doi:10.1111/j.1558-5646.2009.00775.x

LIKELIHOOD-BASED INFERENCE IN ISOLATION-BY-DISTANCE MODELS USING THE SPATIAL DISTRIBUTION OF LOW-FREQUENCY ALLELES

John Novembre[1,2,3] and Montgomery Slatkin[2]

[1]Department of Ecology and Evolutionary Biology, Interdepartmental Program in Bioinformatics, University of California Los Angeles, Los Angeles, California 90095

[2]Department of Integrative Biology, University of California Berkeley, Berkeley, California 94720

[3]E-mail: jnovembre@ucla.edu

Received November 2, 2008
Accepted May 19, 2009

Estimating dispersal distances from population genetic data provides an important alternative to logistically taxing methods for directly observing dispersal. Although methods for estimating dispersal rates between a modest number of discrete demes are well developed, methods of inference applicable to "isolation-by-distance" models are much less established. Here, we present a method for estimating $\rho\sigma^2$, the product of population density (ρ) and the variance of the dispersal displacement distribution (σ^2). The method is based on the assumption that low-frequency alleles are identical by descent. Hence, the extent of geographic clustering of such alleles, relative to their frequency in the population, provides information about $\rho\sigma^2$. We show that a novel likelihood-based method can infer this composite parameter with a modest bias in a lattice model of isolation-by-distance. For calculating the likelihood, we use an importance sampling approach to average over the unobserved intraallelic genealogies, where the intraallelic genealogies are modeled as a pure birth process. The approach also leads to a likelihood-ratio test of isotropy of dispersal, that is, whether dispersal distances on two axes are different. We test the performance of our methods using simulations of new mutations in a lattice model and illustrate its use with a dataset from *Arabidopsis thaliana*.

KEY WORDS: Dispersal, importance sampling, intraallelic genealogy, isolation-by-distance, likelihood, low-frequency alleles.

Patterns of dispersal have long been recognized as important for evolutionary and ecological dynamics. Nevertheless, accurately quantifying patterns of dispersal via direct observation is difficult in most systems. Instead, patterns of genetic variation can be used to obtain indirect estimates of dispersal tendencies (Slatkin 1987). For models in which individuals are in distinct demes, a variety of indirect methods are available, for example, cladistic methods (Slatkin and Maddison 1989; Excoffier et al. 1992) and likelihood-based methods (Rannala and Hartigan 1995; Tufto et al. 1996; Beerli and Felsenstein 2001; Nielsen and Wakeley 2001; Iorio et al. 2005; Hey and Nielsen 2007).

In comparison, methods for estimating dispersal in continuous isolation-by-distance (CIBD) models are less well developed. In CIBD models, individuals are distributed across a continuous habitat and mate preferentially with nearby individuals. If an individual is born at a location (x_0, y_0) then the probability distribution of the location where it leaves its offspring is assumed to be a continuous distribution that is time invariant, identical for every individual, and of the form $k(x - x_0, y - y_0)$ where $k(\cdot, \cdot)$ is a bivariate distribution, such as a Gaussian distribution with zero mean, variances σ_x^2 and σ_y^2, and zero covariance. The standard deviations σ_x and σ_y are measures of the single-generation dispersal distance (but see Rousset 2004 for discussion of how this value should be interpreted). If $\sigma_x^2 = \sigma_y^2$, this parameterization implies dispersal is isotropic; in these cases, σ^2 is often defined such that $\sigma^2 = \sigma_x^2 = \sigma_y^2$. Finally, to avoid the formation of unrealistic spatial clumps of individuals (Felsenstein 1975), population density is typically constrained to be uniform across the habitat. One approach to imposing density regulation is to assume each individual occupies a single node on a large lattice. The resulting models are parameterized by σ^2 and ρ, the density of the population, although in most cases only the joint parameter $\rho\sigma^2$ is identifiable.

Various methods for estimating dispersal in CIBD models have been previously proposed, many of which are moment-based. Rousset (1997, 2000) estimate the product $\rho\sigma^2$ by regressing pairwise estimates of F-statistics on geographic distance. Although this method has been shown to be robust to various violations of the model assumptions (Leblois et al. 2003, 2004), one limitation is that linearity of the regression holds only over a limited, intermediate range of geographic distances that is not known prior to the study. Another moment-based estimator is that of Neigel et al. (1991), which estimates σ^2 by assessing the geographic dispersion of mtDNA haplotypes relative to their estimated time of most recent common ancestor (TMRCA). The method has the advantage of being independent of population density, but it has several drawbacks: it relies on data from only a single nonrecombining locus; it fails to account for uncertainty in the estimated TMRCA; and it assumes that a molecular clock holds. Wilkins and Wakeley (2002) provide a coalescent-based, method-of-moments approach applicable to sequence data, which uses only average pairwise sequence distances among samples.

More recently, efforts have been made to derive maximum-likelihood estimators of dispersal parameters in CIBD models (Rousset and Leblois 2007; Meligkotsidou and Fearnhead 2007). Maximum-likelihood estimators have the advantage of making full use of the data but the potential disadvantage of being difficult to implement. Significant computational challenges exist to implementing likelihood-based methods in population genetics. As a result these methods rely on computational approximations, such as importance sampling (IS), to estimate likelihoods (Stephens and Donnelly 2000).

Here, we propose a novel IS method for maximum-likelihood estimation of dispersal. The proposed method differs from existing likelihood methods in focusing on the geographic distribution of low-frequency alleles only. By restricting the analysis to low-frequency alleles, the computational problem of estimating the likelihood is more tractable than for the full data. The computational gains arise because we need to consider only the ancestry of the carriers of the low-frequency allele, rather than the ancestry of the whole sample. Further, the geographic distribution of low-frequency alleles contains most of the information about recent dispersal. Low-frequency alleles are typically descendants of recent mutations and are geographically clustered in the geographic area where the mutation occurred initially. For a given allele frequency, the tightness of this spatial clustering indicates the levels of dispersal. From a coalescent perspective, copies of a low-frequency allele have recent pairwise coalescent times, and previous studies show that restricted dispersal has a large effect on the distribution of recent pairwise coalescent times (Wilkins 2004; Fearnhead 2007). In addition, by focusing on low-frequency alleles, our method only requires that a population be at demographic equilibrium since the low-frequency alleles in question have arisen. This timescale is much shorter than the time to coalescence of the whole sample (Wiuf 2000; Slatkin 2003).

Our method estimates the likelihood of $\rho\sigma_x^2$ and $\rho\sigma_y^2$ for each locus separately and then combines information from independent loci. It can be applied to datasets with large numbers of independent loci such as single nucleotide polymorphism (SNP) datasets. The likelihood framework provides approximate confidence intervals and allows a likelihood-ratio test for equal dispersal in both directions.

Methods

MODEL

We begin by considering a population genetic sample of L independent biallelic loci from a diploid population of size N that is distributed over a finite area of size A with constant density $\rho = N/A$. Let $\mathbf{n} = (n_1, \ldots, n_L)$ be the number of chromosomes sampled at each of the L loci and let $\mathbf{j} = (j_1, \ldots, j_L)$ be the counts of the derived allele at each of the L loci. Let $\mathbf{X}_l$ be a $2 \times j_l$ matrix containing elements X_{ldk} that represent the dth-dimensional geographical coordinate of the kth copy of the derived allele at locus l, and let $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_L)$. Further, we assume that all copies of the derived allele are descendant from a single, unique mutation event (i.e., all copies of the mutation are identical-by-descent). Although we focus on biallelic loci in our presentation, our method is also to applicable multiallelic loci (e.g., loci that fit an infinite-alleles model of mutation), as long as each allele considered is at low frequency and identical-by-descent.

For alleles that are identical-by-descent, the history of the ancestral lineages of the allele can be summarized by a single intraallelic genealogy for each locus that can be decomposed into two portions. First, we specify a set of times describing events on the genealogy with time measured in units of N generations from the present day. Let $\mathbf{T}_l = (T_{l1}, \ldots, T_{lj_l})$ be a vector of times such that T_{l1} is the time at which the first copy of the derived allele arose at locus l and let T_{li} for $i \geq 2$ be the time points in the past at which the number of ancestral lineages in the intraallelic genealogy decreased from i to $i - 1$ as one looks backwards in time. Let $\mathbf{T} = (\mathbf{T}_1, \ldots, \mathbf{T}_L)$. The second portion of the genealogy is the tree topology describing the lineages involved in each of the $j_l - 1$ coalescent events on the intraallelic genealogy. Let $\mathbf{G} = (G_1, \ldots, G_L)$ be the graphs describing the tree topology of the intraallelic genealogy at each of the L loci.

To model dispersal, we consider a model with independent dispersal along two perpendicular, geographic axes. We consider the per-generation dispersal distribution along each axis to have a mean of 0 with a variance of σ_1^2 along one axis and a variance of σ_2^2 along the other. We assume the higher moments of the distribution are well behaved such that after some small duration of time s (measured in units of N generations) the location of a lineage starting at (x, y) is well approximated by a two-dimensional normal distribution with mean (x, y) and a variance–covariance matrix $[[sN\sigma_1^2, 0], [0, sN\sigma_2^2]]$. The precise time-point at which the approximation becomes valid depends on the timescale of coalescent events in the intraallelic genealogy. In the extreme case, where the allele frequency is so rare that intraallelic coalescent events occur nearly instantaneously, our assumption implies the dispersal distribution must be exactly normal. By assuming that the geographic location of a lineage will be a two-dimensional normal distribution, we are implicitly assuming that the positions of each lineage are following a Brownian motion and that boundary effects are negligible. The lack of boundary effects may be reasonable for low-frequency alleles found centrally within large habitats, because such alleles will likely not have dispersed widely enough to have encountered the boundaries of the habitat.

To denote the unobserved geographic position of each of the single mutation events that gave rise to the first copy of each derived allele, let $\mathbf{Z_l} = (Z_{l0}, Z_{l1})$ be the coordinates of the location at which the mutation event occurred for locus l and let $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_L)$. Due to the constant density of the population, the location at which a mutation occurs is equally likely across the whole habitat, so that the marginal distribution $P(\mathbf{Z}_l = \mathbf{z})$ equals $\frac{1}{A}$ for all $\mathbf{z}$ in the habitat.

Furthermore, we make the approximation that the intraallelic genealogy is independent of the geographical configuration of the lineages. This approximation is also used by Neigel et al. (1991) and Meligkotsidou and Fearnhead (2007) and implicitly assumes weak population density regulation. For our purposes, we note that, even in density-regulated populations, the approximation may be more accurate for low-frequency alleles. In panmictic populations, the number of copies of a low-frequency allele evolves approximately as a linear birth–death process (Slatkin and Rannala 1997), so that each copy of the allele leaves an independent number of descendant copies in the next generation. The extension that we assume here is that because copies of the low-frequency allele reproduce independently of each other, they will also reproduce independently of each other's geographic configuration. Given this approximation, the distribution of topologies and coalescent times for the intraallelic genealogy are described by the birth–death results obtained by Slatkin and Rannala (1997) for randomly mating populations. Specifically, the probability of the vector $\mathbf{T}_l$ can be found by considering $\mathbf{T}_l$ as j_l ordered samples from the density $h(t)$, where

$$h(t) = \frac{2n_l}{(2 + tn_l)^2}. \tag{1}$$

This distribution arises from the equations in Slatkin and Rannala (1997) by setting $f = \frac{n_l}{N}$ and measuring time in units of N generations (see supporting information for more detail). The density implies that:

$$P(\mathbf{T}_l = \{t_1, \ldots, t_{j_l}\}) = j_l! \prod_{i=1}^{j_l} \frac{2n_l}{(2 + t_{li} n_l)^2} \tag{2}$$

for all possible values $\{t_1, \ldots, t_{j_i}\}$. For the topologies G_l, Slatkin and Rannala's results provide the following simple distribution that reflects equiprobability of all labeled tree topologies:

$$P(G_l = g) = \frac{1}{\prod_{i=3}^{j_i} \binom{j_i}{2}} \tag{3}$$

for all possible g. To refer to the model, we use the acronym BBM, as our model is a type of branching Brownian motion.

LIKELIHOOD-BASED INFERENCE

For performing inference on the model described above, we focus on $\mathbf{X}$, the geographic locations of derived alleles for a set of loci. We are interested in inference on σ_1^2 and σ_2^2 although we can only infer the value of these parameters jointly with ρ; thus the identifiable parameters of the model are $\rho\sigma_1^2$ and $\rho\sigma_2^2$. We chose to assume that the area A was known, and so we instead are only interested in inference on $N\sigma_1^2$ and $N\sigma_2^2$, knowing that we can convert them to $\rho\sigma_1^2$ and $\rho\sigma_2^2$ using the known value of A. We define $\theta = (N\sigma_1^2, N\sigma_2^2)$ and use θ to refer to these parameters succinctly. We are also interested in the special case in which $N\sigma_1^2 = N\sigma_2^2$. In this case, there is only one identifiable parameter of the model, which we denote as $N\sigma^2$ or in some cases θ^* to be more compact. Finally, there are the unobserved quantities that are crucial components of the probability model. To summarize these "missing data" at each locus we let $\mathbf{M}_l = (\mathbf{T}_l, G_l, \mathbf{Z}_l)$.

Using the notation $P_\theta(\cdot)$ for the probability of an event given θ, the likelihood can then be written as

$$\begin{aligned} P_\theta(\mathbf{X}) &= \prod_{l=0}^{L} P_\theta(\mathbf{X}_l) \\ &= \prod_{l=0}^{L} \int P_\theta(\mathbf{X}_l \mid \mathbf{M}_l) P_\theta(\mathbf{M}_l)\, d\mathbf{M}_l. \end{aligned} \tag{4}$$

The integration over $\mathbf{M}_l$ is intractable analytically for realistic sample sizes because the space of $\mathbf{M}_l$ is the set of all possible topologies, all possible vectors of intraallelic coalescent times, and all possible geographic origins of the derived allele for locus l.

To approximate the integral over $\mathbf{M}_l$, we use a set of approximation techniques. We use a straightforward Monte Carlo approach to integrate over $\mathbf{T}_l$, an IS approach to integrate over G_l, and an approximate analytical integration for $\mathbf{Z}_l$. The details of each approach are described in supporting information.

From the perspective of how well the whole approximation algorithm performs, the critical part of the algorithm is the IS over G_l. Here, we propose an IS distribution, $P^*(G_l)$ that proceeds by randomly constructing a tree sequentially backwards in time such that topologies that join geographically proximal lineages are favored. Our distribution $P^*(G_l)$ takes two parameters, θ_0 and H. The parameter θ_0 defines a "driving value" of θ, such that the IS distribution will perform best when θ has a value close to θ_0. In practice, we use a single set of m simulated values from $P^*(G_l)$ to evaluate the approximation to equation $P_\theta(\mathbf{X})$ across a range of values of θ. This allows for significant computational speed-ups because we only need to simulate from $P^*(G_l)$ once to calculate a series of points on the likelihood surface around θ_0. The parameter H defines the extent to which geographical proximity influences the sampled topologies. H can be thought of as a "heat" parameter in that as its value increases, the entropy of the importance sampling distribution increases. More specifically, a value of $H = 1$ favors topologies in a close proportion to the contribution the topology will make to the calculation of $P_\theta(\mathbf{X}_l \mid G_l = g_i)$ whereas larger values sample topologies more uniformly. The use of H is designed so that as H approaches ∞, the importance sampler $P^*(G_l)$ will converge on the straightforward Monte Carlo sampler $P(G_l)$. The roles of the H and θ_0 parameters are described in more detail in the supporting information.

Finally, using the approximations to the likelihood, we employ standard optimization routines from the GNU scientific library to maximize the likelihood with respect to θ. We denote the maximum-likelihood estimate (MLE) of θ as $\widehat{\theta} = (\widehat{N\sigma_1^2}, \widehat{N\sigma_2^2})$ and the associated likelihood as $L(\widehat{\theta})$. We also maximize the likelihood for the constrained model in which dispersal is isotropic so that $\sigma_1^2 = \sigma_2^2$. The MLE for the constrained case is denoted as $\widehat{\theta^*} = \widehat{N\sigma^2}$ with likelihood $L(\widehat{\theta^*})$. Given $L(\widehat{\theta})$ and $L(\widehat{\theta^*})$, we can compute the likelihood-ratio test statistic λ for the null hypothesis that $\sigma_1^2 = \sigma_2^2$ as $\lambda = L(\widehat{\theta^*})/L(\widehat{\theta})$.

PERFORMANCE EVALUATION

To formally evaluate the performance of the likelihood-based inference, we take a two-part approach. In both cases, we focus mainly on characterizing the sampling distributions of $\widehat{N\sigma_1^2}$, $\widehat{N\sigma_2^2}$, $\widehat{N\sigma^2}$, and λ because it is the sampling behavior of these statistics that is most relevant to biological applications.

First, we evaluate the performance of our estimation method on data simulated under the same BBM model that is used to define the likelihood function. This step allows us to assess the performance of the algorithm for numerically approximating the likelihood function and producing estimates of θ. Given that the model underlying the likelihood approach is identical to the model simulating the data, we expect that if the algorithm is performing well, we will have a well behaved sampling distribution for the statistics of interest, $\widehat{N\sigma_1^2}$, $\widehat{N\sigma_2^2}$, $\widehat{N\sigma^2}$, and λ.

The second step is to evaluate how the method performs on data from forward simulations from a model of individuals distributed on a lattice. The performance of the method on these simulations will be a result of how well the numerical approximation to the likelihood function performs as well as how accurately the BBM model that the likelihood function is based on summarizes the behavior of the lattice-based model.

For the performance evaluations, we fix the value of H to 2, unless stated otherwise. We also fix the value of θ_0 to twice the value of θ used to simulate data, unless stated otherwise. In practice, both the appropriate values of H and θ_0 will depend on the dataset in question (see Discussion).

Simulation of the BBM model

To simulate data under this model, we first fix the number of loci L, the sample sizes per locus $\mathbf{n}$, and the number of derived alleles observed per locus $\mathbf{j}$. We then perform the following steps for each locus l:

1. Draw a topology from the distribution defined by $P(G_l)$ (eq. 3).
2. Draw a vector of times from the distribution defined by $P(\mathbf{T}_l)$ (eq. 2). See the "Monte Carlo integration over $\mathbf{T}_l$" section of the supporting information for more detail on how to simulate from $P(\mathbf{T}_l)$.
3. Assuming the mutation occurs at a geographic location (0, 0), simulate two independent Brownian motions along the intraallelic genealogy defined by G_l and $\mathbf{T}_l$. The resulting set of geographic locations for the lineages at the present day is stored as a simulated value of $\mathbf{X}_l$.

Simulation of alleles in a finite lattice model

To test our method using an alternative model of dispersal in a continuously distributed population, we simulated from a lattice model in which one individual is at each point in a large lattice. Although this model is still highly idealized, it includes density regulation of the individuals, yet it is simple enough that we could generate sufficient sample data with which to test our method. An alternative approach for simulating from a model with density regulation is the coalsecent-based algorithm of Wilkins and Wakeley (Wilkins and Wakeley 2002; Wilkins 2004).

We assumed a square lattice of $(2l + 1) \times (2l + 1)$ diploid individuals (where l is an arbitrary nonnegative integer), and that each generation consisted of two steps, dispersal of an infinitely large migrant pool followed by a random sampling of alleles from that migrant pool at each lattice point. In each replicate, the population was initially fixed for allele a. Then, at $t = 0$, one of the copies of a in the individual at the center of the lattice mutated to A to create a heterozygote. Then, each copy of A contributed to the migrant pool at the lattice point d_x and d_y steps away in proportion to a discretized and truncated bivariate normal distribution with mean (0, 0), variances σ_x^2 and σ_y^2, and 0 covariance. We truncated the dispersal distribution for d_x and d_y larger than $3\sigma_x$ and $3\sigma_y$ to speed the computations. We also simulated dispersal according to a modified double-exponential distribution that has been motivated by seed dispersal data (Clark 1998): $k(d) \propto e^{-|\frac{d}{\alpha}|^c}$ where $d = \sqrt{d_x^2 + d_y^2}$, α is a scale parameter, and c is a shape parameter. For values of $c < 1$, the tails of this distribution are not exponentially bounded (i.e., the distribution is "fat-tailed"). For these simulations, we truncate the dispersal distribution for x and y larger than $10\sigma_x$ and $10\sigma_y$.

The frequency of A at location (x, y), $p_{x,y}$, is the sum of the contributions to the migrant pool at that location from all extant copies of A. To create the next generation of adults, we assumed individuals are composed of two alleles independently sampled with the frequency of A allele being $p_{x,y}$.

Each replicate continued until A was either lost or fixed. For each set of replicates, we specified a target number of copies, j. Whenever the number of copies of A was exactly j, the locations of those j copies were recorded. At the end of each replicate in which j copies were found at least once, one of the sets of locations was chosen randomly to be the result for that replicate. Replicates were continued until L replicates were obtained in which j copies of A were found at least once. Then, the results for that set of replicates were formatted for analysis by our IS program.

Evaluating a single run of the algorithm

A general property of IS algorithms is that their performance can be evaluated by inspecting the distribution of IS weights (Liu 2002). In particular, the variance of the IS weights is useful because in pathological cases, the weights will vary wildly so that the final approximation will be determined by a few large IS weights. A useful summary statistic based on the variance of the importance sample weights is the effective sample size (*ESS*). Letting $\mathbf{g} = (g_1, \ldots, g_m)$, the *ESS* can be defined as

$$ESS = \frac{m}{1 + \mathrm{Var}_{P^*(G_l)}(w(\mathbf{g}))}$$

where $w(\cdot)$ is defined in the supporting information. The *ESS* statistic can be interpreted as the effective number of independent samples from the target distribution $P_\theta(\mathbf{X}_l \mid G_l)P(G_l)$.

EXAMPLE APPLICATION TO *ARABIDOPSIS THALIANA*

To provide an example application of the method, we analyzed a dataset representing genetic variation from populations of *A. thaliana*. We use a subset of the data presented in Nordborg et al. (2005). Of the 96 accessions presented by Nordborg et al. (2005), we focus on a subset of 49 accessions from Europe (Fig. S2). The 49 accessions were chosen by first taking the subset of 76 accessions in Europe and then excluding accessions at random that represented multiple samples from a single geographic locale. As a result, the set of 49 accessions represent 49 unique geographic locations across Europe. This last fact is important for application of our method because of the assumption in our model that individuals are sampled randomly from across the habitat and obtaining multiple individuals from the same location is unlikely under random spatial sampling.

We next filter the sequence data to obtain sites that are biallelic. We assume minor alleles are derived and limit ourselves to a fixed low-frequency range [i.e., each has six copies of the minor allele segregating (which corresponds to an allele frequency of $6/(2 \cdot 49) \approx 6\%$]. The geographic locations of the low-frequency allele at each locus are used to define $\mathbf{X}$. Here, we present the results for a simple dataset of eight loci from chromosome 3, chosen to be well spaced along the chromosome.

Results

PERFORMANCE OF IMPORTANCE SAMPLING APPROXIMATION

Across a range of exploratory trial values, we found the IS algorithm decreases the Monte Carlo variance relative to using a straightforward Monte Carlo approach. In most cases, the IS algorithm outperforms the Monte Carlo sampler by providing estimates of the likelihood surface that are accurate and suffer from little Monte Carlo sampling error (Fig. 1); however in some rare cases the IS algorithm produces a likelihood surface that is a clear outlier from the majority of other IS replicates. Typically, these aberrant replicates are recognizable by having low values for the *ESS* statistic and/or IS weights with means that are not approximately 1. These rare replicates likely reflect cases in which the importance sampler samples a rare topology with a very large IS weight and so the resulting approximation to the likelihood is dominated by a single replicate. In most cases, increasing the value of H was found to decrease the occurrence of these aberrant runs, although the reduction comes at the cost of increased Monte Carlo variance among the remaining replicates.

Performance on data from the birth-process model

To assess the performance of the IS-based likelihood method, we simulated data under the BBM model that underlies the method. Rather than examining the likelihood surface itself, we focus on the properties of the estimators that would be used in an application to data. In particular, we are interested in the performance of the point-estimates $\widehat{N\sigma_1^2}$, $\widehat{N\sigma_2^2}$, $\widehat{N\sigma^2}$, their associated confidence

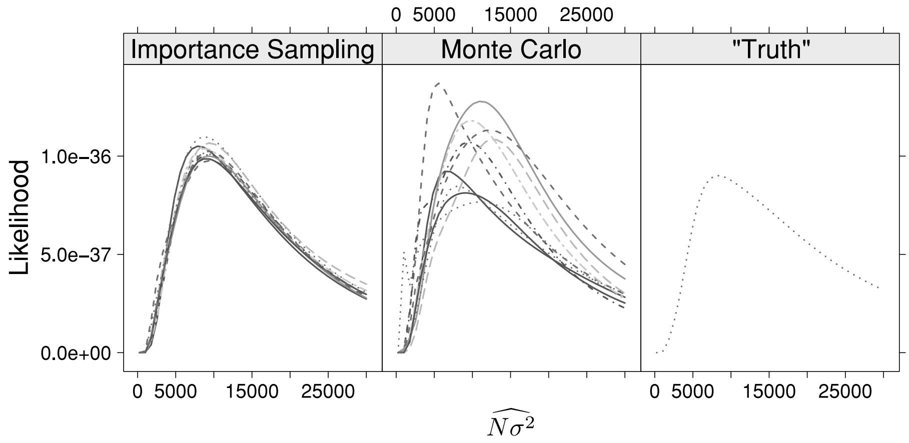

Figure 1. Example of the performance of the importance sampling algorithm relative to the straightforward Monte Carlo sampler. The left panel shows 10 replicate estimates of the θ* likelihood surface using 1000 iterations of the importance sampling algorithm. The central panel shows 10 replicate estimates using 3000 iterations of the random sampling algorithm. The right panel shows a close approximation to the true likelihood surface (obtained by 10 million replicates of the Monte Carlo sampler). The test case is a simulated sample from one locus with 12 minor alleles observed and with $\theta = (10^4, 10^4)$. For the importance sampling algorithm, $H = 1$ and $\theta_0 = (2 \times 10^4, 2 \times 10^4)$.

intervals, and the likelihood-ratio test based on λ. As mentioned above, this step allows us to investigate whether there are any obvious deficiencies in the IS algorithm and to gain insight on the performance of likelihood-based inference for this problem.

We found the sampling variance of the point estimates $\widehat{N\sigma_1^2}$, $\widehat{N\sigma_2^2}$, and $\widehat{N\sigma^2}$, decreases as either the number of low-frequency alleles observed per locus, j, or the total number of loci, L, increases (Fig. 2). For smaller sample sizes the sampling distributions are skewed toward higher values.

Despite the positive skew in the sampling distribution, the estimators appear to be unbiased. The mean of the estimators is consistently close to the true values used in the simulation, even for values of L and j that represent small sample sizes ($j = 3$, $L = 3$). The lack of bias is especially remarkable given the driving value of θ_0 was set to twice the true value of θ used in the simulation. Because $N\sigma_1^2$, $N\sigma_2^2$, and $N\sigma^2$ are scale parameters, proportionally similar results are found when simulations are performed with different values of $N\sigma_1^2$, $N\sigma_2^2$, and $N\sigma^2$ (e.g., the coefficients of variation for each estimator are constant, results not shown).

The sampling distribution of $\widehat{N\sigma^2}$ has a lower sampling variance than that of either $\widehat{N\sigma_1^2}$ or $\widehat{N\sigma_2^2}$, particularly for small sample sizes (Fig. 2 vs. Fig. S3). This result is not unexpected because for $\widehat{N\sigma^2}$ the geographic positions of alleles in both dimensions are informative, whereas for $\widehat{N\sigma_1^2}$ and $\widehat{N\sigma_2^2}$, only the positions of the alleles in a single dimension are informative.

Point estimates of the coverage probabilities for the 2 log-likelihood confidence intervals for $N\sigma_1^2$ and $N\sigma_2^2$ suggest the confidence intervals are slightly too narrow (e.g., Table 1). Across the conditions we investigated, the average coverage probability is 93.5% for both $N\sigma_1^2$ and $N\sigma_2^2$. No clear patterns with regard to how the coverage probability changes with the number of loci or number of copies of the low-frequency allele were observed, although asymptotic likelihood theory suggests the coverage probability will approach 95% as the number of loci increases.

The coverage probabilities for $N\sigma^2$ likewise show no clear relationship to L or the number of copies of the low-frequency allele observed at each locus; however one clear difference is that the confidence intervals are closer to the expected value of 95% (Table 2). The average coverage probability across the conditions we investigated was 95.2%. When we assess the performance of the likelihood ratio test, we find it is well-behaved in the sense that the false positive rate is generally close to 0.05, as intended (Table 3, average P-value is 0.0547 across conditions). The power of the likelihood-ratio test depends on $\log(\sigma_1^2/\sigma_2^2)$ and increases with j (Fig. 3).

These results show how when we simulate data from the same model underlying our method (the BBM model), we observe reasonable performance. The MLEs have low bias, confidence intervals show approximately the correct coverage, and a likelihood-ratio test that is well calibrated with respect to nominal P-values. These results indicate that the IS algorithm for computing the likelihood and subsequent methods for maximizing the likelihood are performing well. However, the BBM model is an approximation to the dynamics of a low-frequency allele that ignores population density regulation. To get a sense of how the method will perform on a model with density regulation, we turn to lattice-based simulations.

Performance on data from the lattice-based model

To assess performance of the method on lattice-based simulations, we again focus on the performance of the point estimates

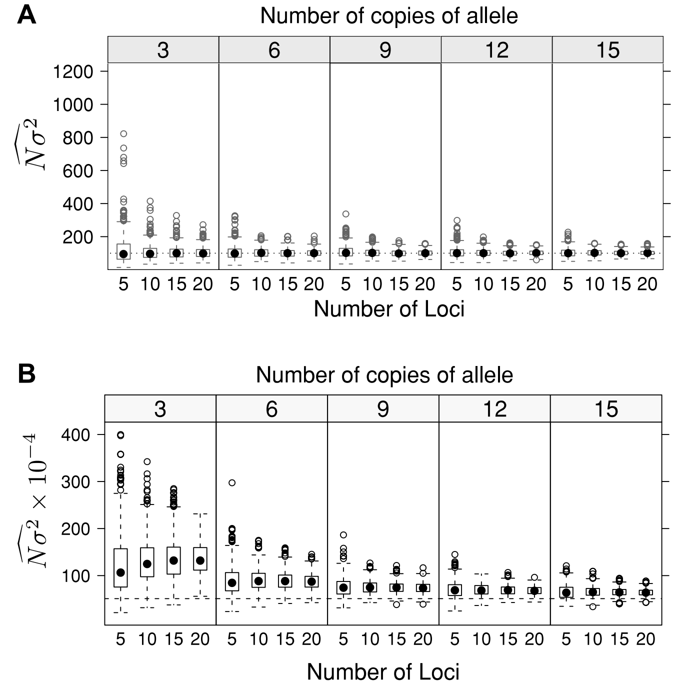

Figure 2. Box plot summaries of the sampling distribution of $\widehat{N\sigma^2}$. Summaries are plotted across a range for the number of loci and the number of copies of the minor allele observed, and the true value of $N\sigma^2$ is indicated by a horizontal dashed line in each panel. (A) Brownian birth process results: Each summary is based on the results of applying the importance sampling algorithm with $M = 2000$, $\theta_0 = (200, 200)$ and $H = 2$–500 datasets obtained by independent simulations from the birth process model with $\theta = (100, 100)$. (B) Lattice-model results: Each summary is based on the results of applying the importance sampling algorithm with $M = 20, 000$, $\theta_0 = (102 \times 10^4, 102 \times 10^4)$ and $H = 2$–500 datasets obtained by independent simulations from a 101 × 101 lattice with $\sigma_1 = \sigma_2 = 5$, such that $\theta = (51 \times 10^4, 51 \times 10^4)$.

Table 1. Point estimates for the coverage probabilities of the 2 log-likelihood confidence intervals for $N\sigma_1^2$.

Simulation model	L	No. of copies of allele				
		3	6	9	12	15
Brownian birth process	5	0.920	0.944	0.948	0.942	0.942
	10	0.938	0.946	0.924	0.928	0.934
	15	0.934	0.926	0.938	0.942	0.940
	20	0.932	0.932	0.916	0.930	0.930
Lattice model	5	0.800	0.836	0.876	0.888	0.924
	10	0.510	0.646	0.756	0.806	0.852
	15	0.356	0.524	0.658	0.732	0.802
	20	0.218	0.460	0.560	0.664	0.756

$\widehat{N\sigma_1^2}$, $\widehat{N\sigma_2^2}$, and $\widehat{N\sigma^2}$, their associated confidence intervals, and the likelihood-ratio test based on λ.

As with data from the BBM model, the sampling variance of the point estimates $\widehat{N\sigma_1^2}$, $\widehat{N\sigma_2^2}$, and $\widehat{N\sigma^2}$ decreases as either the number of loci or the number of low frequency alleles observed per locus increases. When the number of low-frequency alleles is low, we again see a skew toward high values. One important distinction is that for the lattice simulations, we generally observe a modest bias. For example, when we simulate alleles arising on a 101 × 101 lattice with $\sigma_1^2 = \sigma_2^2 = 25$, we find that the estimated values of $\widehat{N\sigma_1^2}$, $\widehat{N\sigma_2^2}$, and $\widehat{N\sigma^2}$ are each biased upwards (Fig. 2, Fig. S3). The bias decreases as the number of copies of the low-frequency allele (j) increases, but it appears to be unaffected by the number of loci (L). One consequence of this bias is that the coverage probabilities of the 95% confidence intervals are poorly behaved—for our simulations, often the lower confidence interval is above the true value of the parameter, resulting in very low coverage of the true value (Tables 1 and 2). Coverage increases as j increases, but as L increases the coverage becomes worse. Presumably as L increases one is getting tighter confidence intervals around a central value that is biased, resulting in poorer coverage; whereas when j increases the bias is reduced, hence increasing the coverage probabilities.

To further investigate bias, we conducted experiments to see how habitat size might affect the estimates. Because the BBM does not allow for edge effects, one might expect that as dispersal increases substantially relative to the scale of the habitat, the method might be biased toward underestimating the levels of dispersal. The downward bias would arise because the limited habitat size forces alleles to be more geographically clustered than they would be in an infinite habitat, and the method confuses this excess clustering for a signature of low levels of dispersal. Our simulations confirm this behavior (Fig. 4A). In contrast, when habitat sizes are large relative to the scale of dispersal, we find the estimates are directly proportional to the true underlying values (with a modest bias upwards, Fig. 4B).

We considered whether the driving value θ_0 could play a role in the upward bias. In nearly all the previous lattice model

Table 2. Point estimates for the coverage probabilities of the 2 log-likelihood confidence intervals for $N\sigma^2$.

Simulation model	L	No. of copies of allele				
		3	6	9	12	15
Brownian birth process	5	0.962	0.954	0.962	0.950	0.938
	10	0.970	0.976	0.960	0.938	0.954
	15	0.956	0.962	0.958	0.962	0.954
	20	0.958	0.952	0.942	0.938	0.914
Lattice model	5	0.766	0.804	0.864	0.856	0.910
	10	0.432	0.530	0.680	0.706	0.778
	15	0.232	0.370	0.492	0.576	0.698
	20	0.142	0.272	0.424	0.518	0.634

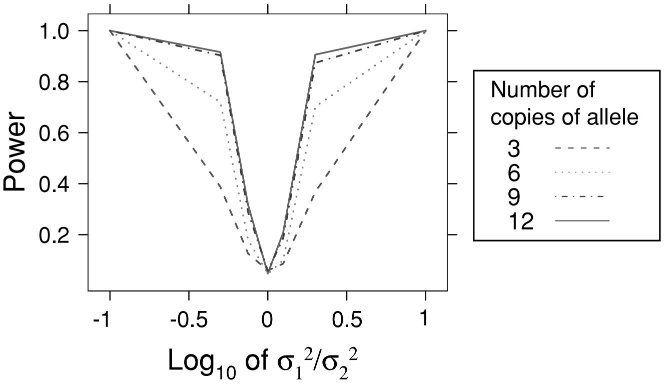

Figure 3. Power of the asymptotic likelihood-ratio test to detect departures from the null hypothesis that $\sigma_1^2 = \sigma_2^2$. The results are based on inference performed on simulated data from the Brownian birth process model where $N\sigma_1^2$ was fixed at 100 and $N\sigma_2^2$ was varied across the values 10, 50, 75, 100, 125, 200, 1000. All simulations were fixed at $L = 20$. Power is estimated from 500 replicate simulations. Similar power curves are found for the lattice model simulations (Fig. S4).

simulations the driving value was arbitrarily chosen to be twice the value of $\theta = (N\sigma_1^2, N\sigma_2^2)$ used to simulate the data (as we did in our simulated data from the BBM model). To assess whether the method is sensitive to this driving value and might bias $\widehat{\theta}$ toward θ_0, we studied how the distribution of $\widehat{N\sigma^2}$ changes as a function of θ_0 (Fig. S5). The results show that unless θ_0 is much less than θ (i.e., $\log_{10}(\theta_0/\theta) < -1$) the results are unaffected by the choice of θ_0. Importantly, even when $\theta_0 = \theta$ (and recalling the ideal choice of θ_0 is θ), we find an upward bias. Combined with the observation that the inference for the BBM model was not biased, this suggests the bias is not a result of properties of the IS approximation for calculating the likelihood.

We next investigated the performance of the likelihood-ratio test on the lattice model data (Table 3 and Fig. S4). We find no clear patterns in how the false positive rate depends on L and j but that overall the false-positive rates are close to the nominal P-value of 0.05 (the average false positive rate across our simulations conditions was 0.044). As in the simulations of the BBM model, the power of the method is proportional to $\log(\frac{\sigma_1^2}{\sigma_2^2})$ and the power increases with j.

We also considered how the method performs for data generated from an alternative dispersal distribution, with more or less kurtosis than the discretized normal distribution. We used the modified double-exponential distribution, in which the parameter c determines the kurtosis of the distribution. For reference, a normal distribution has a kurtosis value of 3; more fat-tailed (leptokurtic) distributions have higher values; and more narrow-tailed (playtkurtic) distributions have lower values. Our results show that as kurtosis increases, the estimates $\widehat{N\sigma^2}$ decrease, maintaining an upward bias, until the distribution becomes very leptokurtic ($c <$ 0.75, kurtosis > 12), at which point the estimates have a downward bias (Fig. 5).

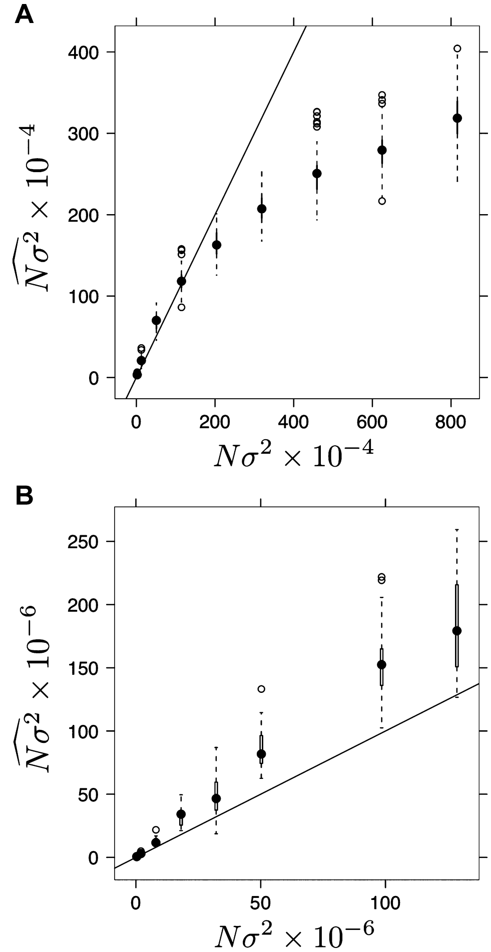

Figure 4. Lattice model results: Effect of habitat size on bias. Each panel shows the distribution of MLEs as a function of $N\sigma^2$. (A) Lattice of size 101 × 101. Hundred replicates per value of $N\sigma^2$. (B) Lattice of size 401 × 401. Twenty-five replicates per value of $N\sigma^2$. All simulations had $L = 10$, $j = 9$.

Finally, we note that for the lattice model one can convert estimates of $N\sigma^2$ to estimates of the "neighborhood size," $4\rho\pi\sigma^2$

Table 3. Point estimates for the false positive rate of the asymptotic likelihood-ratio test that $\sigma_1^2=\sigma_2^2$.

Simulation model	L	No. of copies of allele				
		3	6	9	12	15
Brownian birth	5	0.044	0.054	0.062	0.042	0.056
process	10	0.068	0.070	0.060	0.052	0.050
	15	0.072	0.054	0.044	0.060	0.054
	20	0.060	0.046	0.058	0.048	0.040
Lattice model	5	0.054	0.054	0.056	0.052	0.028
	10	0.058	0.054	0.018	0.036	0.028
	15	0.072	0.046	0.040	0.048	0.038
	20	0.071	0.070	0.038	0.056	0.046

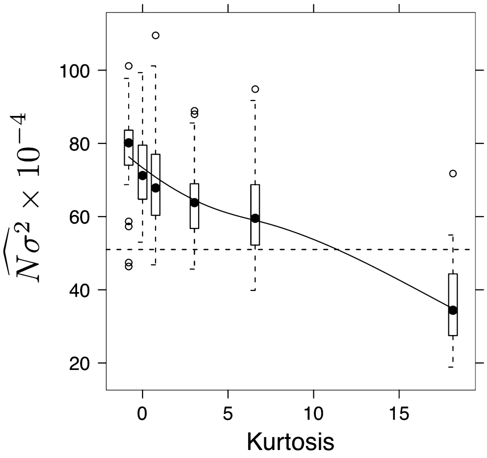

Figure 5. **Lattice model results: Effect of kurtosis on bias. The distribution of MLEs is plotted as a function of the kurtosis observed in the modified double-exponential dispersal distribution described in the text. Simulations used *c* = (0.5, 0.75, 1, 1.5, 2, 4) to produce the six different levels of kurtosis, and in each case α was chosen to result in a constant standard deviation of 5. Kurtosis was calculated based on the distribution produced after discretization and truncation (see main text). The lattice size was 101 × 101 so that the true underlying value of $N\sigma^2 = 51 \times 10^4$. Twenty-five replicates per value of *c*. All simulations had $L = 10$, $j = 9$.**

(Wright 1943) by dividing $\widehat{N\sigma^2}$ by the total area of the population (A) and multiplying by 4π. For example, suppose we obtain the estimate of 6×10^6 for $N\sigma^2$ on a 200 × 200 lattice. In this case, the neighborhood size would be estimated as 1885 individuals.

APPLICATION TO *A. THALIANA*

In the sample *Arabidopsis* dataset, the copies of the minor allele are distributed across a spatial area of hundreds of kilometers in each dimension (Fig. S6), although the exact extent varies largely from locus to locus. Locus 147 and Locus 7014 have the most compact distributions, whereas Locus 2123 has the most widespread. The variability from locus to locus in the distribution of the minor allele translates to high variability in the likelihood surfaces for $N\sigma^2$ at each locus (Fig. 6). The curvature of each individual likelihood surface and the variability of the likelihood surface among loci make obvious that precise estimation of $N\sigma^2$ is difficult using single loci. The joint likelihood curve (Fig. 7) has a much more narrow confidence interval ([3.1×10^6, 11.27×10^6]) than any individual locus and an MLE of 5.9×10^6. Assuming $N = 50{,}000$ the corresponding value of σ would be 10 km. The likelihood surface for $N\sigma_1^2$ and $N\sigma_2^2$ (Fig. 8) shows the MLE is close to the line defined by $\sigma_1^2 = \sigma_2^2$ and in turn there is no significant support to reject the null hypothesis that $\sigma_1^2 = \sigma_2^2$.

Discussion

The IS approach taken here works well for approximating the likelihood of $N\sigma_1^2$, $N\sigma_2^2$, and $N\sigma^2$ in the BBM model. Estimates of each parameter are unbiased, confidence intervals for each parameter have roughly the correct coverage probabilities, and the likelihood-ratio test of $\sigma_1^2 = \sigma_2^2$ has the expected false positive rate. These patterns are true even for small datasets, which is particularly noteworthy because the statistical properties of likelihood

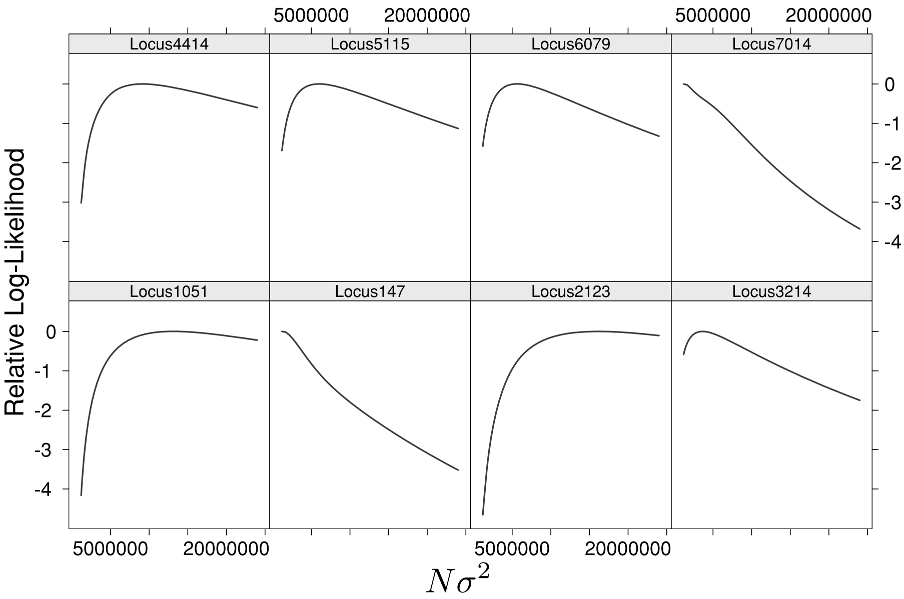

Figure 6. **The approximate log-likelihood curves for $N\sigma^2$ for each of the eight loci in the example dataset. The log-likelihood curves are translated so that the maximum value across the range of 2.5×10^6 km^2–2.5×10^7 km^2 is positioned at 0 on the *y*-axis. In some cases, the MLE lies outside of this range (e.g., Locus 147 and Locus 7014).**

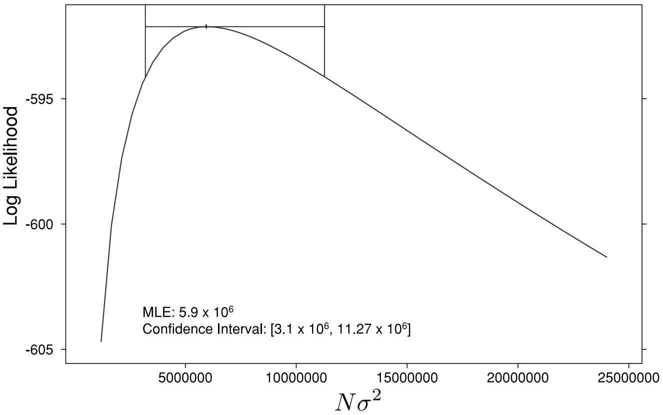

Figure 7. The joint likelihood curve across all loci for $N\sigma^2$. The horizontal line intersects the curve at the curve's maximum value and the two vertical lines demarcate the 2 log-likelihood support/confidence interval.

estimators are assured only as sample sizes become large. In addition, these patterns were observed across a large number of simulated datasets that were analyzed in a batch without any special adjustments to parameters that govern the implementation of the method (H, θ_0, and M). In addition, the computations proceed quickly (for a single locus with nine copies of the low-frequency allele, 20,000 iterations of the IS algorithm complete in a few seconds on a 3 GHz processor with 16GB RAM). The favorable performance of the inference method under these settings is evidence that the technical challenges of approximating the likelihood function of the BBM and finding its maxima are overcome by the IS algorithm used here.

Despite the robust performance of the IS algorithm observed in the BBM simulations, Monte Carlo techniques such as IS should always be used with care. The appropriate settings for H, θ_0, and M will necessarily depend on the dataset in question. Here, we found values of $H = 2$ work well and we found similar results for values of $H = 2 - 10$ (unpublished results, note: values > 10 not tested). For θ_0, we showed that the choice is not crucial as long as its value is not drastically smaller than the underlying θ for the dataset (Fig. S5). Because θ is not known a priori a reasonable approach suggested by Figure S5 is to run the algorithm iteratively using as θ_0 the $\widehat{\theta}$ value from the previous run until the estimate $\widehat{\theta}$ does not change. Finally, the larger the value of M the better the approximation will be, and it is useful to rerun the algorithm several times. The *ESS* statistic will indicate whether the estimates are suffering from large Monte Carlo sampling error if the value of M is too small.

To assess the effects of model misspecification, we also simulated data from a lattice model of dispersal. We found our method still performs well, in that the estimated parameters are proportional to their true values, especially for large habitat sizes where boundary effects are minimized (Fig. 4). However, we do find a general upward bias—inferred average dispersal distances tend to be larger than the true values. Given the lack of bias for data simulated under the BBM model, we conclude that the bias is due to discrepancies between the dynamics in the lattice and BBM models, and not to the IS algorithm itself. The bias appears to be caused by the fact that low-frequency mutations are dispersing slightly farther than expected given the branch lengths of the intraalleleic genealogy. Or to restate the same conclusion, given the geographic locations, the branch lengths of the intraallelic genealogy tend to be somewhat shorter than expected under the linear birth–death model.

The observation of upwardly biased estimates of dispersal in a CIBD model is not unique to our study. Meligkotsidou and Fearnhead (2007) observe a similar bias, and both our methods use the same approximation; namely, the probability distribution for events in the genealogical process occurs independently of the locations of the sampled lineages. This approximation implicitly ignores population density regulation, which is a key feature of the lattice model. Coalescent-based models that explicitly include population density regulation (Barton and Depaulis 2002; Wilkins and Wakeley 2002; Wilkins 2004), are much more challenging computationally, and further work is necessary before they can be adapted for likelihood-based inference. A further consideration is that, in practice, the extent of density regulation may vary across species. For example, Meligkotsidou and Fearnhead (2007) found the performance of their estimator improved in data derived from populations that have been expanding, and we would expect a similar pattern to hold for our estimator.

We also assessed the effects of habitat size and the kurtosis of the dispersal distribution. If the habitat size is small relative to $N\sigma^2$, dispersal parameters will be underestimated. This problem arises because of boundary effects, whereby alleles are being constrained in how far they can disperse, an aspect that is not

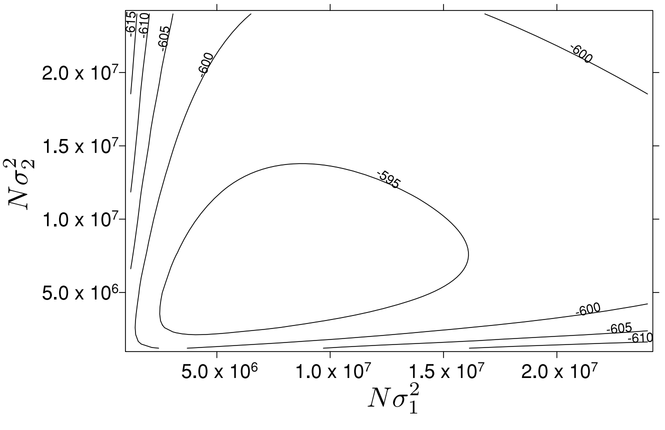

Figure 8. The joint likelihood surface across all loci for $N\sigma_1^2$ and $N\sigma_2^2$.

captured by our model. In practice, this suggests that loci with low-frequency alleles that are found near the edges of a habitat should be excluded from datasets prior to applying our approach, or at least analyzed separately with the intention of quantifying boundary effects. We also found that if the dispersal distribution is strongly leptokurtic, dispersal parameters may be underestimated. This perhaps occurs because the long-range migration events that make a dispersal distribution fat-tailed are unlikely to have been sampled in the time because a low-frequency allele arose by mutation. As a result, low-frequency alleles are more clumped than they would be if dispersal distances were normally distributed. Although the inference algorithm could in principle be based on a different dispersal distribution, that would increase the computational cost because our approach exploits properties of the normal distribution for the peeling algorithm (see supporting information).

Given these considerations, we can tentatively conclude that barring boundary effects and fat-tailed dispersal distributions, our method will often overestimate the dispersal parameter, and thus it provides an upper bound for the average dispersal distances. There are many complicating factors that we have not addressed, including misspecification of the derived allele, density-dependent dispersal, directional bias in dispersal, and nonrandom spatial sampling, that will affect the results from using our method, and its true accuracy cannot be established without detailed modeling of dispersal in a study species. Nevertheless, our method does provide a computationally feasible approach for estimation based on the dispersal of low frequency, relatively young mutations, and those estimates are at least of the right order of magnitude for the models of dispersal considered here.

ACKNOWLEDGMENTS

The authors wish to thank Eric Anderson for helpful discussions regarding this work. This work was funded in part by the National Institutes of Health (NIH-GM-40282) (MS) and a Howard Hughes Pre-doctoral fellowship (JN).

LITERATURE CITED

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

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

Clark, J. S. 1998. Why trees migrate so fast: confronting theory with dispersal biology and the paleorecord. Am. Nat. 152:204–224.

Excoffier, L., P. E. Smouse, and J. M. Quattro. 1992. Analysis of molecular variance inferred from metric distances among dna haplotypes—application to human mitochondrial-DNA restriction data. Genetics 131:479–491.

Fearnhead, P. 2007. On the choice of genetic distance in spatial-genetic studies. Genetics 177:427–434.

Felsenstein, J. 1973. Maximum-likelihood estimation of evolutionary trees from continuous characters. Am. J. Hum. Genet. 25:471–492.

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

———. 1981. Evolutionary trees from gene frequencies and quantitative characters: finding maximum likelihood estimates. Evolution 35:1229–1242.

Hey, J., and R. Nielsen. 2007. Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. Proc. Natl. Acad. Sci. USA 104:2785–2790.

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

Leblois, R., A. Estoup, and F. Rousset. 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., F. Rousset, and A. Estoup. 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.

Liu, J. S. 2002. Monte Carlo strategies in scientific computing. Springer, New York.

Meligkotsidou, L., and P. Fearnhead. 2007. Postprocessing of genealogical trees. Genetics 177:347–358.

Neigel, J. E., R. M. Ball, and J. C. Avise. 1991. Estimation of single generation migration distances from geographic-variation in animal mitochondrial-DNA. Evolution 45:423–432.

Nielsen, R., and J. Wakeley. 2001. Distinguishing migration from isolation: A Markov chain Monte Carlo approach. Genetics 158:885–896.

Nordborg, M., T. T. Hu, Y. Ishino, J. Jhaveri, C. Toomajian, et al. 2005. The pattern of polymorphism in *arabidopsis thaliana*. PLoS Biol 3:e196.

Rannala, B., and J. A. Hartigan. 1995. Identity by descent in island-mainland populations. Genetics 139:429–437.

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

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

———. 2004. Genetic structure and selection in subdivided populations (MPB-40) (Monographs in Population Biology). Princeton Univ. Press, Princeton, NJ.

Rousset, F., and R. Leblois. 2007. Likelihood and approximate likelihood analyses of genetic structure in a linear habitat: performance and robustness to model misspecification. Mol. Biol. Evol. 24:2730–2745.

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

———. 2002. A vectorized method of importance sampling with applications to models of mutation and migration. Theor. Popul. Biol. 62:339–348.

———. 2003. The Age of Alleles in *Modern Developments in Theoretical Population Genetics: The Legacy of Gustave Malecot'*. (edited by Slatkin, M. and Veuille, M.) Oxford Univ. Press, Oxford, pp. 233–260.

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

Slatkin, M., and B. Rannala. 1997. Estimating the age of alleles by use of intraallelic variability. Am. J. Hum. Genet. 60:447–458.

Stephens, M., and P. Donnelly. 2000. Inference in molecular population genetics. J. R. Stat. Soc. B 62:605–635.

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

Wilkins, J. F. 2004. A separation-of-timescales approach to the coalescent in a continuous population. Genetics 168:2227–2244.

Wilkins, J. F., and J. Wakeley. 2002. The coalescent in a continuous, finite, linear population. Genetics 161:873–888.

Wiuf, C. 2000. On the genealogy of a sample of neutral rare alleles. Theor. Popul. Biol. 58:61–75.

Wright, S. 1943. Isolation by distance. Genetics 28:114–138.

Associate Editor: D. Posada

Supporting Information

The following supporting information is available for this article:

Figure S1. Example intra-allelic genealogy.

Figure S2. Geographic distribution of samples from *Arabidopsis thaliana*.

Figure S3. Box-plot summaries of the sampling distribution of $\widehat{N\sigma_1^2}$.

Figure S4. Lattice model results: Power of the asymptotic likelihood-ratio test to detect departures from the null hypothesis that $\sigma_1^2 = \sigma_2^2$.

Figure S5. Lattice model results: Effect of driving value parameter on bias.

Figure S6. The geographic positions of the minor alleles at each of the eight loci in the example dataset.

Supporting Information may be found in the online version of this article.
(This link will take you to the article abstract).

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.