 (0.11)	0.16 (0.70)

NOTE.—Estimation is by maximum PAC likelihood except for cases [67] and [69]. Except as noted, $n_t = 10$, $n_p = 512$ and sample sizes were 5 loci per sample, 20 demes sampled, and 20 genes sampled per deme. Thirty such samples were analyzed in each case, except 120 samples for cases [67]–[70]. The parameter ranges explored were $2N\mu \in [0.0125, 0.5]$, $2Nm \in [2.5, 60]$ (except for $g = 0.2$ where $2Nm \in [2.5, 40]$ was sufficient), and $g \in [0.05, 0.8]$ except $g \in [0.2, 0.999]$ when true $g = 0.75$.

[a] Second step of 512 points after case [67].

[b] Second step of 512 points after case [68]. It took ≈ 20 min per sample on 2.6 GHz 64-bit processors. First and second iterations required about 1,000 and 110 more time, respectively, for IS computation than for PAC likelihood.

PAC likelihood methods, increasing the number of parameter points (cases [69] and [70]) markedly improved $N\sigma^2$ estimation. For the other parameters, the distribution of estimates are shown in fig. 4; PAC likelihood was again only slightly, though consistently, better than ML. Thus, the PAC likelihood heuristics again appears as an excellent substitute to likelihood estimation.

The effect of a bounded stepwise mutation process was tested again (case [72]), and it was again found that it yielded a reduction in mutation rate estimates and little effect on other parameters.

Lower Mutation Rate

Performance was assessed for a lower mutation rate ($\mu = 10^{-4}$), still with relatively high dispersal rates and small deme sizes (table 7 and fig. 3). For all cases in the table, samples were simulated for an array of 1,000 demes of 40 haploid individuals, with $m = 0.25$, 0.5, or 0.75 and $\mu = 10^{-4}$ ($N\mu = 0.004$, $Nm = 10$–30). Estimation was by maximum PAC likelihood with $n_t = 10$, and $D\sigma^2$ estimates were compared with those obtained by the moment method. Except as noted, sample sizes were 5 loci per sample, 20 demes sampled, and 40 genes sampled per deme; sampled ranges were $2N\mu \in [0.00125, 0.25]$; $2Nm \in [2.5, 60]$, $g \in [0.05, 0.8]$ when true $g = 0.5$ and $g \in [0.2, 0.999]$ when true $g = 0.75$. Genetic diversity remains high in these simulations. For example, in case [78], the probability of identity in the samples was 0.245.

With respect to $N\mu$ estimation, for $n_m = 60$, the bias appears high, but only following the previous decision to measure biases relative to $n_mN\mu$ rather than relative to $n_dN\mu$, the mutation rate scaled by the true total size of the population. For the highest dispersal, the bias is much weaker when assessed relative to $n_mN\mu$, and thus estimation performs more in accordance to the general definition of the algorithm. Expectedly, the biases are reduced when n_m is increased to 200, though still large in the highest dispersal case. Again, data simulated under a bounded SMM (case [84]) yield lower estimates of mutation rate.

Dispersal rate estimates are also substantially biased but can be interpreted as low-bias estimates neither of n_mNm nor of n_dNm. In particular, they do not scale as n_d in the highest dispersal case. F_{ST}-based estimates of Nm could well be better than those derived under $n_m = 60$. The biases on Nm and g seem to compensate each other, yielding low relative biases on $N\sigma^2$ estimation. Most biases are reduced when the number of demes is increased, but additionally increasing the number of loci has little effect, which again indicates that large biases are mostly due to mis-specification.

$N\sigma^2$ estimates could be compared with those obtained by a moment method (Rousset 1997). Following the classical bias-variance trade-off of likelihood estimators, the PAC likelihood estimator generally has lower variance but higher bias than the moment estimator. The PAC likelihood estimator may have lower MSE overall, as one might expect under well-specified models, but the trend is not clear cut, which leaves room to speculate what would be the "best" method in practical conditions. The comparison could have been more favorable to the likelihood method in simulation conditions with relatively large μ/m as mutation is expected to bias results of the moment method in that case.

Downloaded from https://academic.oup.com/mbe/article/24/12/2730/978152 by Bibliothèque Universitaire de médecine - Nîmes user on 16 June 2021

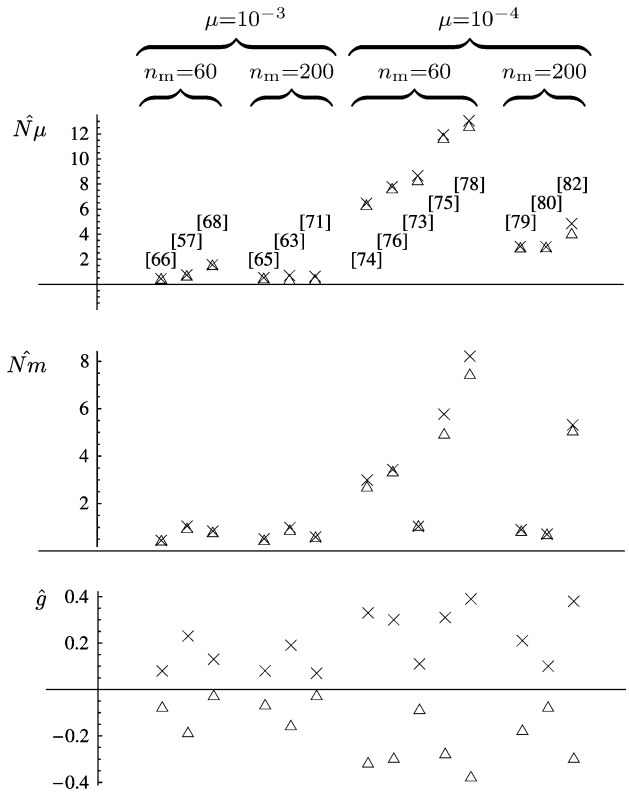

FIG. 3.—Interaction of mutation with n_m. This figure compares selected results from tables 6 and 7. (Relative) bias and $\sqrt{\text{MSE}}$ are shown as in Fig. 2. For each (μ, n_m) combination, cases are ranked by increased $N\sigma^2$ values.

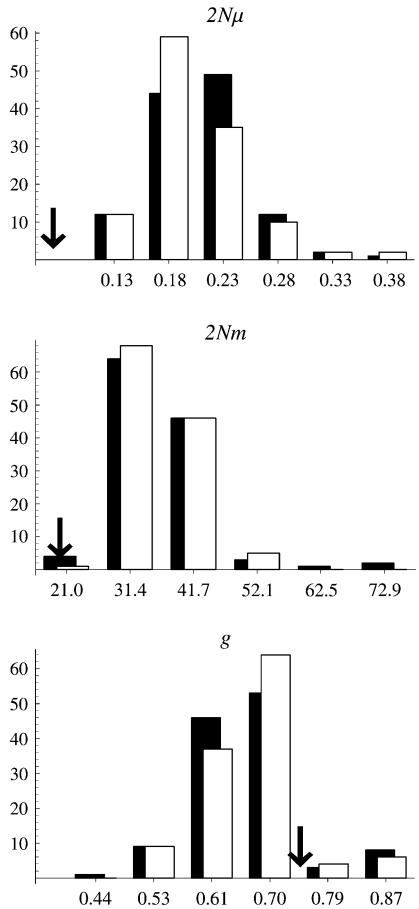

FIG. 4.—Distributions of estimates by PAC likelihood and IS estimation. The PAC likelihood distributions (case [70]) are laid over the likelihood distributions (case [69]). The arrows mark the position of the parameter values.

$N\sigma^2$ estimation is better ceteris paribus with $n_m = 60$ (case [78]) than with $n_m = 200$ (case [82]) for the same data, although the opposite effect of mis-specification holds for the other parameters. This suggests that specifying a large number of demes is less important for good estimation of $N\sigma^2$ than for other parameters.

As usual, it is not a priori obvious to which extant a relatively poor performance is due to numerical issues. In particular, for high g values, a minor error in finding the g MLE results in a high error on $N\sigma^2$ estimates. This effect was already apparent in cases [67]–[70], where increasing the density of points analyzed markedly improved $N\sigma^2$ estimation. We have further tested the effect of numerical parameters in cases showing the poorest $N\sigma^2$ estimation relative to the moment method. For $n_m = 60$, increasing n_t to 100 (case [77] vs. [76]) had no notable effect on the conclusions, whereas when $n_m = 200$, increasing n_t to 50 (case [83] vs. [82]) substantially reduced the MSE. Thus, mis-specification is the main determinant of poor $N\sigma^2$ estimation for low n_m, whereas a higher number of replicates become necessary for $N\sigma^2$ estimation with high n_m. The latter conclusion was confirmed when the mutation model is also mis-specified (cases [84]–[86]). The performance notably improves when n_t is increased from 10 to 50, mainly due to improvement of a few outlying estimates. As before, PAC likelihood performs as least as well as ML in this case; MLEs for $N\mu$ and $N\sigma^2$ actually have higher MSE ($P < 0.027$).

Application to Real Data

Watts et al. (2007) have compared genetic and demographic estimates of $D\sigma^2$ in the damselfly *Coenagrion mercuriale* along a linear habitat. The demographic estimate of $4D\sigma^2$ (for D here being a density of diploid individuals) derived from a mark–recapture study and corrected for variance in reproductive success was 277,894 individuals.m. Indirect estimates obtained from a sample of 240 individuals and 14 loci by several variants of the regression method based on pairwise comparison of individual genotypes (Rousset 2000) ranged within 179,058–242,816, with a synthetic CI 66,015–392,866.

In simulation conditions fitted to these data with respect to sampling design, gene diversity, dispersal distribution, and total population size (Watts et al. 2007, case σ = 130), the relative bias and root MSE of $D\sigma^2$ estimates yielded by the $\hat{e}$ regression estimator were 0.93 and 2.55 (reduced to 0.67 and 1.31 when 3 outliers are taken out of 200 replicates), and the other regression estimator considered yielded some negative estimates (for ease of comparison and as previously discussed in Leblois et al. 2003, bias and MSE of the more Gaussian-distributed $1/(D\sigma^2)$ were instead reported in Watts et al.). These biases are

Downloaded from https://academic.oup.com/mbe/article/24/12/2730/978152 by Bibliothèque Universitaire de médecine - Nîmes user on 16 June 2021

Table 7
Performance of Estimation under Lower Mutation Rate

	$N\mu$ relative bias (relative $\sqrt{\text{MSE}}$)	Nm	Relative bias (relative $\sqrt{\text{MSE}}$)	g	Bias ($\sqrt{\text{MSE}}$)	$N\sigma^2$	Relative bias (relative $\sqrt{\text{MSE}}$)	
							PAC	Regression
$n_m = 60$								
[73][a]	8.19 (8.66)	10	1.01 (1.04)	0.75	−0.09 (0.11)	280	0.16 (0.54)	0.003 (0.50)
[74][a]	6.22 (6.46)	20	2.66 (2.99)	0.5	−0.32 (0.33)	120	0.05 (0.27)	−0.08 (0.44)
[75][a]	11.56 (11.94)	—	4.89 (5.77)	0.75	−0.28 (0.31)	560	0.13 (0.43)	−0.04 (0.44)
[76][a]	7.54 (7.78)	30	3.30 (3.43)	0.5	−0.30 (0.30)	180	0.36 (0.50)	−0.08 (0.20)
[77][b]	7.95 (8.19)	—	3.82 (3.97)	—	−0.34 (0.34)	—	0.31 (0.46)	−0.08 (0.20)
[78][a]	12.53 (13.05)	—	7.41 (8.21)	0.75	−0.38 (0.39)	840	0.09 (0.45)	−0.06 (0.55)
$n_m = 200$								
[79]	2.83 (3.01)	—	0.79 (0.90)	—	−0.18 (0.21)	60	−0.12 (0.26)	−0.08 (0.37)
[80]	2.85 (2.99)	10	0.66 (0.71)	—	−0.08 (0.10)	280	0.02 (0.44)	0.003 (0.50)
[81] (20 loci)	2.84 (2.89)	—	0.61 (0.63)	—	−0.07 (0.08)	—	−0.06 (0.16)	−0.04 (0.25)
[82][a]	3.96 (4.84)	30	5.02 (5.31)	—	−0.30 (0.38)	840	0.18 (0.70)	−0.06 (0.55)
[83][c]	3.79 (4.07)	—	3.66 (3.68)	—	−0.24 (0.25)	—	0.12 (0.46)	−0.06 (0.55)
Samples generated under SMM								
[84]	2.37 (2.85)	10	0.73 (0.80)	—	−0.08 (0.11)	280	0.18 (0.92)	−0.09 (0.57)
[85][d]	1.95 (2.45)	—	0.74 (0.79)	—	−0.07 (0.10)	—	0.17 (0.80)	−0.09 (0.57)
[86][e]	2.12 (2.71)	—	0.71 (0.76)	—	−0.08 (0.10)	—	0.13 (0.67)	−0.09 (0.57)
[87][f] IS	1.73 (1.82)	—	0.87 (0.91)	—	−0.06 (0.08)	—	0.34 (0.72)	
Versus [86] (subset)[g]	1.39 (1.51)	—	0.84 (0.87)	—	−0.07 (0.09)	—	0.18 (0.45)	

NOTE.—In each case, 200 samples were analyzed by the moment method, and 30 samples were analyzed by maximum PAC likelihood, except for cases [84] and [86] (60 samples) $n_p = 512$, except as noted.

[a] For analyses with $n_m = 60$, as well as case [82], where large biases were observed and difficult to anticipate, 2 steps of 512 points were computed as described in the text. In the first step, estimates were deduced from 512 points in the range $2N\mu \in [0.00125, 0.25]$, $2Nm \in [2.5, 450]$, and $g \in [0.05, 0.999]$.

[b] Two-steps procedure, $n_t = 100$ after first step of analysis of case [76].

[c] Two-steps procedure, $n_t = 50$ after first step of analysis of case [82].

[d] Two-steps procedure, the first step being case [84].

[e] As case [85] but with $n_t = 50$ in the second step.

[f] In case [87], the first 10 samples of case [86] were analyzed by ML with $n_p = 256$ and $n_t = 50$. This computation takes about 13 CPU years on 2.8-GHz processors.

[g] Same 10 samples as in case [86].

largely small-sample ones as could be seen by comparison with an estimate from 40,000 loci. More important though, the CI deduced jointly from the 2 regression estimators were little affected by such biases and had good coverage properties (Watts et al. 2007). To analyze the same simulated data by PAC likelihood, individual genotypes have to be binned in artefactual demes. Here, 80 such demes were defined exactly as described below for the actual data analysis. Geometric dispersal is still assumed in the statistical model, which now implies mis-specification of the dispersal distribution. Despite this, estimation performance is substantially improved as the maximum PAC likelihood has a lower relative root MSE of 0.54 (bias is 0.8%; from 60 replicates).

We have reanalyzed the damselfly data by PAC likelihood. Here, the linear habitat was divided in n_m spatial units of width $3500/(n_m - 1)$ m, the patch of habitat (the "Lower Itchen Complex" in Watts et al. 2007, fig. 1) being about 3500 m long. For $n_m = 80$, several gradually more focused (in parameter space) analyses led to $4D\sigma^2 \hat{=} 2159$, the unit being individuals.(bin width). When translated back to individuals.m, this is $4D\sigma^2 \hat{=} 95,645$. Several independent, less focused replicate analyses yielded likelihood ratio CI ≈50,000–140,000 (fig. 5 shows one such computation, where the Nb estimate is 92,039). Similar computations yielded $4D\sigma^2 \hat{=} 123,676$ and $113,303$ individuals.m for $n_m = 5$ and 20, respectively. Thus, as did n_m in the simulations, the bin width has little effect on $D\sigma^2$ estimates, although it affects more the other estimators.

Discussion

Performance of Estimation

In this work, we have investigated the performance of ML estimation of mutation and dispersal parameters under isolation by distance in a linear habitat, using de Iorio and Griffiths' IS algorithm. We have focused on the effect of mis-specification of the number of demes. In the same conditions, we have also found that the maximum PAC likelihood approximation is practically as efficient as ML analysis.

Beyond the simulation results reported in this ms, we have considered some additional approximations that would ease computations for large arrays of demes. In particular, approximation of the remote ancestry of a sample by Kingman's coalescent has been considered in 2-dimensional models (Cox 1989; Cox and Durrett 2002; Zähle et al. 2005), but for ancestors of genes uniformly sampled on the lattice, rather than in a small part of it as considered here. Even for uniform sampling, both analysis (Cox 1989) and simulations (Wilkins 2004) suggest it is not appropriate for linear habitats. In agreement with these results, we could not achieve good performance by such approximations while simultaneously reducing computation time by a notable extent (details not shown).

Expectedly, there is good performance of ML in favorable conditions (no model mis-specification, large sample size). In less favorable conditions, performance is affected differentially for different parameters. Estimation of g is often very poor (e.g., fig. 4). In general, the dispersal rate

Downloaded from https://academic.oup.com/mbe/article/24/12/2730/978152 by Bibliothèque Universitaire de médecine - Nîmes user on 16 June 2021

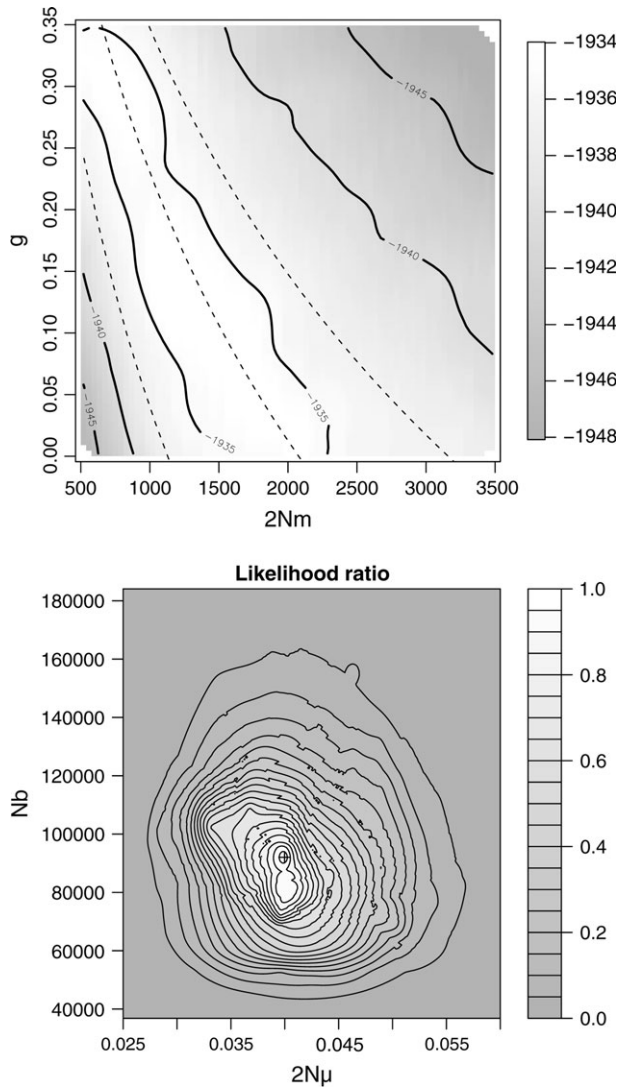

FIG. 5.—PAC likelihood surface for the *Coenagrion mercuriale* data set. Top: a contour plot of likelihood for $2N\mu = 0.4$. Dashed lines are lines of equal Nb values (50,000, 92,039, and 140,000 from left to right). PAC likelihood analysis was computed for 2197 points with $n_t = 100$ (requiring 52 CPU hours on 2.66 GHz processors). Bottom: a profile likelihood ratio plot from the same PAC likelihood computation, derived by profiling over (Nm, g) values for given neighborhood size (Nb). The likelihood ratio relative to the maximum is shown. Confidence regions are given by the χ^2 approximation for the profile likelihood ratio (Cox and Hinkley, 1974, pp. 322 sqq). With 2 df, the 95% confidence region for ($2N\mu$, Nb) is simply bounded by the 0.05 level for the profile likelihood ratio (i.e., the outer level in this plot), whereas the 95% CI for Nb is given by the 0.1465 level of the 1-dimensional profile. The point estimate is marked by a +.

Nm appears easier to estimate to the extent that ~60% relative biases are deemed acceptable, but larger biases can again result from mis-specification. Large biases occur even though sampled demes exchange essentially no migrants with demes not accounted in the statistical model. Positive biases of Nm estimates and negative biases of g estimates compensate each other to yield $N\sigma^2$ estimates with low relative MSE. Thus, the largest $N\sigma^2$ MSEs are obtained when g is well estimated (i.e., for $g = 0.75$ in tables 6 and 7). Figure 5 also shows that there is more information about $N\sigma^2$ than about Nm and g separately.

Estimates of the mutation rate $N\mu$ are generally biased upward, to some extent by small sample bias, but in particular when fewer subpopulations are considered in the statistical model than in simulation of the data and when the mutation rate is low. Local diversity contains information about the mutation rate scaled by the total population size (Nagylaki 1983; Slatkin 1987; Strobeck 1987), so the likelihood must depend on the total size of the population. However, the dependence of diversity on total size may be only perceptible for small mutation rates. For high mutation rates and low dispersal, the probability of identity within demes (or a few demes apart) depends little on the total size of the array (see, e.g., comparison of Maruyama's (1970b) finite lattice results to Nagylaki's (1974) infinite lattice results in fig. 2 of Cox and Durrett 2002), so that there may be little statistical information to distinguish between a 40-demes and a 1,000-demes array. This may explain why distribution of $N_T\mu$ estimates appear closer to $Nn_m\mu$ than to the true $N_T\mu$ value for the higher mutation rate (10^{-3}) and lower dispersal, whereas the reverse holds for higher dispersal and lower mutation rate (10^{-4}, table 7).

Similar trends are observed for Nm estimates but are not always so easily understood. For high mutation ($\mu = 10^{-3}$), relatively good estimation of Nm can be achieved. For $\mu = 10^{-4}$ and large dispersal, no migration rate appears well estimated when deme number is mis-specified. This does show how easily likelihood methods could be misused in realistic conditions.

There would be considerable difficulties, both conceptual and statistical, in trying to estimate the number of demes itself. Over the timescale of genealogical processes, assuming a fixed number of demes is often no more than a convenient device. Even in the ideal case considered in the simulations, comparing the PAC likelihoods of the fitted parameter values under models with different number of demes does not point to good estimates of this number. In the 3 cases from table 7 were the comparison was possible, the data were equally well fitted under the 60-demes model as under the 200-demes model; the fitted PAC likelihoods were, if anything, slightly lower for larger number of demes, thus pointing away from the true value of 1,000 demes.

Because any pure simulation study may miss important factors affecting the performance of estimation, comparisons with demographic estimates are also important to evaluate the possible impact of factors ignored in the simulations and eventually to force us to consider additional factors. In the present case, the maximum PAC likelihood estimate is ~3 times lower than the demographic estimate, and its CI excludes this estimate. As discussed by Watts et al. (2007), the demographic estimate reported in that study is a worst-case overestimate for comparisons with genetic estimates, in that no attempt was made to correct for variations in population density over years. The genetic point estimates differ in a manner consistent with the expected bias of the regression estimators under simulation conditions fitted to the conditions of the population studied, but more importantly the CI obtained by the regression method overlaps widely with the one given by PAC

Downloaded from https://academic.oup.com/mbe/article/24/12/2730/978152 by Bibliothèque Universitaire de médecine - Nîmes user on 16 June 2021

likelihood. Hence, one explanation consistent with all available evidence is that both genetic methods estimate, with different small-sample biases, the same effective $D\sigma^2$ and that the demographic estimate was too high. Although discrepancies between the different methods (in particular, asymptotic bias) could still be sought, they would be of the order of differences in confidence limits (20% for the lower bound, 145% for the upper bound). Further comparisons would be necessary to demonstrate systematic differences of this magnitude.

We have assumed the same mutation rate for all loci. It is unclear how variation in mutation rate would affect the analyses, and estimating one mutation parameter per locus would be both highly impractical and would increase the MSE of the other estimates. A tentative solution to this problem could be to use a random effect model for mutation, that is, to integrate the likelihood over a distribution of mutation rates, of which some parameters would be estimated.

Predicting Mis-Specification Biases

Although our analysis has highlighted the biases resulting from mis-specification of the number of demes, some of these biases appear small compared with the accuracy sometimes expected (Whitlock and McCauley 1999) from analyses of spatial genetic structure. How far this conclusion will remain true when a wider range of biological scenarios is considered? It would be helpful to be able to predict biases by relatively simple arguments.

In an idealized world, spatial patterns would contain no information about mutation rates, and dispersal rates could be estimated independently of mutation. To some extent, this is what occurs with moment methods based on probabilities of identity of pairs of genes: local diversity depends on the mutation rate but F_{ST} and related quantities are relatively independent of mutation (Crow and Aoki 1984; Slatkin 1991), particularly at a local geographical scale (Rousset 1996). Thus, it is to some extent possible to estimate dispersal rate without good estimates of mutation rates. The present results, as those of Beerli (2004), suggest a similar behavior, in that mutation rate estimation is more affected by mis-specification. However, cases where mis-specification also notably affects estimation of dispersal were also pointed out.

Attempts to estimate simultaneously dispersal and mutation by moment methods could also result in biased estimates of both parameters (an example will be presented below). It is therefore tempting to try to predict the biases of MLEs from the analytical theory for pairs of genes, and the number of demes to be considered in the statistical model might be predicted from such theory. Bias prediction was considered by Slatkin (2005), but the approximations of diversity by expected coalescence times he considered do not describe well genetic identity at loci with high mutation rates. In addition, it may not be possible to fit exactly all probabilities of identity in a large array of demes to a model with few parameters. Validating any prediction procedure is bound to be complex.

Nevertheless, the simple example of the island model can be used to support such a logic. A way of predicting biases in the island model is to compute expected values of within- and among-deme probabilities of identity for the actual number of population and to find the numerical values of the mutation and migration rates which would give the same probabilities of identity for the assumed number of demes in the estimation model. These computations are straightforward (e.g., Nagylaki 1983; Rousset 2004, pp. 27, 224).

Thus, in the demographic conditions of case [51] ($n_d = 100$ demes, $N = 400$, $m = 0.01$), but for an island model of dispersal, if the mutation probability is 10^{-4} (for a KAM with 10 alleles) the probabilities of identity within and among demes are 0.269 and 0.181, and the mutation and migration probabilities which yield the same probabilities in a 10-demes model are 9.1×10^{-4} and 0.0083. The predicted relative biases are therefore 8.1 and −0.17. The observed biases were close: 8.8 and −0.20 out of 60 replicate samples of 5 loci (further simulation details not shown). The observed biases are thus well predicted and close to those of a moment method using the information contained in probabilities of identity. Likewise, if the mutation probability is 10^{-3}, the probabilities of identity within and among demes are 0.196 and 0.108, and the mutation and migration probabilities which yield the same probabilities in a 10-demes model are 0.0053 and 0.0048, yielding predicted relative biases 4.3 and −0.52, respectively. The observed biases were again close: 4.04 and −0.46 out of 60 replicate samples of 5 loci.

Beyond illustrating a case where the probabilities of identity provide good prediction of MLE biases, these examples also illustrate the simple expectation, consistent with the other simulation results, that the relative bias on mutation estimation will be of the order of the n_d/n_m ratio when the mutation rate is low and lower for higher mutation rates. In the latter case, however, the bias on the migration rate can be large and not so easily interpreted.

Therefore, a higher number of demes might need to be considered for lower mutation rates, which could be a serious practical problem for some types of markers. The comparison of $N\mu$ biases in table 6 versus table 7 supports this idea. Local diversity is more sensitive to total size when dispersal is less localized (higher m or g values). So, by the same logic, a higher number of demes should be considered when dispersal rates are higher, which is indeed observed in our simulations. Mis-specification effects could be important in 2-dimensional applications and more generally when the probability of identity is more dependent on the total size of the population than in a linear habitat.

Finally, the variation in local diversity in KAM versus SMMs is at most that resulting from a 2-fold variation in mutation rate (Rousset 1996), so one could expect the mutation model to have little impact on estimator performance beyond an at most 2-fold effect on mutation rate estimation, which is indeed what was observed when stepwise mutation data were analyzed under a KAM statistical model.

Conclusion

The present work has shown that ML can be applied to allelic type data from moderately large networks of populations. Maximum PAC likelihood is of potential utility for

Downloaded from https://academic.oup.com/mbe/article/24/12/2730/978152 by Bibliothèque Universitaire de médecine - Nîmes user on 16 June 2021

larger networks. Its performance was practically identical to that of ML estimation and even superior in most cases for an identical computation effort. The current implementation effectively allows ML analyses of systems of $n_m = 10$ demes in a few hours, and maximum PAC likelihood analyses of larger arrays (up to 200 demes in this study) can yield reasonably accurate estimates within a week. When the true number of demes is unknown, the assessment of performance yields mixed results. The number of demes that has to be considered in the statistical model to achieve good performance depends on the scale of dispersal and the mutation rate, which may limit the range of realistic applications. Mis-specification biases for mutation rates are relatively easily understood but less so for dispersal parameters. The composite parameter $D\sigma^2$ was relatively little affected by mis-specification of the number of demes, but it may be difficult to overcome mis-specification biases in the estimation of other dispersal parameters.

Acknowledgments

This study was made possible by access first to the computing facilities of the CINES (Montpellier, France), then to a PC cluster of the University of Montpellier 2, and finally to the ISEM cluster. We thank J.-B. Ferdy for substantial help in using this cluster, as well as V. Ranwez, K. Belkhir, and J. Maizi. The MNHN cluster was also used. We thank J.-M. Cornuet for access to his unpublished work. This is publication ISEM 07-119.

Literature Cited

Arbogast BS, Edwards SV, Wakeley J, Beerli P, Slowinski JB. 2002. Estimating divergence times from molecular data on phylogenetic and population genetic timescales. Ann Rev Ecol Syst. 33:707–740.

Bahlo M, Griffiths RC. 2000. Inference from gene trees in a subdivided population. Theor Popul Biol. 57:79–95.

Barton NH, Gale KS. 1993. Genetic analysis of hybrid zones. In: Harrison RG, editor. Hybrid zones and the evolutionary process. Oxford: Oxford University Press. p. 13–45.

Beerli P. 2004. Effect of unsampled populations on the estimation of population sizes and migration rates between sampled populations. Mol Ecol. 13:827–836.

Beerli P. 2006. Comparison of Bayesian and maximum-likelihood inference of population genetic parameters. Bioinformatics. 22:341–345.

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

Beerli P, Felsenstein J. 2001. Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. Proc Natl Acad Sci USA. 98:4563–4568.

Broquet T, Johnson CA, Petit É, Burel F, Fryxell JM. 2006. Dispersal kurtosis and genetic structure in the American marten, *Martes americana*. Mol Ecol. 15:1689–1697.

Cornuet JM, Beaumont MA. 2007. A note on the accuracy of PAC-likelihood inference with microsatellite data. Theor Popul Biol. 71:12–19.

Cox DR, Hinkley DV. 1974. Theoretical statistics. London: Chapman & Hall.

Cox JT. 1989. Coalescing random walks and voter model consensus times on the torus in $\mathbb{Z}^d$. Ann Probab. 17:1333–1366.

Cox JT, Durrett R. 2002. The stepping stone model: new formulas expose old myths. Ann Appl Probab. 12:1348–1377.

Cressie NAC. 1993. Statistics for spatial data. New York: Wiley.

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

de Iorio M, Griffiths RC. 2004a. Importance sampling on coalescent histories. Adv Appl Probab. 36:417–433.

de Iorio M, Griffiths RC. 2004b. Importance sampling on coalescent histories. II. Subdivided population models. Adv Appl Probab. 36:434–454.

de Iorio M, Griffiths RC, Leblois R, Rousset F. 2005. Stepwise mutation likelihood computation by sequential importance sampling in subdivided population models. Theor Popul Biol. 68:41–53.

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

Ellegren H. 2000. Microsatellite mutations in the germline: implications for evolutionary inference. Trends Genet. 16:551–558.

Estoup A, Rousset F, Michalakis Y, Cornuet JM, Adriamanga M, Guyomard R. 1998. Comparative analysis of microsatellite and allozyme markers: a case study investigating microgeographic differentiation in brown trout (*Salmo trutta*). Mol Ecol. 7:339–353.

Fearnhead P, Donnelly P. 2001. Estimating recombination rates from population genetic data. Genetics. 159:1299–1318.

Fenster CB, Vekemans X, Hardy OJ. 2003. Quantifying gene flow from spatial genetic structure data in a metapopulation of *Chamaecrista fasciculata* (Leguminosae). Evolution. 57:995–1007.

Fields Development Team. 2006. Fields: tools for spatial data. Boulder (CO): National Center for Atmospheric Research. http://www.cgd.ucar.edu/software/fields.

Golub GH, van Loan CF. 1996. Matrix computations. Baltimore (MD): John Hopkins University Press. 3rd ed.

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

Gusmão L, Sánchez-Diz P, Calafell F, et al. (42 co-authors). 2005. Mutation rates at Y chromosome specific microsatellites. Hum Mutat. 26:520–528.

Herbots HM. 1997. The structured coalescent. In: Donnelly P, Tavaré S, editors. Progress in population genetics and human evolution. New York: Springer-Verlag. pp. 231–255.

Kimura M, Weiss GH. 1964. The stepping stone model of population structure and the decrease of genetic correlation with distance. Genetics. 49:561–576.

Leblois R, Estoup A, Rousset F. 2003. Influence of mutational and sampling factors on the estimation of demographic parameters in a "continuous" population under isolation by distance. Mol Biol Evol. 20:491–502.

Leblois R, Rousset F, Estoup A. 2004. Influence of spatial and temporal heterogeneities on the estimation of demographic parameters in a continuous population using individual microsatellite data. Genetics. 166:1081–1092.

Li N, Stephens M. 2003. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics. 165:2213–2233. Correction: 167: 1039.

Malécot G. 1975. Heterozygosity and relationship in regularly subdivided populations. Theor Popul Biol. 8:212–241.

Maruyama T. 1970a. Effective number of alleles in a subdivided population. Theor Popul Biol. 1:273–306.

Downloaded from https://academic.oup.com/mbe/article/24/12/2730/978152 by Bibliothèque Universitaire de médecine - Nîmes user on 16 June 2021

Maruyama T. 1970b. Stepping stone models of finite length. Adv Appl Probab. 2:229–258.
Meng XL, Wong WH. 1996. Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. Stat Sin. 6:831–860.
Nagylaki T. 1974. The decay of genetic variability in geographically structured populations. Proc Natl Acad Sci USA. 71:2932–2936.
Nagylaki T. 1983. The robustness of neutral models of geographical variation. Theor Popul Biol. 24:268–294.
Notohara M. 1990. The coalescent and the genealogical process in geographically structured population. J Math Biol. 29:59–75.
Press WH, Flannery BP, Teukolsky SA, Vetterling WT. 1988. Numerical recipes in C. Cambridge: Cambridge University Press.
R Development Core Team. 2004. R: A language and environment for statistical computing. Vienna, Austria. R Foundation for Statistical Computing. [Internet]. [cited 2007 Oct 19]. http://www.r-project.org.
Raufaste N, Bonhomme F. 2000. Properties of bias and variance of two multiallelic estimators of F_{ST}. Theor Popul Biol. 57:285–296.
Rousset F. 1996. Equilibrium values of measures of population subdivision for stepwise mutation processes. Genetics. 142:1357–1362.
Rousset F. 1997. Genetic differentiation and estimation of gene flow from *F*-statistics under isolation by distance. Genetics. 145:1219–1228.
Rousset F. 2000. Genetic differentiation between individuals. J Evol Biol. 13:58–62.
Rousset F. 2004. Genetic structure and selection in subdivided populations. Princeton (NJ): Princeton University Press.
Rousset F. 2007. Inferences from spatial population genetics. In: Balding DJ, Bishop M, Cannings C, editors. Handbook of statistical genetics. Chichester, UK: Wiley. pp. 945–979.
RoyChoudhury A, Stephens M. 2007. Fast and accurate estimation of the population-scaled mutation rate, θ, from microsatellite genotype data. Genetics. 176:1363–1366.
Sacks J, Welch WJ, Mitchell TJ, Wynn HP. 1989. Design and analysis of computer experiments. Stat Sci. 4:409–435.
Sawyer S. 1977. Asymptotic properties of the equilibrium probability of identity in a geographically structured population. Adv Appl Probab. 9:268–282.
Slatkin M. 1987. The average number of sites separating DNA sequences drawn from a subdivided population. Theor Popul Biol. 32:42–49.
Slatkin M. 1991. Inbreeding coefficients and coalescence times. Genet Res. 58:167–175.
Slatkin M. 1993. Isolation by distance in equilibrium and non-equilibrium populations. Evolution. 47:264–279.
Slatkin M. 1994. Gene flow and population structure. In: Real LA, editor. Ecological Genetics. Princeton (NJ): Princeton University Press. pp. 3–17.
Slatkin M. 2005. Seeing ghosts: the effect of unsampled populations on migration rates estimated between sampled populations. Mol Ecol. 14:67–73.
Stephens M, Donnelly P. 2000. Inference in molecular population genetics (with discussion). J R Stat Soc. 62:605–655.
Strobeck C. 1987. Average number of nucleotide differences in a sample from a single subpopulation: a test for population subdivision. Genetics. 117:149–153.
Sumner J, Estoup A, Rousset F, Moritz C. 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.
Vigouroux Y, Jaqueth JS, Matsuoka Y, Smith OS, Beavis WD, Smith JSC, Doebley J. 2002. Rate and pattern of mutation at microsatellite loci in maize. Mol Biol Evol. 19:1251–1260.
Waples RS, Gaggiotti O. 2006. What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. Mol Ecol. 15:1419–1439.
Watts PC, Rousset F, Saccheri IJ, Leblois R, Kemp SJ, Thompson DJ. 2007. Compatible genetic and ecological estimates of dispersal rates in insect (*Coenagrion mercuriale*: Odonata: Zygoptera) populations: analysis of 'neighbourhood size' using a more precise estimator. Mol Ecol. 16:737–751.
Welch WJ, Buck RJ, Sachs J, Wynn HP, Mitchell TJ, Morris MD. 1992. Screening, prediction, and computer experiments. Technometrics. 34:15–25.
Whitlock MC, McCauley DE. 1999. Indirect measures of gene flow and migration: $F_{st} \neq 1/(4Nm + 1)$. Heredity. 82: 117–125.
Wilkins JF. 2004. A separation-of-timescales approach to the coalescent in a continuous population. Genetics. 168: 2227–2244.
Winters JB, Waser PM. 2003. Gene dispersal and outbreeding in a philopatric mammal. Mol Ecol. 12:2251–2259.
Wood JW, Smouse PE, Long JC. 1985. Sex-specific dispersal patterns in two human populations of highland New Guinea. Am Nat. 125:747–768.
Zähle I, Cox JT, Durrett R. 2005. The stepping stone model, II: genealogies and the infinite sites model. Ann Appl Probab. 15:671–699.

Marcy Uyenoyama, Associate Editor

Accepted September 18, 2007

Downloaded from https://academic.oup.com/mbe/article/24/12/2730/978152 by Bibliothèque Universitaire de médecine - Nîmes user on 16 June 2021