DOI: 10.1534/genetics.103.024182

Multilocus Methods for Estimating Population Sizes, Migration Rates and Divergence Time, With Applications to the Divergence of *Drosophila pseudoobscura* and *D. persimilis*

Jody Hey*,[1] and Rasmus Nielsen†

**Department of Genetics, Rutgers, State University of New Jersey, Piscataway, New Jersey 08854 and †Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, New York 14853*

Manuscript received November 3, 2003
Accepted for publication February 16, 2004

ABSTRACT

The genetic study of diverging, closely related populations is required for basic questions on demography and speciation, as well as for biodiversity and conservation research. However, it is often unclear whether divergence is due simply to separation or whether populations have also experienced gene flow. These questions can be addressed with a full model of population separation with gene flow, by applying a Markov chain Monte Carlo method for estimating the posterior probability distribution of model parameters. We have generalized this method and made it applicable to data from multiple unlinked loci. These loci can vary in their modes of inheritance, and inheritance scalars can be implemented either as constants or as parameters to be estimated. By treating inheritance scalars as parameters it is also possible to address variation among loci in the impact via linkage of recurrent selective sweeps or background selection. These methods are applied to a large multilocus data set from *Drosophila pseudoobscura* and *D. persimilis*. The species are estimated to have diverged ~500,000 years ago. Several loci have nonzero estimates of gene flow since the initial separation of the species, with considerable variation in gene flow estimates among loci, in both directions between the species.

FOR many species, existence is a complicated mix of multiple populations that are dynamic in size, location, and levels of gene exchange. Sometimes an individual population diverges from others to a sufficient degree that evolution thereafter proceeds largely independently. What is it about these populations that go on to become new species that sets them apart from others that do not? Clearly the accumulation of endemic mutations under an extended period of allopatry can enable this process (Dobzhansky 1936; Muller 1940; Mayr 1942). But if populations are not completely separated, then divergence entails a competition between unifying and diversifying genetic processes. Genetic drift will enhance divergence, while gene flow will retard it. Natural selection, enabled by gene flow, can act to reduce divergence if selective sweeps pass across populations. But on the other hand, natural selection that leads to population-specific selective sweeps or that acts otherwise to retard gene exchange can promote divergence.

One classic finding on gene flow and genetic drift that helps to focus our intuition is that a modest level of gene flow (one gene copy per generation, on average) will prevent substantial divergence at a locus (Wright 1931). This point also begets a key corollary: that only a modest level of natural selection against gene flow may be sufficient to enable divergence. But despite these guidelines, empirical questions on population divergence can be fairly intractable. Perhaps the clearest case of this is the common situation when a measure of differentiation has been obtained for a pair of populations, such as a genetic distance or an estimate of Wright's F_{st} (Wright 1922). Given such a number, one can then estimate how long ago the populations diverged (assuming no gene flow), or one can estimate the gene flow rate, assuming the populations are at equilibrium and have been separated and are exchanging genes at that rate for a very long period of time. In short, one can fit a model that assumes the populations will become increasingly divergent (model I, for isolation), or one can fit a model that assumes the populations will never diverge more than they have already, because of gene flow (model M, for migration). Neither one is of much use if the goal is to develop a full picture that includes estimates of separation time *and* gene flow.

Investigators have considered nonequilibrium models of population splitting; with gene flow, however, there are significant challenges (Latter 1973; Takahata and Slatkin 1990; Takahata 1995; Wakeley 1996b; Wakeley and Hey 1998). The problem is that the two different models (I and M) can lead to similar gene tree topologies and can be fit equally well to most kinds of data summaries (Slatkin and Maddison 1989; Takahata and Slatkin 1990). However, the two models do

[1]*Corresponding author:* Department of Genetics, Rutgers, State University of New Jersey, 604 Allison Rd., Piscataway, NJ 08854-8082. E-mail: hey@biology.rutgers.edu

not have identical predictions and data summaries and likelihood methods that exploit these differences can be used to distinguish them (Wakeley 1996a; Nielsen and Slatkin 2000; Nielsen and Wakeley 2001).

One fairly complete approach is to include both isolation and migration [the "isolation with migration" (IM) model] and to apply a probabilistic method for fitting the model to a data set. Nielsen and Wakeley (2001) developed a likelihood/Bayesian framework for fitting a six-parameter version of the IM model to data from a single, nonrecombining locus drawn from two populations or closely related species. At the heart of the method is an expression for the distribution of model parameters Θ, given the data, X:

$$f(\Theta|X) = cf(\Theta)\int_G f(X|G, \Theta)f(G|\Theta)\,dG. \qquad (1)$$

Here, $c = 1/\Pr(X)$, the inverse of the probability of the data. In the course of calculations c is treated as a constant to ensure that the total probability for all values of Θ sums to one. $f(\Theta)$ is the prior probability density function of the parameters and $f(G|\Theta)$ is the prior distribution of genealogies (rooted ultrametric trees with branch lengths). In this framework, inferences regarding Θ are based on the posterior distribution of Θ, $f(\Theta|X)$.

The most challenging aspect of (1) hinges on the unknown genealogy (*i.e.*, the gene tree) that underlies the data. For any particular genealogy, G, and set of parameters, Θ, it is possible to evaluate $f(G|\Theta)$ using coalescent theory (Kingman 1982a,b; Hudson 1983; Tavare 1984). Also for a given mutation model it is possible to calculate the probability of the data, for a given genealogy and set of parameters values, $f(X|G, \Theta)$. The genealogy for each locus consists of a tree with a branching pattern (topology), in which all of the DNA sequences in the data set for that locus are represented at the tips. Each genealogy also includes values for all of the branch lengths, as well as times of migration events. In effect G is a nuisance parameter that must be integrated out to gain insight into the demographic parameters. Nielsen and Wakeley (2001) implemented a Markov chain Monte Carlo (MCMC) approach that jointly approximates the integration over genealogies in the course of also approximating the full expression for $f(\Theta|X)$.

In principle the prior distribution of Θ can be set to reflect actual prior information regarding Θ; however, for most purposes $f(\Theta)$ is set to a constant value over a prescribed range of values. By setting this prior distribution to be uniform, $f(\Theta|X)$ is proportional to the likelihood of the parameters, given the data. Thus, for example, if (1) can be evaluated, then the mode of the posterior distribution provides a maximum-likelihood estimate of Θ.

The method of Nielsen and Wakeley (2001) was designed for data from just a single nonrecombining locus, and it can be slow to converge on the correct posterior densities for some data sets. Here we present some extensions of the method that alleviate these problems and greatly increase the applicability of the method. We demonstrate the methods using previously published data from *Drosophila pseudoobscura* and *D. persimilis*.

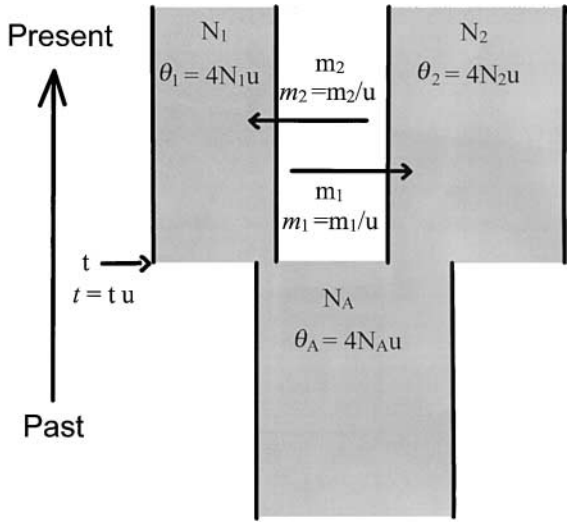


Figure 1.—The isolation with migration (IM) model is depicted with two parameters sets. The basic demographic parameters are constant effective population sizes ($\mathbf{N}_1$, $\mathbf{N}_2$, and $\mathbf{N}_A$), gene flow rates per gene copy per generation ($\mathbf{m}_1$ and $\mathbf{m}_2$), and the time of population splitting at **t** generations in the past. The second set of parameters is scaled by the neutral mutation rate **u**, and it is these parameters that are actually used in the model fitting.

MODEL

Consider a general IM model in which a population gives rise to two populations, after which there may be gene exchange between the two populations. This model has six major parameters: the population sizes of the three populations (for populations 1 and 2 and the ancestral population), two migration rates, and a time point at which the ancestral population gave rise to populations 1 and 2. In Figure 1, two versions of these parameter sets are shown. One includes the basic demographic parameters of population size, migration rate, and time of population splitting. In this framework, the genetic process of drift and mutation occurs on a timescale of generations. However, when we fit the model to genetic data we usually do not have direct access to a timescale that is in units of generations or years. For this reason the method (and others like it) must scale the parameters either by the rate of genetic drift or by the mutation rate.

In the present framework, each population is represented by a population mutation rate, $\theta = 4\mathbf{Nu}$, where **N** is the effective size of a population of diploid individuals, and **u** is the neutral mutation rate, rather than by population size. For migration, the rates are expressed as the

rate of migration for each gene copy, per mutation event, or $m = \mathbf{m}/\mathbf{u}$. In this way, the product of a population mutation rate and a migration rate is equal to the more familiar population migration rate $\mathbf{M} = 2\mathbf{N}\mathbf{m} = \theta m/2$. The time parameter also is expressed in terms of mutations, $t = \mathbf{t}\mathbf{u}$. If we wish, we can convert to a measure of time that is on the same scale as the process of genetic drift, $\mathbf{T} = \mathbf{t}/2\mathbf{N} = 2t/\theta$. Thus also shown in Figure 1 are parameters scaled to the neutral mutation rate, **u**. This parameterization differs from the original description of the method, which used $\theta_1 = 4\mathbf{N}_1\mathbf{u}$, $\mathbf{r} = \mathbf{N}_2/\mathbf{N}_1$, $\mathbf{a} = \mathbf{N}_A/\mathbf{N}_1$, $\mathbf{M}_1 = 2\mathbf{N}_1\mathbf{m}_1$, $\mathbf{M}_2 = 2\mathbf{N}_2\mathbf{m}_2$, and $\mathbf{T} = \mathbf{t}/2\mathbf{N}_1$ (Nielsen and Wakeley 2001). Throughout the article, quantities that are in units of individuals (**N**) or generations (**t**) or that are rates per generation (**u** and **m**) are in boldface type, whereas parameters that are used in the method, including demographic parameters that are scaled by the mutation rate, are expressed in italics (*e.g.*, m and t).

Multiple loci: A key assumption of the method is that the locus being studied has been evolving neutrally and that it has been drawn at random from all loci, with respect to genealogical history. In other words, the locus should not have been drawn in such a way that it is likely to have an atypical gene tree depth or to have experienced an atypical amount of gene flow. But even if these assumptions apply, different unlinked loci will vary widely in their histories. This normal stochastic variance among loci can be very great, and it is a major difficulty for phylogeographic studies that use just one locus (Hey and Machado 2003). In principle this can be overcome, and parameter estimates greatly improved, by extending the method to simultaneously include multiple loci.

The extension of the method to multiple independent (*i.e.*, effectively unlinked) loci is fairly straightforward, as the joint density function for the parameters can be expressed as a function of the product of the densities calculated for each individual locus. Similar to expression (1), the joint posterior density can be written as

$$f(\Theta|X_1, X_2, \ldots, X_n) = cf(\Theta)\prod_{i=1}^{n}\int_{G_i} f(X_i|G_i)f(G_i|\Theta)\,dG_i, \quad (2)$$

where Θ still refers to the vector of parameters of the model, X_i refers to the data for locus i, and G_i is the genealogy for locus i. As in the single-locus case, G_i is described by the topology of an ultrametric tree, its associated coalescence times, and the times of migrations on each branch of the tree. For notational convenience, from now on we use the notation $G = (G_1, G_2, \ldots)$ and $X = (X_1, X_2, \ldots)$.

Expression (2) cannot be solved analytically. However, it is possible to estimate the posterior probability density by simulating a Markov chain using the Metropolis-Hastings algorithm (see, *e.g.*, Gilks *et al.* 1996). The basic idea in this method is to simulate a Markov chain with state space on the set of all possible values of G and Θ and stationary distribution $f(G, \Theta|X)$. Then $f(\Theta|X)$ can be approximated by sampling values of Θ from this chain at stationarity. The simulation initiates with a starting set of parameter values and a starting set of genealogies that are consistent with the data and proceeds by iteratively updating each of the genealogies and the parameter values in turn according to the appropriate Metropolis-Hastings criterion that ensures that the Markov chain has the desired stationary distribution. Over the course of a long simulation a record is kept of the time that the chain spends at each of the possible values for each parameter. After a sufficiently long run, the distribution of residence times for a given parameter should be a good approximation of the marginal posterior density of that parameter.

Under this multilocus framework the general expression for the Metropolis-Hastings criterion seems fairly complex, with an update, from parameter values Θ and genealogies $G_1, G_2 \ldots G_n$ to Θ^* and $G_1^*, G_2^*, \ldots G_n^*$, accepted with probability

$$\min\left\{1, \prod_{i=1}^{n}\frac{f(X_i|\Theta^*, G_i^*)f(G_i^*|\Theta^*)f(\Theta^*)q[(\Theta^*, G_i^*) \rightarrow (\Theta, G_i)]}{f(X_i|\Theta, G_i)f(G_i|\Theta)f(\Theta)q[(\Theta, G_i) \rightarrow (\Theta^*, G_i^*)]}\right\} \quad (3)$$

[see expression (3) of Nielsen and Wakeley 2001]. However, for most parameters the quantity simplifies considerably, and in all cases where the parameter being updated affects the probability associated with more than one locus, the criterion is a product of terms given in Nielsen and Wakeley (2001).

To include multiple loci it is necessary to extend the parameter set. But since under the model all of the basic demographic parameters apply to all loci, the only additional parameters necessary are locus-specific mutation scalars. Thus for locus i, we let u_i represent the relative mutation rate for locus i such that the population mutation rate at that locus is θu_i. One way to implement such scalars is to pick one locus as a standard with a mutation rate scalar of 1 and to have the scalars for other loci vary as parameters to be estimated. However, this would cause parameter estimates to vary depending on the locus selected as the standard, and it might also be the case that convergence would be slow because of the strong correlation among parameters in the proposal kernel. An alternative that we have chosen is to let all loci have scalars that are free to vary, subject to the constraint that their products are equal to 1. For example, when there are two loci, there will be two mutation rate scalars, u_1 and u_2, that vary reciprocally so that at all times during the Markov chain simulation $u_1 = 1/u_2$ and vice versa. We use a log uniform prior on the inheritance scalars subject to the constraint $\Pi_{i=1}^{n} u_i = 1$. At the beginning of a Markov chain simulation with n loci, all n scalars are set to 1. Updates to these scalars are considered in turn along with updates

to the other parameters. For each update, two loci, i and j, are selected at random from the n loci and their scalars, u_i and u_j, are replaced by $u_i^* = du_i$ and $u_j^* = u_j/d$, respectively. d is drawn at random from a uniform log scale distribution such that u_i^* and u_j^* fall between $1/x$ and x (x is a very large number). If we envision there being n different population mutation rates for population 1, $\theta_{1,1}, \ldots, \theta_{1,n}$ with $\theta_{1,i} = \theta_1 u_i$, then the value of θ_1 is equal to the geometric mean of the individual locus values and remains constant before and after an update of the scalars. Because both the ratios of update densities and the ratios of the priors are proportional to $u_i^* u_j^*/(u_i u_j) = 1$, the Metropolis-Hastings acceptance probability is simply

$$\min\left\{1, \frac{f(X_i|u_i^*)f(X_j|u_j^*)}{f(X_i|u_i)f(X_j|u_j)}\right\}. \quad (4)$$

A benefit of mutation scalars implemented in this way is that they can be easily applied regardless of the mutation model and regardless of the length of individual loci. Thus, in general a locus represented by long DNA sequences will reveal more polymorphisms and greater divergence and correspondingly high values for the mutation scalar relative to a short locus that is included in the same analysis. If just two loci of identical underlying mutation rates per base pair, but of different sequence lengths, are included in an analysis, then we would expect the estimate of the mutation rate scalar of the longer locus to be greater than one and to be close to the reciprocal of the estimate of the shorter locus. In most applications described below, the infinite-sites model is used, but regardless of the mutation model the scalars can be applied to a mixture of loci of various mutation models, including the Hasegawa-Kishino-Yano (HKY) model (Hasegawa *et al.* 1985; Palsbøll *et al.* 2004) as well as the stepwise mutation model (Hey *et al.* 2004).

Inheritance scalars as constants: When multiple genes are studied it is often the case that different loci have different modes of inheritance (*e.g.*, autosomal, hemizygous, or sex limited). The issue raised by these cases is that the effective population size of nonautosomal loci is correspondingly reduced by their lower effective level of ploidy and number of carriers in the population. Thus it is common to multiply estimates of θ for a hemizygous locus by a factor of ¾, and a sex limited locus by a factor of 4, to bring them up to par with estimates for autosomal loci. These types of adjustments can be readily included within the model. For locus i with an expected effective number of gene copies of h_i, relative to an autosomal locus, the population mutation rate parameter during the Markov chain is set to θh_i. Thus at all points in the Markov chain involving calculations for locus i, and using θ_1, θ_2, and θ_A, the products of these parameters and h_i are used instead. The effect of this is to compress or stretch the distribution of θ by a factor of $1/h_i$. For example, if an estimated value is obtained for a single locus assuming $h = 1$, then the same data using an inheritance scalar of $h = ¼$ will return an estimated population mutation rate of four times that value.

Inheritance scalars as parameters: A difficulty with inheritance scalars is that their values are generally taken to be known directly as a consequence of the mode of inheritance. The usual assumption is that males and females each contribute equally to the effective population size of the population. An alternative approach to asserting a particular value for h is to allow the inheritance scalar to be a parameter in the model. Under such a framework, the value of h for each locus would be free to vary during the course of the Markov chain as a function of the data and the model, and posterior distributions would be returned for $h_1, h_2, \ldots h_n$ just as for the other parameters. There are two sorts of biological justifications for this. First, the assumptions regarding sex ratios and effective numbers of males and females that underlie the conventional values of these scalars may not hold. Second, and perhaps more importantly, there are other, selective reasons why the effective number of gene copies experienced by different loci may vary systematically among populations—over and above that variation caused by the mode of inheritance. Recurrent selective sweeps (Maynard Smith and Haigh 1974; Gillespie 2000) or background selection (Charlesworth *et al.* 1993) can cause a locus to steadily experience a reduced effective population size that is different from that of other loci.

When implementing inheritance scalars as parameters we are faced with a situation similar to that for the locus-specific mutation scalars. As in that case, we have used an updating scheme in which pairs of inheritance scalars are changed in such a way that the product of all inheritance scalars is 1. The Metropolis-Hastings criterion for updating the inheritance scalars for locus i and j, from h_i and h_j to dh_i and h_j/d, is

$$\min\left\{1, \frac{f(G_i|dh_i)f(G_j|h_j/d)}{f(G_i|h_i)f(G_j|h_j)}\right\}. \quad (5)$$

For equilibrium models, in which population sizes and migration rates are constant over an effectively infinite period of time, all branches on the genealogy will scale with both effective population size and mutation rate. In this situation, modifiers of effective population size (*i.e.*, inheritance scalars) and mutation rate (*i.e.*, mutation rate scalars) cannot be independently estimated (*i.e.*, the model is not identifiable when both inheritance and mutation scalars are free to vary as parameters). Thus, for example, with data from a single population and multiple loci that vary in polymorphism levels, one could estimate a set of mutation scalars (*i.e.*, one for each locus) or a set of inheritance parameters, but the two sets would be identical and one could not estimate both. However, when a model includes population splitting, the two sets of parameters become separa-

ble. Polymorphism within populations depends upon inheritance mode and mutation rate, while the divergence between populations depends directly upon mutation rate, but not directly upon the mode of inheritance. This is why some models of population structure must consider the mode of inheritance separately from mutation rate (Wang and Caballero 1999; Laporte and Charlesworth 2002) and why multilocus models of population splitting must include both types of scalars (Wang *et al.* 1997).

Because the two mutation and inheritance scalars both apply to the population mutation rates (θ_1, θ_2, and θ_A) the two types of parameters are expected to negatively covary to a large extent, particularly if the time of population splitting has been recent relative to the depth of gene trees within populations. Similarly gene flow between populations blurs the demarcation between variation that arises between populations, due to population splitting, and variation that arises within populations. It is likely that data sets drawn from populations that have had either very recent separation from an ancestral population or substantial gene flow since separation will not have the divergence necessary to support the estimation of both mutation rate scalars and inheritance scalars.

Metropolis coupling: A major difficulty of MCMC estimation of probability densities is not knowing the length of time needed for convergence (*i.e.*, for the simulated values to accurately approximate the true density). The method offers no guarantee that the chain will sufficiently traverse the state space in reasonable time, and there has been some debate on whether an investigator should run multiple chains and on how long chains should be to have some confidence that the results have converged on the correct answer (Gelman and Rubin 1992a,b; Geyer 1992a,b). Usually with single-locus data sets of total sample size <50, chains of 20 million steps prove sufficient for repeatable, albeit rough, point estimates, although considerably longer chains are needed for highly precise estimates. The same cannot be said of data sets with multiple loci, for which convergence may often be very slow. The fundamental problem is that for many loci, $f(\Theta|G, X)$ tends to be centered on a specific value of Θ (*i.e.*, the posterior density of the parameters is well determined, conditional on G). However, the unconditional posterior density, $f(\Theta|X)$, may nonetheless have a large variance. This seems especially to be an issue for the parameter t and leads to reduced mixing and slow convergence of the chain.

To offset this difficulty we have implemented a Metropolis-coupled version of the algorithm in which multiple chains are run simultaneously, with all chains but one having heated stationary distributions (Geyer 1991). These heated chains will not individually return the correct posterior distributions but they will explore the parameter space far more quickly than will the nonheated chain. Increased mixing in the nonheated chain is obtained by symmetrically swapping parameter and genealogy states between chains at rates determined by a Metropolis criterion that is a function of the difference in overall probabilities between the chains and the difference in heating values of the chains (Geyer 1991). For a simulation with Metropolis coupling among k chains, each chain will be approximately a fraction $1/k$ as long as a single chain run for the same length of time. The advantage gained is that the overall rate of mixing on the primary chain may be vastly improved. In practice we have found this method solves the difficulties of inadequate mixing that arise sometimes with data sets that include multiple loci.

Mutation models: In the original description, the method was limited to the infinite-sites mutation model (Kimura 1969). Recently it has been extended to the HKY model (Hasegawa *et al.* 1985; Palsbøll *et al.* 2004). This means that it can be used for loci, like the mtDNA, that do not have recombination, but generally do show evidence of homoplasy. The method has also been extended to include the stepwise mutation model (Hey *et al.* 2004).

Computer program development: A computer program (available from the authors) was written to implement the method with the enhancements. In addition to basic debugging, we employed three types of checking: ensuring that the posterior distributions are identical to the prior distributions when $f(X|G, \Theta)$ is set to 1, for all G and Θ; comparing results with simpler models for which posterior densities can be calculated directly or that can be assessed using other programs (Nielsen 1997; Wilson and Balding 1998; Nielsen and Wakeley 2001); and applying the method to data sets simulated under the IM model. This last method is the most complete but it is laborious because there is not a necessary relationship between the parameters used to simulate a data set and the posterior densities that are estimated from that simulated data. Rather, multiple simulated data sets need to be analyzed so that a set of posterior densities can be assessed in relation to the true parameter values used for the simulations. Figure 2 shows the marginal posterior densities estimated from each of 20 independent five-locus simulations. For each of the six demographic parameters, the posterior densities vary about the true value used in the simulation. To test whether the locations of these distributions, considered together, are consistent with the true values of the parameters (*i.e.*, the values used in the simulations), we used Fisher's approach to combining probabilities from independent tests of the same hypothesis (Fisher 1954). For each posterior density we determined p_i, $i = 1, \ldots, 20$, the chance that a parameter value is more extreme (*i.e.*, departs more from the mean of the distribution) than the actual true value. That is, if x is the area of the curve to the left of the true value then $p_i = 2x$ if $x < 0.5$ and $p_i = 2(1 - x)$ if $x > 0.5$. If the p_i's are uniformly distributed, then $a = -2\sum_{i=1}^{20}\text{Log}(p_i)$ is

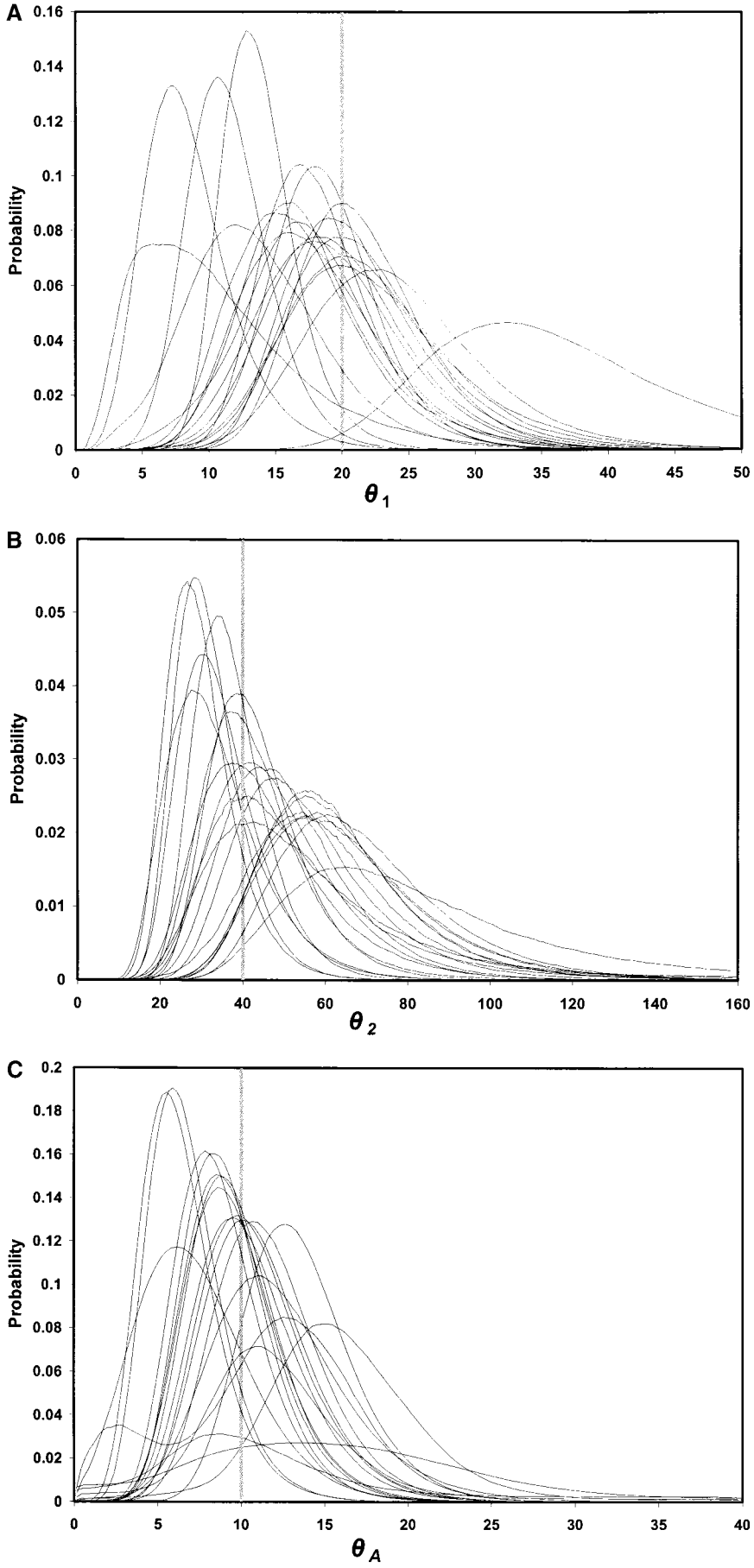
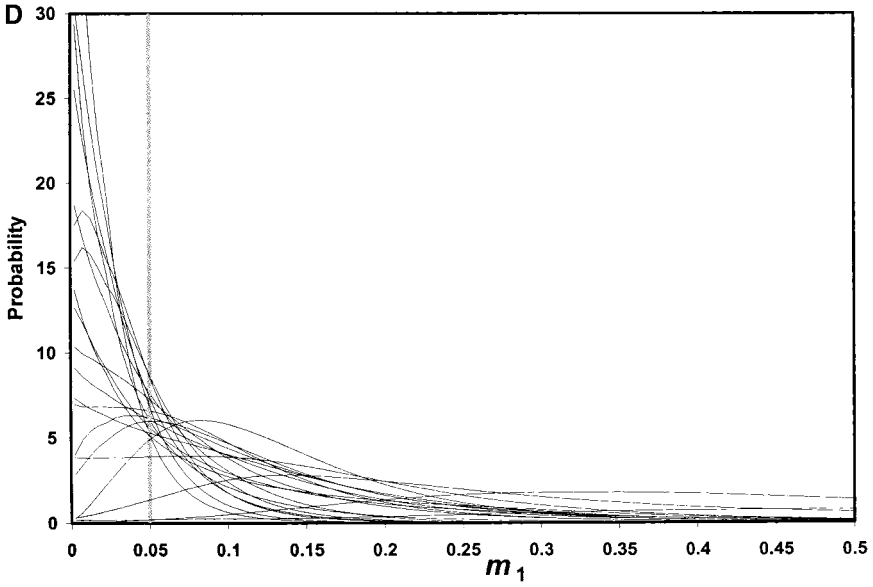
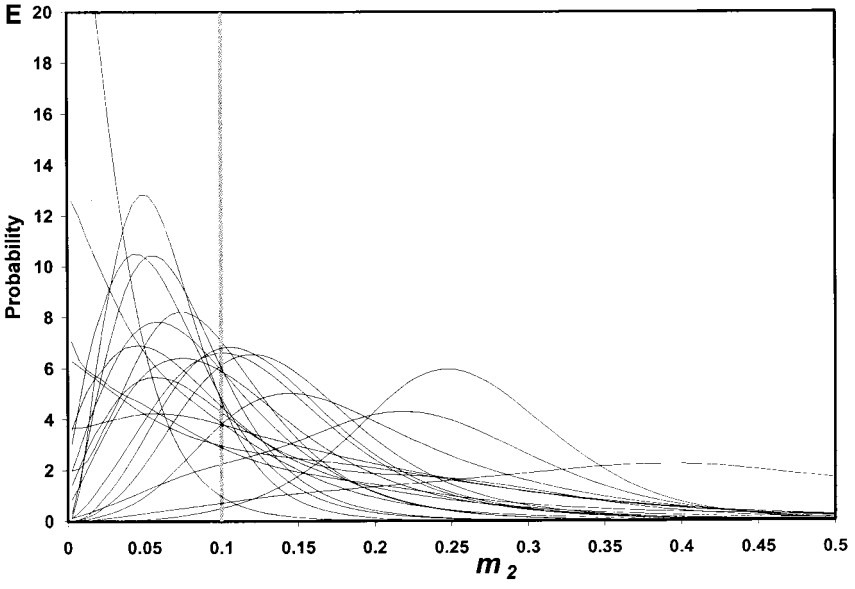


FIGURE 2.—The marginal densities obtained by fitting the IM model to simulated data. The input parameters for the simulations were as follows: $\theta_1 = 20$; $\theta_2 = 40$; $\theta_A = 10$; $m_1 = 0.05$ ($2\mathbf{N}_1\mathbf{m}_1 = 0.5$); $m_2 = 0.1$ ($2\mathbf{N}_2\mathbf{m}_2 = 2$); and $t = 5$ ($\mathbf{t}/2\mathbf{N}_1 = 0.5$). For each simulated data set, coalescent simulations were done for each of five loci with identical mutation rates under an infinite-sites mutation model, each with sample sizes of 10 for each of the two populations. Each simulated data set was analyzed using wide uniform prior distributions for each parameter and four chains (three heated chains, in addition to the primary chains) joined by Metropolis coupling. Each analysis began with a burn-in period of 300,000 steps followed by a primary chain of 6,000,000 steps. The curves for parameters θ_1 through t are shown in A–F, respectively. The true parameter values used in the simulations are shown as shaded vertical bars.





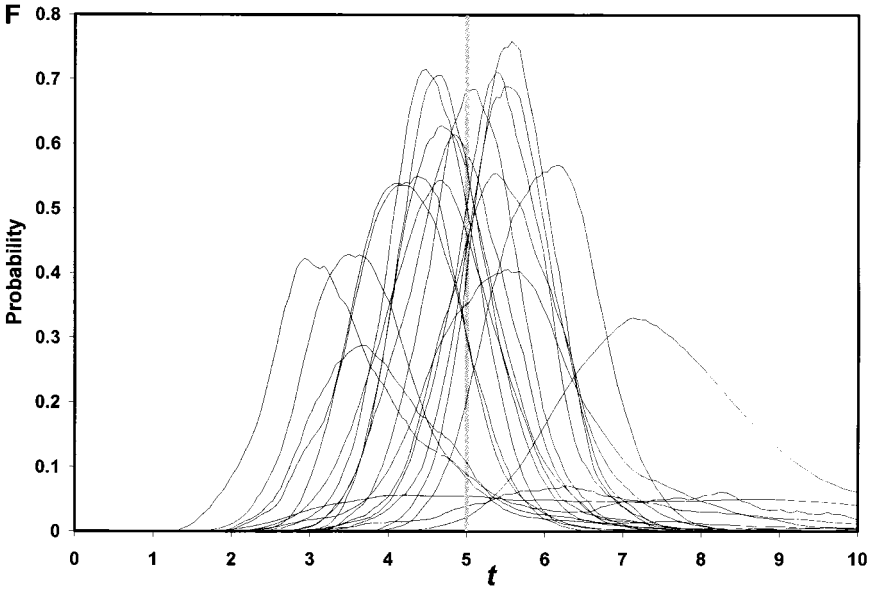


FIGURE 2.—*Continued.*

χ^2 distributed with 40 d.f. (*i.e.*, two times the number of densities). In this distribution 90% of the probability mass falls above 29.05, 50% falls above 39.3, and 10% falls above 51.8 (Rohlf and Sokal 1981). If the values of the true parameters in the coalescent simulations are consistent with the shape and locations of the posterior densities (*i.e.*, p_i is uniformly distributed), then a for each parameter should be drawn from this χ^2 distribution. The values of a for the different parameters are θ_1, 45.5; θ_2, 48.2; θ_A, 29.5; m_1, 43.1; m_2, 36.8; and t, 41.5. Although these values are not entirely independent of each other, they all fall in the middle of the χ^2 distribution, and their mean (40.8) is quite close to the 50% point of the χ^2 distribution (39.3). While results from only a limited number of simulations have been shown here, they do suggest that the method is reliable and that the credibility intervals established by the method may be interpreted as classical confidence intervals.

Figure 2 is also useful for providing a sense of how much data, under the infinite-sites mutation model, may be needed to return posterior probabilities that have useful confidence intervals and that might be expected to reveal gene flow, if it indeed had been ongoing. For example, most of the simulated data sets did not reveal a nonzero peak for m_1 (true value of 0.05 corresponding to $\mathbf{M}_1 = 0.5$), but most data sets did reveal a nonzero peak for m_2 (true value of 0.1, corresponding to $\mathbf{M}_2 = 2$). For the other parameters, most curves lie fairly close to the true value, but many curves also easily span values that are double or half the true value. Thus while these modestly sized simulated data sets provide a good approximate view of the true history, the simulations also suggest that larger multilocus data sets would be required to achieve narrow credible intervals for most parameters.

APPLICATIONS AND RESULTS

We applied these methods to the divergence of *D. pseudoobscura* and *D. persimilis* (Dobzhansky and Epling 1944). This well-studied species pair is well known as the focus of much of the research by Dobzhansky and colleagues over many years (Lewontin *et al.* 1981). When the species are crossed, hybrid females are fertile while hybrid males and some hybrid backcross females are sterile (Dobzhansky 1936). The species are partially sympatric in the western part of North America (from California to British Columbia; Dobzhansky and Epling 1944) and are known to hybridize at a low frequency in nature (Dobzhansky 1973; Powell 1983). Recently a set of inbred lines from each species was sequenced at 16 different portions of the genome. Analyses showed that loci varied significantly in their patterns of variation, strongly suggesting the presence of gene flow at some loci, but not at others (Wang and Hey 1996; Wang *et al.* 1997; Machado *et al.* 2002; Machado and Hey 2003). However, it has not been possible until now to quantify the gene flow, together with the other relevant population size and divergence time parameters.

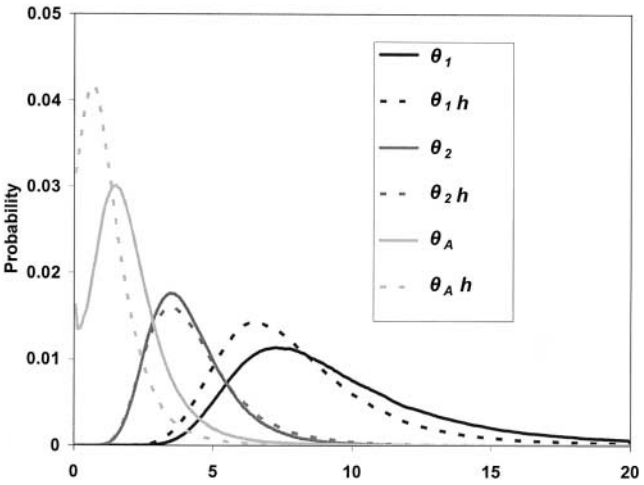


Figure 3.—The marginal densities for the population mutation rate parameters obtained by fitting the IM model to a three-locus data set (loci 4002, 2003, and X010). Distributions were obtained by integrating the full-likelihood surface over all of the other model parameters. Values for populations *D. pseudoobscura, D. persimilis*, and the ancestral species (θ_1, θ_2, and θ_A, respectively) are shown as dashed lines for runs in which the inheritance scalars are free to vary along with all other parameters in the model (also denoted h). The solid lines are for runs in which the inheritance scalars as set to specific values ($h = 1$ for loci 4002 and 2003 and $h = 0.75$ for locus X010).

Inclusion of inheritance scalars as parameters: To see the impact of including inheritance scalars as parameters, in a multilocus context, we fit the IM model to a data set for the three loci that showed zero or little evidence of recombination (see below). One locus (4002) showed no evidence of recombination, and two loci (2003 and X010) were consistent with zero recombination, provided that one sequence was removed from each sample set (the perSALEM sequence in the case of 2003 and psMATH10 in the case of X010). Two other loci (the mtDNA and *eyeless*) could have been included with these other three; however, both showed genealogical histories that departed markedly from others, suggesting the action of natural selection (Machado and Hey 2003).

We fit the IM model for both the case of constant inheritance scalars ($h = 1$ for autosomal loci 4002 and 2003, and $h = 0.75$ for X-linked locus X010) and the case when inheritance scalars were free to vary along with the other parameters. As shown in Figure 3, both with and without inheritance parameters, the positions of the peaks of the marginal posterior densities for the population size parameters suggest that *D. pseudoobscura* has had a larger effective population than *D. persimilis*. In this case, the effect of including the inheritance parameters is to shift the positions of the peaks a modest amount. A similarly modest effect is observed on t, and regardless of the inheritance parameters, both migra-

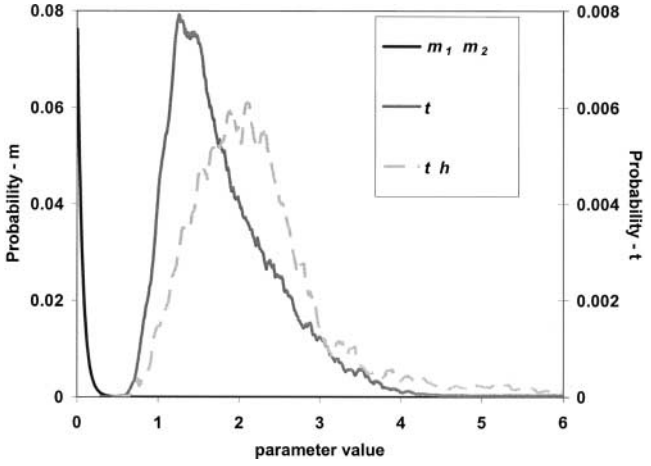


FIGURE 4.—The marginal densities for the migration and time parameters. Distributions were obtained by integrating the full-likelihood surface over all of the other model parameters. Both migration rate parameters, m_1 and m_2, revealed nearly identical peaks, both with and without inclusion of inheritance parameters (h). For t, the solid line shows the case when inheritance values were preset constants and the dashed line shows the case where inheritance terms are free to vary as parameters.

tion rate parameters were nearly identical and showed strong peaks at zero (Figure 4). The effect on mutation rate scalars is more dramatic, and the curves that result by inclusion of inheritance scalars are farther apart from each other and considerably flatter than without them (Figure 5). The curves for the inheritance scalars themselves are shown in Figure 6, with estimated peak locations at 2.21 (locus 2003), 1.09 (locus 4002), and 0.39 (locus X010). As expected the geometric mean of these values is near one (0.98). The estimate for a given locus reflects the departure of locus-specific effective population size from this geometric mean. The estimates vary considerably from the case when inheritance scalars are set as constants: 1 for autosomal loci (loci 2003 and 4002) and ¾ for the X-linked locus X010. Although the curves overlap, they are consistent with the loci having different effective population sizes, possibly by the action of recurrent selective sweeps in partially linked regions of the chromosome (GILLESPIE 2000) or by background selection (CHARLESWORTH *et al.* 1993).

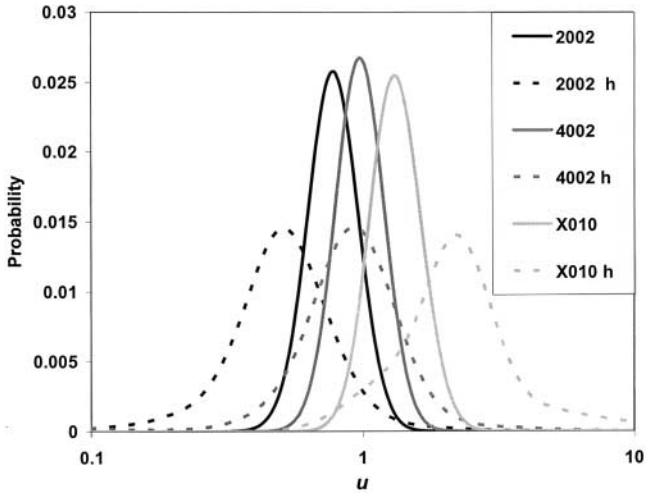


FIGURE 5.—The marginal densities for the mutation rate scalars. Distributions were obtained by integrating the full-likelihood surface over all of the other model parameters. Values are shown for each of the three loci in the data set. As in Figures 3 and 4, results are shown for both the case when inheritance scalars are set as constants and the case when they are free to vary along with the other parameters (the latter are designated h).

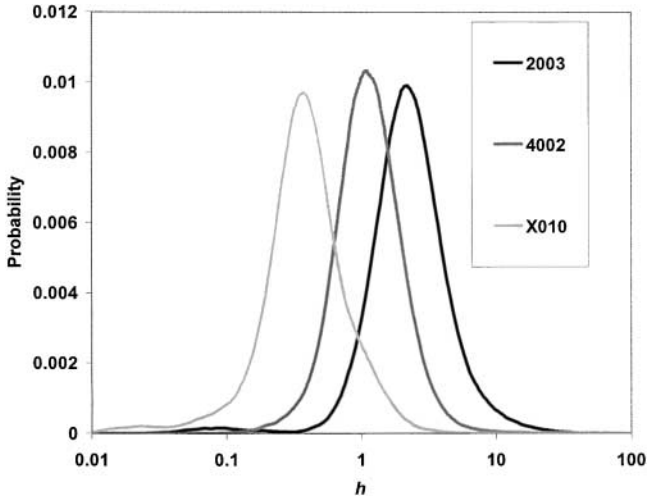


FIGURE 6.—The marginal densities for the inheritance parameters for the runs in which they are free to vary. Distributions were obtained by integrating the full-likelihood surface over all of the other model parameters. Values are shown for each of the three loci in the data set.

Together these three loci seem to fit the circumstances under which separate estimates of both mutation scalars and inheritance scalars can be obtained. In the first place, the amount of divergence relative to the depth of genealogies within species is not low (though neither is it very high). Estimates of divergence time in units of 2**N** generations (*i.e.*, $2\ t/\theta$) are 0.34 for *D. pseudoobscura* and 0.73 for *D. persimilis*. Also the migration rates have probably been very low or zero in both directions for these loci (Figure 4).

Loci with recombination: The fitting of the IM model assumes that the genealogical history of a locus is strictly bifurcating and thus does not include recombination or gene conversion. Furthermore, it is difficult to include recombination in a genealogically based, likelihood framework for historical model fitting (KUHNER *et al.* 2000; NIELSEN 2000). However, there are ways to use the method with data that come from recombining genomes by taking advantage of the imprint left by recombination on the pattern of haplotype variation at a locus. One approach is to limit analyses to loci that do not

show evidence of recombination by the "four-gamete" criterion (Hudson 1985), as was done in the example described. The pitfall here is that we expect such loci to have shorter genealogies, on average. This is because those genes that happen to have shorter gene trees will also have had less opportunity for recombination within the span of their genealogical history. In the case of 4002, 2003, and X010, the data suggest that there has not been gene flow (Figure 4). If indeed the IM model is roughly appropriate in these cases, then it is probably not the case that these loci have had histories much shorter than other loci, simply because, in the absence of gene flow, the depth of the gene tree must extend at least to t. Finally, the estimate of t per kilobase pair of sequence (Table 1) is not lower for these loci than for others.

Another approach for a locus that shows evidence of recombination is to break the data into blocks of sequence each of which does not show evidence of recombination. The algorithm of Hudson and Kaplan (1985) can be used to identify sequence blocks across which all sequences are consistent with a model of no recombination. Then together all such blocks found within a locus can be included in a multilocus fitting of the IM model (*i.e.*, each block is a "locus"). This will violate the assumption that different loci have segregated independently, because the different loci will have had highly correlated histories because of tight linkage, and this is expected to lead to poorly estimated (and more sharply peaked) densities. However, it may still be the case that the mode of the posterior density has an expected value that is the same as a proper maximum-likelihood estimate. For each of the loci that showed evidence of recombination by the four-gamete criterion, the data were divided into multiple portions, with each portion treated as an independent locus in a run of the IM program. The locations of peak heights for each of the main parameters are shown in Table 1. Four loci (*Adh*, *4003*, *3002*, and *Rh1*) revealed estimates of population migration rates ($\mathbf{M}_1$ and $\mathbf{M}_2$) >0.4. This is consistent with other analyses that indicated that gene flow was limited mostly to loci that are not near chromosomal inversions that distinguish *D. persimilis* and *D. pseudoobscura* (Machado *et al.* 2002; Machado and Hey 2003). Interestingly, low but nonzero levels of gene flow are suggested at several X chromosomal loci, which do carry large inversions.

Yet another approach that allows a portion of the data from recombining loci to be included in a multilocus analysis is to take from each separate locus one randomly selected block of sequence identified as nonrecombining by the four-gamete criterion and to not use the remainder of the data. We used this approach in a multilocus analysis of all the loci listed in Table 1. The parameter estimates from this multilocus analysis are near those of the means of the values estimated for each of the loci individually. Overall the analyses suggest that these two species have had low levels of gene flow in the time since they began diverging.

mtDNA: Table 1 also shows the results of analysis on data from the mtDNA (Machado and Hey 2003). The estimated gene tree for these data differed dramatically from those of other loci, with *D. pseudoobscura* and *D. persimilis* sharing multiple complete haplotypes (despite high levels of polymorphism) and with a high level of divergence between sequences from these species and a third species, *D. pseudoobscura bogotana* (Machado and Hey 2003). In this case, fitting of the IM model suggests that *D. persimilis* has had a much larger effective population size than *D. pseudoobscura* and a high level of gene flow from *D. persimilis* to *D. pseudoobscura* in the coalescent (*i.e.*, going backward in time, the direction is the reverse when considered forward in time). Given that the mtDNA includes a large number of completely linked loci, it is probably the case that natural selection has shaped this history (Machado and Hey 2003).

DISCUSSION

The study of population divergence has often been limited to two quite different general models. One class of models assumes that divergence is the result of an equilibrium between genetic drift, mutation, and limited gene flow, acting over a very long period of time, while the second class does not include gene flow, but instead supposes that divergence is the result of population splitting at some point in the past. Sewall Wright's classic work on population subdivision (Wright 1922, 1931, 1951) embraces the first class of equilibrium models, as do stepping-stone models (Kimura and Weiss 1964). In recent years methods for simultaneously estimating migration rates and population sizes have been developed for equilibrium gene flow models (Beerli and Felsenstein 1999; Bahlo and Griffiths 2000). New methods have also been developed for estimating historical population sizes and divergence times for the nonequilibrium isolation model, assuming no migration (Wakeley and Hey 1997; Nielsen 1998; Nielsen *et al.* 1998; Nielsen and Slatkin 2000; Rannala and Yang 2003; Wilson *et al.* 2003). However, for many questions concerning the divergence of populations, investigators need methods that permit assessments of both population splitting and gene flow simultaneously (Slatkin and Maddison 1989; Takahata and Slatkin 1990). For example, Nielsen and Hey (2003) showed that likelihood models that do not take migration into account are not adequate to describe the history of some human populations.

A new tool for the study of divergence: We have extended the original MCMC method of Nielsen and Wakeley (2001) to include multiple loci with locus-specific inheritance scalars. With inclusion of multiple loci the parameters fall into three distinct categories: the primary demographic parameters (including θ_1, θ_2,

TABLE 1

Parameter estimates

Gene	No. loci[a]	Length[b]	θ_1	θ_2	θ_A	m_1[c]	m_2[c]	t	u_Σ[d]	$(t \times u_\Sigma)$/kbp[e]	$\theta_1 \times u_\Sigma$[f]	$\theta_2 \times u_\Sigma$[f]	**R**[g]	**A**[h]	$\mathbf{M}_1$[i]	$\mathbf{M}_2$[i]
per	9	1470	1.5	0.6	0.3	0.19	0.32	0.87	11.1	6.57	16.2	6.5	0.40	0.23	0.13	0.09
X008	10	997	3.6	0.3	0.5	0.08	0	0.25	16.1	4.04	57.9	4.0	0.07	0.15	0.14	0
X009	6	698	0.8	1.1	0.5	0	0.10	0.70	6.9	6.92	5.6	7.8	1.40	0.65	0	0.05
Hsp82	3	1957	3.0	0.5	0.9	0	0.07	0.57	4.3	1.25	13.1	2.2	0.17	0.29	0	0.02
X010	1	871	4.5	3.1	2.0	0	0	4.76	1	5.46	4.5	3.1	0.69	0.45	0	0
2003	1	522	5.4	1.7	1.0	0	0	0.57	1	1.09	5.4	1.7	0.33	0.18	0	0
rh1	9	1443	2.4	1.7	0.3	0.38	0	0.34	10.6	2.50	25.1	18.2	0.73	0.14	0.45	0
bcd	8	1371	1.2	0.6	0.5	0	0	0.32	11.3	2.64	13.2	6.9	0.52	0.38	0	0
2002	6	915	5.5	0.6	0.3	0	0	0.76	7.0	5.81	38.7	4.3	0.11	0.05	0	0
2001	4	677	4.2	1.8	0.5	0	0	0.31	4.4	2.01	18.4	7.9	0.43	0.11	0	0
3002	11	660	0.9	0.4	0.5	0.02	5.00	0.02	17.6	0.53	16.4	6.6	0.41	0.58	0.01	0.94
4003	3	619	5.5	3.8	1.0	0.30	0	0.35	3.1	1.75	16.9	11.6	0.69	0.18	0.82	0
Adh	20	3448	0.8	1.1	0.6	0.17	7.93	0.16	28.6	1.33	22.0	30.7	1.39	0.77	0.06	4.25
4002	1	825	5.5	2.3	0.8	0	0	1.45	1	1.76	5.5	2.3	0.41	0.15	0	0
Mean[j]	6.6	1177	3.2	1.4	0.7	0.08	0.96	0.82		3.12			0.55	0.31	0.12	0.38
Joint[k]	14		3.0	1.6	0.7	0.06	0.14	0.62					0.54	0.23	0.09	0.11
mtDNA[l]	1	1826	13.4	1017	1.8	0.20	0.82	2.29	1	1.25	13.4	1017.6	76.15	0.13	1.30	414.65

Parameter estimates are obtained from the location of the peaks of the marginal posterior distributions. *D. pseudoobscura* was designated as species 1 and *D. persimilis* was designated as species 2. The average sample sizes were 16 sequences for *D. pseudoobscura* and 13 sequences for *D. persimilis* (Machado *et al.* 2002). Results are shown for each individual locus. For those loci that showed evidence of recombination, the data were divided into segments as described in the text. Also shown are the mean parameter estimates for the 14 X-linked and autosomal loci. The joint estimate is based on including the leftmost segment of each of the 14 loci in a single model. Results for the mtDNA sequences are shown separately, because of the unique history of this locus (Machado and Hey 2003).

[a] For individual genes, the number of "loci" is the number of apparently nonrecombining segments into which the data were divided to meet the four-gamete criterion for each segment.

[b] The average length of complete sequences.

[c] Migration rate estimates were identified as being at 0 when the highest observed value of the marginal posterior density was at the lower limit of resolution. The HKY model was used for the mtDNA.

[d] u_Σ is the sum of the mutation rate scalars for the different segments.

[e] Divergence time in units of mutations per kilobase pairs of sequence (*i.e.*, t/length).

[f] The product of θ estimates and the sum of the mutation rate scalars is an estimate of θ for the entire locus.

[g] The ratio of the estimates of population mutation rates, θ_2/θ_1, reflects the size of *D. persimilis* relative to that of *D. pseudoobscura.*

[h] The ratio of the estimates of population mutation rates, θ_A/θ_1, reflects the size of the ancestral species relative to that of *D. pseudoobscura.*

[i] Population migration rate estimates, $\mathbf{M}_1 = 2\mathbf{N}_1\mathbf{m}_1 = m_1 \times \theta_1/2$ and $\mathbf{M}_2 = 2\mathbf{N}_2\mathbf{m}_2 = m_2 \times \theta_2/2$.

[j] Mean parameter estimates for the 14 X-linked and autosomal loci.

[k] Results of fitting the model to all 14 loci. Those loci that showed evidence of recombination were represented only by the leftmost segment.

[l] Results for the mtDNA, including data from both *ND4* and *COI* (Machado and Hey 2003).

θ_A, m_1, m_2, and t); the mutation scalars; and, if implemented as parameters, the inheritance scalars. To facilitate mixing of the Markov chain, we have also implemented Metropolis coupling (Geyer 1991). In addition, the original limitation to the infinite-sites mutation model has been overcome by the inclusion of the HKY mutation model (Palsbøll *et al.* 2004) and the stepwise mutation model, as well as a model that includes loci that have both an infinite-sites portion and a stepwise portion (Hey *et al.* 2004). With these extensions the method offers a versatile tool for addressing questions that, while traditionally quite difficult, are critical for our understanding of basic evolutionary processes of divergence.

Inheritance scalars as parameters: With the inclusion of inheritance parameters, it may be possible for investigators to study the effects of selection, via linkage, within the IM model. If directional selection acts steadily, either as recurrent selective sweeps or as background selection, then the levels of polymorphism in a region will be a function of local gene density and recombination levels, both of which may be shared between closely related species. For example, both in the *D. simulans* complex (Kliman *et al.* 2000) and between *D. pseudoobscura* and its sister species (Machado *et al.* 2002), polymorphism levels per base pair are correlated across loci between species. One reason for this may be the action of selection. Traditionally, studies of polymorphism levels as a function of linkage have been limited to intraspecific comparisons (together with an outgroup

to control for mutation rate; Begun and Aquadro 1992). Also traditionally, the fitting of demographic models like the IM model has required that loci conform to the neutral model. The invocation of inheritance parameters may allow more complete studies that include selection *and* demography.

Simplified models as special cases: Another advantage of a highly parameterized IM model is that it includes a number of boundary cases that are often of interest. Thus, by prescribing migration rates of zero, the model becomes a conventional isolation model (Takahata and Nei 1985; Hey 1994; Wakeley and Hey 1997). If migration rates are nonzero and the time of splitting is specified to be very long ago, then the model becomes a simple two-island model and can be used to study the countervailing forces of genetic drift and gene flow and the equilibrium between them. If one of the descendant populations has a size of zero then the model becomes one of instantaneous population size change (*i.e.*, at t) for the remaining population. Finally if it is specified that $t = 0$, such that there has effectively not been a separation, then the model becomes one of a single constant-size population. The model can also be simplified, and the number of parameters reduced, by specifying that two or all three population mutation rates are identical or that the two migration rates are identical. All of these variations are included in the computer program that realizes the method.

The divergence of *D. pseudoobscura* and *D. persimilis*: Previous studies on these species have shown that different loci vary widely in their genealogical history and apparent phylogenetic history, in ways that are broadly consistent with a model in which gene flow occurs but is prevented at some loci by natural selection (Wang *et al.* 1997; Machado *et al.* 2002; Machado and Hey 2003). Those loci within or near inversions that distinguish *D. pseudoobscura* from *D. persimilis* show little or no evidence of gene flow, in contrast to loci on other regions of the genome. This is consistent both with the genetic map locations of loci that contribute to male hybrid sterility (Noor *et al.* 2001b) and with models in which chromosomal inversions played a formative role by limited gene flow early in the divergence of the species (Noor *et al.* 2001a; Rieseberg 2001).

By fitting the IM model we are now able to put numbers to the gene flow rates, without confounding them with our estimates of population sizes and divergence time. The estimates of the population migration rates ($\mathbf{M}_1$ and $\mathbf{M}_2$) vary considerably (particularly so when the mtDNA is considered; Table 1). However, with the exceptions of the mtDNA and the *Adh* locus, population migration rate estimates are <1. Interestingly, gene flow does not appear to be one sided, with approximately equal numbers of loci suggesting gene flow in each of the two directions, although for any given locus with nonzero gene flow estimates usually gene flow is indicated in only one direction.

We can also inquire of the date at which *D. pseudoobscura* and *D. persimilis* began to diverge. From Table 1, the mean estimated value of t since population splitting is 3.12 mutations per kilobase pair. To convert to absolute time, we can use the estimated absolute divergence time between *D. pseudoobscura* and the more distantly related species, *D. miranda*, of 2 million years (Aquadro *et al.* 1991; Wang and Hey 1996). The mean divergence between these species over the 14 loci in Table 1 is 21.2 changes per kilobase pair (Machado *et al.* 2002). Thus the estimated rate of divergence, per year, per kilobase is 5.3×10^{-6} changes per year and the estimated time of common ancestry between *D. pseudoobscura* and *D. persimilis* is 589,000 years (*i.e.*, $3.12/5.3 \times 10^{-6}$). Given the phenotypic similarity between these species (Dobzhansky 1944), the fact that they can produce fertile hybrids, and the estimates of gene flow between them, it is perhaps surprising that the divergence time estimate is so great. If we roughly adjust for generations (*e.g.*, eight per year in Drosophila) then there may have been 20 times the number of generations separating these two species of Drosophila as currently separates humans and chimpanzees [*i.e.*, assuming 6 million years divergence, at 25 years per generation (Chen and Li 2001; Brunet *et al.* 2002)]. The contrast suggests a slow divergence process between *D. pseudoobscura* and *D. persimilis*, notwithstanding the presence of gene flow.

We thank John Wakeley for helpful comments throughout this work.

LITERATURE CITED

Aquadro, C. F., A. L. Weaver, S. W. Schaeffer and W. W. Anderson, 1991 Molecular evolution of inversions in Drosophila pseudoobscura: the amylase gene region. Proc. Natl. Acad. Sci. USA **99:** 305–309.

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

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

Begun, D. J., and C. F. Aquadro, 1992 Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. Nature **356:** 519–520.

Brunet, M., F. Guy, D. Pilbeam, H. T. Mackaye, A. Likius *et al.*, 2002 A new hominid from the Upper Miocene of Chad, Central Africa. Nature **418:** 145–151.

Charlesworth, B., M. T. Morgan and D. Charlesworth, 1993 The effect of deleterious mutations on neutral molecular evolution. Genetics **134:** 1289–1303.

Chen, F.-C., and W.-H. Li, 2001 Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. Am. J. Hum. Genet. **68:** 444–456.

Dobzhansky, T., 1936 Studies of hybrid sterility. II. Localization of sterility factors in *Drosophila pseudoobscura* hybrids. Genetics **21:** 113–135.

Dobzhansky, T., 1944 Chromosomal races in *Drosophila pseudoobscura* and *Drosophila persimilis*, pp. 47–144 in *Contributions to the Genetics, Taxonomy, and Ecology of Drosophila pseudoobscura and Its Relatives*, edited by T. Dobzhansky and C. Epling. Carnegie Institute of Washington, Washington, DC.

Dobzhansky, T., 1973 Is there gene exchange between *Drosophila pseudoobscura* and *Drosophila persimilis* in their natural habitats? Am. Nat. **107:** 312–314.

Dobzhansky, T., and T. Epling, 1944 Taxonomy, geographic distribution and ecology of *Drosophila pseudoobscura* and its relatives, pp. 1–46 in *Contributions to the Genetics, Taxonomy, and Ecology of Drosophila pseudoobscura and Its Relatives*, edited by T. Dobzhansky and C. Epling. Carnegie Institute of Washington, Washington, DC.

Fisher, R. A., 1954 *Statistical Methods for Research Workers*. Oliver & Boyd, Edinburgh.

Gelman, A., and D. B. Rubin, 1992a Inference from iterative simulation using multiple sequences. Stat. Sci. **7:** 457–472.

Gelman, A., and D. B. Rubin, 1992b Rejoinder: replication without contrition. Stat. Sci. **7:** 503–511.

Geyer, C. J., 1991 Markov chain Monte Carlo maximum likelihood, pp. 156–163 in *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, edited by E. M. Keramidas. Interface Foundation of North America, Seattle.

Geyer, C. J., 1992a Practical Markov chain Monte Carlo. Stat. Sci. **7:** 473–511.

Geyer, C. J., 1992b Rejoinder. Stat. Sci. **7:** 502–503.

Gilks, W. R., S. Richardson and D. J. Spiegelhalter, 1996 *Markov Chain Monte Carlo in Practice*. Chapman & Hall, Boca Raton, FL.

Gillespie, J. H., 2000 Genetic drift in an infinite population. The pseudohitchhiking model. Genetics **155:** 909–919.

Hasegawa, M., H. Kishino and T. Yano, 1985 Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. **22:** 160–174.

Hey, J., 1994 Bridging phylogenetics and population genetics with gene tree models, pp. 435–449 in *Molecular Approaches to Ecology and Evolution*, edited by B. Schierwater, B. Streit, G. Wagner and R. Desalle. Birkhäuser-Verlag, Basel, Switzerland.

Hey, J., and C. A. Machado, 2003 The study of structured populations—new hope for a difficult and divided science. Nat. Rev. Genet. **4:** 535–543.

Hey, J., Y.-J. Won, A. Sivasundar, R. Nielsen and J. A. Markert, 2004 Using nuclear haplotypes with microsatellites to study gene flow between recently separated Cichlid species. Mol. Ecol. **13:** 909–919.

Hudson, R. R., 1983 Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. **23:** 183–201.

Hudson, R. R., 1985 The sampling distribution of linkage disequilibrium under an infinite allele model without selection. Genetics **109:** 611–631.

Hudson, R. R., and N. L. Kaplan, 1985 Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics **111:** 147–164.

Kimura, M., 1969 The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics **61:** 893–903.

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

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

Kingman, J. F. C., 1982b On the genealogy of large populations. J. Appl. Probab. **19A:** 27–43.

Kliman, R. M., P. Andolfatto, J. A. Coyne, F. Depaulis, M. Kreitman *et al.*, 2000 The population genetics of the origin and divergence of the *Drosophila simulans* complex species. Genetics **156:** 1913–1931.

Kuhner, M. K., J. Yamato and J. Felsenstein, 2000 Maximum likelihood estimation of recombination rates from population data. Genetics **156:** 1393–1401.

Laporte, V., and B. Charlesworth, 2002 Effective population size and population subdivision in demographically structured populations. Genetics **162:** 501–519.

Latter, B. D., 1973 The island model of population differentiation: a general solution. Genetics **73:** 147–157.

Lewontin, R. C., J. A. Moore, W. B. Provine and B. Wallace, 1981 *Dobzhansky's Genetics of Natural Populations I–XLIII*. Columbia University Press, New York.

Machado, C. A., and J. Hey, 2003 The causes of phylogenetic conflict in a classic Drosophila species group. Proc. R. Soc. Lond. Ser. B **270:** 1193–1202.

Machado, C., R. M. Kliman, J. M. Markert and J. Hey, 2002 Inferring the history of speciation from multilocus DNA sequence data: the case of *Drosophila pseudoobscura* and its close relatives. Mol. Biol. Evol. **19:** 472–488.

Maynard Smith, J., and J. Haigh, 1974 The hitch-hiking effect of a favourable gene. Genet. Res. **23:** 23–35.

Mayr, E., 1942 *Systematics and the Origin of Species*. Columbia University Press, New York.

Muller, H. J., 1940 Bearings of the Drosophila work on systematics, pp. 185–268 in *The New Systematics*, edited by J. Huxley. Clarendon Press, Oxford.

Nielsen, R., 1997 A likelihood approach to populations samples of microsatellite alleles. Genetics **146:** 711–716.

Nielsen, R., 1998 Maximum likelihood estimation of population divergence times and population phylogenies under the infinite sites model. Theor. Popul. Biol. **53:** 143–151.

Nielsen, R., 2000 Estimation of population parameters and recombination rates from single nucleotide polymorphisms. Genetics **154:** 931–942.

Nielsen, R., and J. Hey, 2003 Discussion on the paper by Wilson, Weale and Balding. J. R. Stat. Soc. Ser. A Stat. Soc. **166:** 188.

Nielsen, R., and M. Slatkin, 2000 Likelihood analysis of ongoing gene flow and historical association. Evolution **54:** 44–50.

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

Nielsen, R., J. L. Mountain, J. P. Huelsenbeck and M. Slatkin, 1998 Maximum-likelihood estimation of population divergence times and population phylogeny in models without mutation. Evolution **52:** 669–677.

Noor, M. A., K. L. Grams, L. A. Bertucci and J. Reiland, 2001a Chromosomal inversions and the reproductive isolation of species. Proc. Natl. Acad. Sci. USA **98:** 12084–12088.

Noor, M. A. F., K. L. Grams, A. Bertucci, Y. Almendarez, J. A. Reiland *et al.*, 2001b The genetics of reproductive isolation and the potential for gene exchange between *Drosophila pseudoobscura* and *D. persimilis* via backcross hybrid males. Evolution **55:** 512–521.

Palsbøll, P. J., M. Bérubé, A. Aguilar, G. Notarbartolo di Sciara and R. Nielsen, 2004 Discerning between recurrent gene flow and recent divergence under a finite-site mutation model applied to North Atlantic and Mediterranean Sea fin whale (Balaenoptera physalus) populations. Evolution **58:** 670–675.

Powell, J. R., 1983 Interspecific cytoplasmic gene flow in the absence of nuclear gene flow: evidence from Drosophila. Proc. Natl. Acad. Sci. USA **80:** 492–495.

Rannala, B., and Z. Yang, 2003 Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics **164:** 1645–1656.

Rieseberg, L. H., 2001 Chromosomal rearrangements and speciation. Trends Ecol. Evol. **16:** 351–358.

Rohlf, F. J., and R. R. Sokal, 1981 *Statistical Tables*. W. H. Freeman, San Francisco.

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

Takahata, N., 1995 A genetic perspective on the origin and history of humans. Annu. Rev. Ecol. Syst. **26:** 343–372.

Takahata, N., and M. Nei, 1985 Gene genealogy and variance of interpopulational nucleotide differences. Genetics **110:** 325–344.

Takahata, N., and M. Slatkin, 1990 Genealogy of neutral genes in two partially isolated populations. Theor. Popul. Biol. **38:** 331–350.

Tavare, S., 1984 Line-of-descent and genealogical processes, and their applications in population genetics models. Theor. Popul. Biol. **26:** 119–164.

Wakeley, J., 1996a Distinguishing migration from isolation using the variance of pairwise differences. Theor. Popul. Biol. **49:** 369–386.

Wakeley, J., 1996b Pairwise differences under a general model of population subdivision. J. Genet. **75:** 81–89.

Wakeley, J., and J. Hey, 1997 Estimating ancestral population parameters. Genetics **145:** 847–855.

Wakeley, J., and J. Hey, 1998 Testing speciation models with DNA sequence data, pp. 157–175 in *Molecular Approaches to Ecology and Evolution*, edited by R. DeSalle and B. Schierwater. Birkhäuser Verlag, Basel, Switzerland.

Wang, J. L., and A. Caballero, 1999 Developments in predicting the effective size of subdivided populations. Heredity **82:** 212–226.

Wang, R. L., and J. Hey, 1996 The speciation history of *Drosophila pseudoobscura* and close relatives: inferences from DNA sequence variation at the *period* locus. Genetics **144:** 1113–1126.

Wang, R. L., J. Wakeley and J. Hey, 1997 Gene flow and natural selection in the origin of *Drosophila pseudoobscura* and close relatives. Genetics **147:** 1091–1106.

Wilson, I. J., and D. J. Balding, 1998 Genealogical inference from microsatellite data. Genetics **150:** 499–510.

Wilson, I. J., M. E. Weale and D. J. Balding, 2003 Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. J. R. Stat. Soc. Ser. A Stat. Soc. **166:** 155–188.

Wright, S., 1922 Coefficients of inbreeding and relationship. Am. Nat. **56:** 330–338.

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

Wright, S., 1951 The genetical structure of populations. Ann. Eugen. **15:** 323–354.

Communicating editor: M. K. Uyenoyama