# Estimation of Admixture Proportions: A Likelihood-Based Approach Using Markov Chain Monte Carlo

**Lounès Chikhi,*^,†^ Michael W. Bruford*^,‡^ and Mark A. Beaumont*^,§^**

*Institute of Zoology, Regent's Park, London NW1 4RY, United Kingdom, †School of Biological Sciences, Queen Mary and Westfield College, University of London, London E1 4NS, United Kingdom, ‡School of Biosciences, Cardiff University, Cardiff CF10 3TL, United Kingdom and §School of Animal and Microbial Sciences, University of Reading, Reading RG6 6AJ, United Kingdom*

Manuscript received September 20, 2000
Accepted for publication April 24, 2001

ABSTRACT

When populations are separated for long periods and then brought into contact for a brief episode in part of their range, this can result in genetic admixture. To analyze this type of event we considered a simple model under which two parental populations ($P_1$ and $P_2$) mix and create a hybrid population (H). After that event, the three populations evolve under pure drift without exchange during $T$ generations. We developed a new method, which allows the simultaneous estimation of the time since the admixture event (scaled by the population size $t_i = T/N_i$, where $N_i$ is the effective population size of population $i$) and the contribution of one of two parental populations (which we call $p_1$). This method takes into account drift since the admixture event, variation caused by sampling, and uncertainty in the estimation of the ancestral allele frequencies. The method is tested on simulated data sets and then applied to a human data set. We find that (i) for single-locus data, point estimates are poor indicators of the real admixture proportions even when there are many alleles; (ii) biallelic loci provide little information about the admixture proportion and the time since admixture, even for very small amounts of drift, but can be powerful when many loci are used; (iii) the precision of the parameters' estimates increases with sample size ($n = 50$ *vs.* $n = 200$) but this effect is larger for the $t_i$'s than for $p_1$; and (iv) the increase in precision provided by multiple loci is quite large, even when there is substantial drift (we found, for instance, that it is preferable to use five loci than one locus, even when drift is 100 times larger for the five loci). Our analysis of a previously studied human data set illustrates that the joint estimation of drift and $p_1$ can provide additional insights into the data.

DURING their history, populations can be separated for long periods and then brought into contact for a brief episode in part of their range, resulting in genetic admixture (Bernstein 1931; Chakraborty 1986). This process is frequent in human populations where movements have brought together populations that were historically separated for varying amounts of time. This can be seen, for instance, in South America where many groups are essentially mixed populations containing varying amounts of contributions from European, African, and native American stocks (*e.g.*, Roberts and Hiorns 1965; Chakraborty 1986). Admixture occurs widely and in many species and has certainly taken place a great many times since the last glaciations when populations expanded from different refugia (Taberlet *et al.* 1998; Hewitt 2000). On a smaller time scale, humans have caused extensive admixture through transfers of plants and animals, both inadvertently (as in the case of commensal species) and deliberately (as in restocking of rivers with nonnative fishes). Admixture has also been quite common during the process of domestication and the creation of new breeds.

The interest for admixture estimation and admixed populations thus ranges from evolutionary to more applied issues. The study of admixed populations can provide information on (i) the inheritance of complex genetic disease and, in particular, the mapping of the genes involved (Chakraborty and Weiss 1988; McKeigue *et al.* 2000). In biogeography it could (ii) help identify the relative contributions of different glacial refugia to current populations. In conservation biology it could also (iii) help define which source populations, and in which proportion, should be used when reintroduction programs are defined.

Even though one could use admixture methods to estimate the relative contributions of subspecies meeting in hybrid zones, it is important to stress that the studies of hybrid zones and of admixed populations are often quite different. Whereas hybrid zone studies deal with spatial phenomena, admixture studies usually disregard this aspect and concentrate on the estimation of admixture proportions (see, for instance, Goodman *et al.* 1999 for an example where the difference is analyzed).

Recent theoretical advances substantially improved

*Corresponding author:* Lounès Chikhi, Department of Biology, University College London, Wolfson House, 4 Stephenson Way, London NW1 2HE, United Kingdom. E-mail: l.chikhi@ucl.ac.uk

the ability to use genetic information from present-day populations to draw inferences about past demographic events (*e.g.*, Slatkin and Hudson 1991; Rogers and Harpending 1992; Wilson and Balding 1998; Beaumont 1999). The coalescent theory (Kingman 1982a,b) provided population geneticists with both a statistical framework and a simple way to simulate samples taken from populations evolving under different demographic models (Hudson 1990). However, until a few years ago, all coalescent-based methods were applied to summary statistics. In practice, the coalescent theory was used to simulate genealogical trees under different demographic models and the simulated data sets were used to estimate the distribution of an appropriate statistic ($n_A$, the number of alleles, $H_e$, the expected heterozygosity, etc.). Although powerful, these methods were criticized because they do not make full use of the genetic information present in the allelic distribution (Felsenstein 1992). Clearly, any method based on a transformation of the original data can lead to a loss of information and should therefore be less powerful than methods that use the probability of observing the exact sample configurations (*i.e.*, the likelihood of the sample). This is particularly relevant for genetic data where the information available is inherently limited due to correlation between the data points.

One could naively use coalescent-based simulations to estimate how often a particular allelic configuration is observed. Practically, however, this is not possible because the number of possible genealogies becomes astronomical very quickly. As a consequence, even for moderate sample sizes ($n > 10$), the likelihood is impossible to evaluate by direct simulation. One could also consider using an analytical approach to derive an expression for the likelihood. Unfortunately, this expression is practically impossible to solve as soon as the number of alleles and the sample size become large (see, however, methods). Griffiths and Tavaré (1994) and Kuhner *et al.* (1995) were the first to propose solutions to this problem using Monte Carlo methods.

In this article, we apply a full-likelihood and coalescent-based approach to the admixture problem. We derive the likelihood function and compare results from this analytical approach with approximations obtained using the method of Griffiths and Tavaré (1994). We demonstrate the advantage of the latter. To integrate over nuisance parameters in the model (such as the ancestral gene frequencies), we then use the Metropolis-Hastings algorithm (a step for which there are no analytical results). Because of the large amount of time required, most previous full-likelihood (Bayesian) methods were tested on small data sets (usually one population, sample size ≤50). In this study we chose to simulate data sets that are closer to those currently available (*i.e.*, total sample sizes = 150 and 600, see below). We tested the performance of our method on a wide range of parameters for both sample sizes and with either biallelic loci (similar to many allozymes) or 10-allele loci (similar to microsatellite or mtDNA data). Finally, we applied the method to a published human data set.

## METHODS

**The model:** The admixture model shown in Figure 1 assumes that two independent parental populations, $P_1$ and $P_2$, of size $N_1$ and $N_2$, mixed some time $T$ in the past (measured in generations) with respective proportions $p_1$ and $p_2$ ($= 1 - p_1$), creating a hybrid population H of size $N_h$. At the time of hybridization, the gene frequency distributions of $P_1$ and $P_2$ are, respectively, the two vectors $x_1$ and $x_2$, and that of the hybrid population is $p_1x_1 + p_2x_2$. After admixture, $P_1$, $P_2$, and H evolve independently (with no migration) by pure drift (no mutations) until the present time. Even though $T$, the time since admixture (in generations) is the same for the three populations, the time scaled by the effective size of each population can be different for the three populations and is thus called $t_1 = T/N_1$, $t_2 = T/N_2$, and $t_h = T/N_h$. The parameters of the model are thus $p_1$, $t_1$, $t_2$, $t_h$, $x_1$, $x_2$. Note that Thompson (1973) analyzed the same model using a Brownian motion approximation to represent drift.

**A Bayesian approach:** We are interested in making inferences about a parameter (or a set of parameters) $\Psi$ of a statistical model by using the information provided by the observation of the data, $D$. This is given by a probability density function (pdf), which describes the probability distribution of $\Psi$ given the data $p(\Psi|D)$. We can use Bayes' theorem to write

$$p(\Psi|D) = \frac{p(\Psi)p(D|\Psi)}{p(D)}. \tag{1}$$

The first term is the pdf of $\Psi$ before the data are obtained and is therefore called the *prior* as opposed to $p(\Psi|D)$, which is the *posterior*. Practically, $p(\Psi)$ summarizes our belief, knowledge, or lack of knowledge about $\Psi$. The second term represents the probability of observing the data under the statistical model. Seen as a function of $\Psi$, $p(D|\Psi)$ ($= L(\Psi)$) is the likelihood function (Edwards 1972). The last term represents the probability of the data. It is often impossible to evaluate but is a constant given the data. As a consequence, this term can be ignored and we need only to know $p(\Psi|D)$ up to this multiplicative constant. When $\Psi$ is a set of parameters, we can obtain the distribution of any specific parameter by averaging across all others, and this is called the marginal pdf.

By taking a Bayesian (or full-likelihood) approach we consider that all relevant information about the parameter(s) is contained in the posterior pdf, and we are thus interested in the complete distribution rather than in point estimates. However, summary statistics such as point estimates can convey convenient information

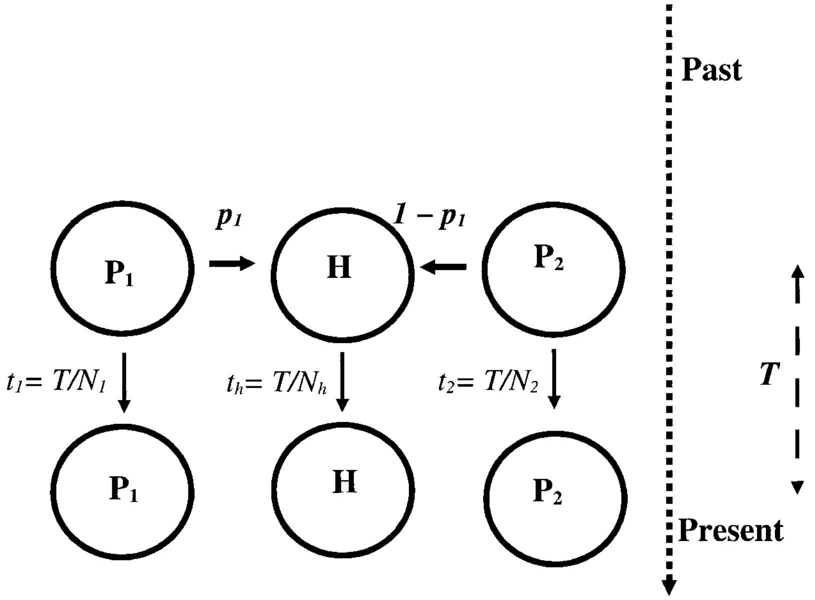

FIGURE 1.—The admixture model. We assume a single admixture event, $T$ generations ago (see text). The three populations are allowed to have different sizes $N_1$, $N_2$, and $N_h$. The contribution of parental population 1 is $p_1$.

about the pdf for comparison and are provided as well. For instance, the standard deviation (SD) is given when useful because it is a commonly used measure of dispersion. However, dispersion is better described using the width between the 5 and 95% quantiles. This is often referred to as the 95% *credible* or *equal-tail probability* interval (CI or ETPI, respectively). To avoid confusion with the 95% *confidence* interval we use ETPI. For approximately symmetric distributions, using the mean, the median, or the mode provides very similar results. However, for distributions that are highly skewed toward small values, the mode can be very difficult to estimate. This proved particularly true for the $t_i$'s. We therefore decided to use the median that is the most widely used point estimate (see GELMAN *et al.* 1995 for further discussions on the choice of a point estimator), keeping in mind that it is the full posterior pdf that we regard as relevant.

The Bayesian procedure requires that we provide a prior on all parameters of the model. Although this step may be difficult in some problems, since it involves some subjectivity (GELMAN *et al.* 1995), a lack of knowledge can be represented by a flat prior so that the posterior will in fact be proportional to the likelihood function. Because of this, we also use the term likelihood for posterior pdf in some circumstances. We chose flat priors for $p_1$, $t_1$, $t_2$, and $t_h$. For $x_1$ and $x_2$, we chose a prior in which all possible allele frequencies have equal probability; this is given by a uniform Dirichlet distribution. This choice has the advantage of making no specific assumption on how genetically distant the parental populations are and thereby encompasses any possible history of the parental populations.

**The full likelihood:** However, the posterior $p(p_1, t_1, t_2, t_h, x_1, x_2 \mid D)$ [corresponding to $p(\Psi|D)$ in Equation 1] is not available in a closed form. The likelihood function $p(D \mid p_1, t_1, t_2, t_h, x_1, x_2)$ can be written as (see O'RYAN *et al.* 1998 for details)

$$\begin{aligned} p(D|p_1, t_1, t_2, t_h, x_1, x_2) &= p(a_1, a_2, a_h|p_1, t_1, t_2, t_h, x_1, x_2) \\ &= \sum_{c_1,c_2,c_h} \sum_{f_1,f_2,f_h} ABC, \qquad (2) \end{aligned}$$

where

$$A = p(a_1|f_1)p(a_2|f_2)p(a_h|f_h)$$

$$B = p(c_1|t_1, n_1)p(c_h|t_h, n_h)p(c_2|t_2, n_2)$$

$$C = p(f_1|x_1, c_1)p(f_h|p_1x_1 + (1 - p_1)x_2, c_h)p(f_2|x_2, c_2).$$

$a_1$, $a_2$, and $a_h$ are the sample frequency counts in present-day samples of P$_1$, P$_2$, and H; $f_1$, $f_2$, and $f_h$ are the founder frequency counts in P$_1$, P$_2$, and H; $c_1$, $c_2$, and $c_h$ are the number of coalescences in the genealogical history; and $n_1$, $n_2$, and $n_h$ are the sample sizes of P$_1$, P$_2$, and H.

The first term ($A$) was first derived by SLATKIN (1996) for two alleles and by NIELSEN *et al.* (1998) for any number of alleles (Equation 9; see also O'RYAN *et al.* 1998 for an independent derivation) and represents the probability of observing a particular allelic configuration in a sample given the allelic configuration of the founders just after admixture. The second term was derived by TAVARÉ (1984, Equation 6.1) and represents the probability of observing $c_i$ coalescence events given the time (scaled by the effective size) since admixture and the sample sizes. Finally, the third term is specific to our model and represents the probability of the allelic configuration in the founders [the sample size of which is given by $n_i - c_i$ for $i = \{1, 2, h\}$] given the allelic distribution in the ancestral parental population and the amount of admixture. The summation is over the number of coalescent events in the genealogy of each population, which determines the size of the sample of founder lineages, and the number of different frequency counts for each sample of this size.

It is, however, computationally expensive to estimate this likelihood directly, because the number of allelic configurations among the founders that is compatible with the data can be very large. An alternative approach is therefore to use sampling methods to estimate the likelihood. Equation 2 can be rewritten in a more general form as

$$p(D|\Psi) = \int_{G,c} p(D|G)\ p(G|c)\ p(c|\Psi)\, dGdc, \tag{3}$$

where $G$ represents all possible genealogies and consists of a sequence of $c$ coalescent events going back from time 0 to time $T$ and where the allele frequency count among the lineages is recorded at each event. Following the notation of Stephens and Donnelly (2000), the integral denotes summation over all numbers of coalescent events and allelic configurations at each coalescent event. This rewriting becomes helpful because it is possible to sample from $p(G|c)p(c|\Psi)$ using standard methods of simulation from the coalescent (Hudson 1990). From the standard theory of Monte Carlo sampling (*e.g.*, Ripley 1987) we can then estimate (3) as the average of $p(D|G)$ for each realized $G$. Unfortunately $p(D|G)$ will be 0 for most realized $G$. To circumvent this problem we used the method introduced by Griffiths and Tavaré (1994), which proved extremely efficient in analyzing the case of pure drift (O'Ryan *et al.* 1998; Beaumont and Bruford 1999; Ciofi *et al.* 1999; see also Felsenstein *et al.* 1999 for a review).

**The method of Griffiths and Tavaré:** To circumvent the problem of analyzing all possible genealogies and allelic configurations, Griffiths and Tavaré (1994) used a Monte Carlo approach to evaluate the likelihood at specific parameter values; as noted by Felsenstein *et al.* (1999) this is equivalent to importance sampling (IS; see Ripley 1987). In this approach (see Stephens and Donnelly 2000 for extensive discussion) Equation 3 can be rewritten as

$$p(D|\Psi) = \int_{G,c} p(D|G)\frac{p(G|c)}{p^*(G|c)}p^*(G|c)p(c|\Psi)\, dGdc. \tag{4}$$

Thus (2) as can be approximated by simulating $K$ times from $p^*(G|c)p(c|\Psi)$ and estimating (4) as

$$p(D|\Psi) = \frac{1}{K}\sum_{1\ldots K} p(D|G)\frac{p(G|c)}{p^*(G|c)} \tag{5}$$

for all realized $G$ and $c$. In fact, the scheme of Griffiths and Tavaré always guarantees that $p(D|G) = 1$, because the genealogical history is constructed backward from the data as described below.

More specifically, the $G$ and $c$ are sampled according to the following scheme. If we call $S_k$ the state with $k$ lineages, the state $S_{k-1}$ is chosen (going backward in time) according to the transition probabilities

$$\begin{aligned} p(S_{k-1}|S_k) &= \frac{(n_{Ai}-1)}{(k-m)} \quad \text{if } S_{k-1} = S_k - A_i, \quad i = 1\ldots m \\ &= 0 \qquad\qquad \text{otherwise} \end{aligned} \tag{6}$$

(Griffiths and Tavaré 1994; O'Ryan *et al.* 1998), where $m$ is the number of allelic types, $A_i$ is the $i$th allele, $n_{Ai}$ is the number of $A_i$ alleles in the current state, and $S_k - A_i$ means that the allelic configuration is identical to $S_k$ apart from the fact that $n_{Ai}$ is reduced by 1. The waiting time until the next coalescent event is sampled from an exponential distribution (Kingman 1982a,b; Hudson 1990). The equivalent probability under the coalescent model for each step in the chain is $(n_{Ai} - 1)/(k - 1)$, and therefore $p(G|c)/p^*(G|c)$ can be obtained by multiplying at each step the ratio of these quantities, $(k - m)/(k - 1)$. In our model, the chain stops when the cumulative coalescence times become greater than the time of the admixture event. The state at that time represents the allelic configuration among the founder lineages and is a random draw from the ancestral frequencies of the parental populations. Therefore, to have an estimate of the likelihood of the sample, it is then necessary to multiply the final probability [the $\prod(k - m)/(k - 1)$] by the probability of observing this founding state, which is a multinomial draw from the ancestral parental frequencies. If a chain has more coalescent events than $n - k$ (*i.e.*, giving rise to fewer than $k$ founders), $p(G \mid \Psi) = 0$ by construction.

This chain is run a reasonably large number of times and the likelihood is averaged across these runs. A comparison of simulated *vs.* analytical results on small data sets and comparing results obtained with different numbers of runs on larger data sets shows that 500 runs is large enough to estimate the likelihood when drift only is considered (see appendix and O'Ryan *et al.* 1998).

To summarize, the method of Griffiths and Tavaré allows us to calculate the likelihood $p(D \mid p_1, t_1, t_2, t_{\mathrm{h}}, x_1, x_2)$ for specific values of $p_1$, $t_1$, $t_2$, $t_{\mathrm{h}}$, $x_1$, and $x_2$. Since we are interested in obtaining the posterior distribution $p(p_1, t_1, t_2, t_{\mathrm{h}}, x_1, x_2 \mid D)$ [equivalent to $p(\Psi|D)$ in Equation 1] and, in particular, some of the marginals such as $p(p_1 \mid D)$, we need a method to sample from the posterior distribution. Markov chain Monte Carlo (MCMC) is a sampling-based method that enables us to do so.

**Markov chain Monte Carlo methodology:** In Monte Carlo simulations, samples $X_i$ ($i = 1\ldots n$) of a random variable $X$ are drawn from a distribution $\pi(.)$ and then used to evaluate functions of $X$. When the distribution of interest is impossible to evaluate either because no closed form is known or because it is difficult to sample from, it is possible to construct a Markov chain having $\pi(.)$ as its equilibrium distribution. One method to do so is by using the Metropolis-Hastings algorithm (Metropolis *et al.* 1953; Hastings 1970), which is described here. If we call $X_t$ the current state of a Markov chain in the parameter space defined by the model of interest, the algorithm requires that we first choose a candidate for the next step of the chain, $X_{t+1}$, by using a proposal distribution $q(.|X_t)$. The chain then moves from state $X_t$ to the candidate $X_{t+1}$ with probability

$$\alpha = \min\left(1, \frac{\pi(X_{t+1})\, q(X_t/X_{t+1})}{\pi(X_t)\, q(X_{t+1}/X_t)}\right). \tag{7}$$

Note that we need only to be able to estimate $\pi(.)$ at some specific values and up to a multiplicative constant (*i.e.*, provided by the IS scheme above). If the candidate state is not accepted the chain remains in its current state and a new candidate state is randomly chosen from the proposal distribution. Provided that some conditions are met (irreducibility of the chain; *e.g.*, ROBERTS 1996), the proposal distribution $q(.|X_t)$ is to a large extent unimportant and the chain will sample from $\pi(.)$ once equilibrium is reached. Practically the choice of $q(.|X_t)$ is crucial if one wants the chain to reach equilibrium in a reasonable amount of time (see below). We applied the MCMC algorithm to the parameter space defined by our admixture model, *i.e.*, $p_1$, $t_1$, $t_2$, $t_{\text{h}}$, $x_1$, and $x_2$.

Different proposal (or updating) distributions were tested during the development of the method. We finally updated $p_1$ by taking a normal random deviate around $p_1$ with a standard deviation 0.05. We also found it efficient to update $p_1$ 10% of the time rather than at every step. The other parameters were updated the rest of the time. A lognormal distribution with mean $t_i$ and standard deviation $s = 1/2\sqrt{3n_{\text{loc}}}$ on a log scale was used for $t_1$, $t_2$, and $t_{\text{h}}$, where $n_{\text{loc}}$ is the number of loci. The ancestral parental allelic frequencies were updated by first selecting an allele at random, thus defining a partition of two sets of alleles: the allele itself and all the others. A β-distribution with parameters $v$ and $w$ was then used to update the chosen allele frequency. $v$ was chosen to be 1 while $1/(1 + w)$ was equated to the smallest frequency of the partition (see Appendix in CIOFI *et al.* 1999).

**Testing for convergence and analysis of the output:** A key issue in MCMC simulation is to determine when equilibrium has been reached, *i.e.*, when to stop the simulation to have a reasonable approximation of the posterior or likelihood curve. This is a serious problem, since even very long runs that appear to have converged may in fact be misleading (see STEPHENS and DONNELLY 2000 for examples). A number of diagnostic methods have been proposed (reviewed by BROOKS and GELMAN 1998), which rely on running either a number of short chains each with starting points widely dispersed within the parameter space (GELMAN *et al.* 1995) or one very long chain (RAFTERY and LEWIS 1996). We used the former method, which is based on the analysis of the variance observed for each parameter within ($V_{\text{w}}$) and between ($V_{\text{b}}$) the chains. This is done by computing $\sqrt{(V_{\text{b}} + V_{\text{w}})/V_{\text{w}}}$. GELMAN *et al.* (1995) suggest that values $<1.1$ (*i.e.*, when the variance between chains is $< \sim 5\%$ that observed within chains) are a good indication that equilibrium is reached (see BEAUMONT 1999).

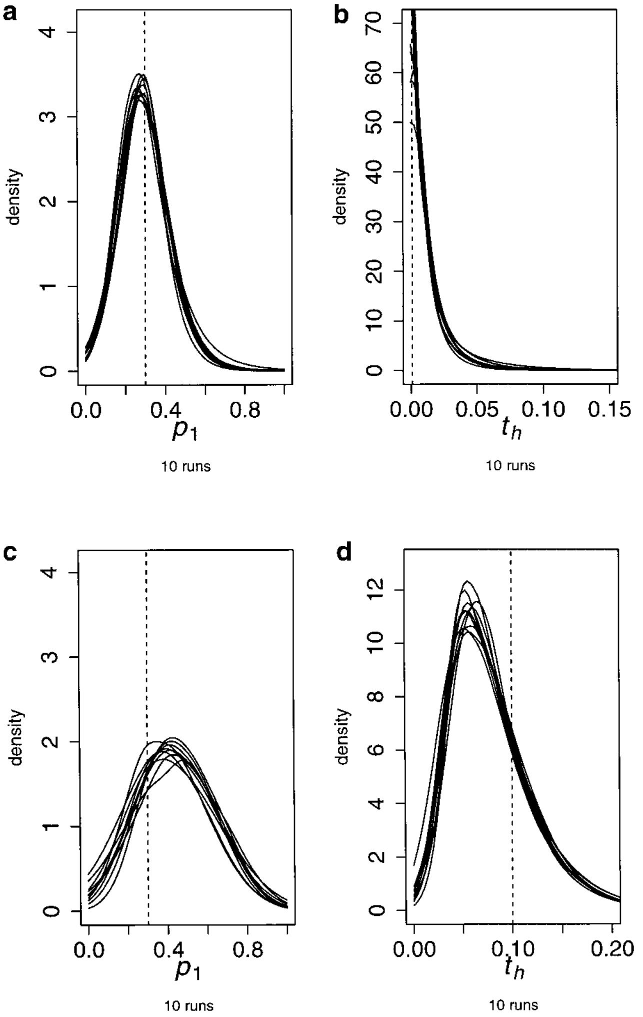

FIGURE 2.—Convergence of the MCMC for $p_1$ and $t_{\text{h}}$. The results of 10 runs are presented for $n = 200$ and for the two extreme values of $t_i$ (= 0.001 and 0.1) used in the simulations. Each curve represents the posterior pdf for 1 run. The curves are close enough to suggest that equilibrium is reached in all cases. The values of the Gelman convergence statistic were all between 1.01 and 1.06 for all parameters (see text). The pdf's are obtained using the locfit package for R. The vertical dashed lines represent the values of the parameter with which the data were simulated. (a) pdf's of $p_1$ for $t_i = 0.001$; (b) pdf's of $t_{\text{h}}$ for $t_i = 0.001$; (c) pdf's of $p_1$ for $t_i = 0.1$; (d) pdf's of $t_{\text{h}}$ for $t_i = 0.1$.

We ran 10 independent chains for independent loci for each of the three tested values of $t_i$. This was done for loci with 10 alleles and a sample size of 200 genes per population (see next paragraph for the exact procedure). In all cases, we found that running the chain for 50,000 steps was enough to produce values of the statistic $<1.1$ (see Figure 2, which represents 10 runs for $p_1$ and $t_{\text{h}}$ for $t_i = 0.001$ and $t_i = 0.1$). We did not need to repeat the diagnostic analysis for the smaller sample size ($n =$

50, see below) or number of alleles since equilibrium is reached more quickly.

For each run 10,000 points were collected for all parameters of the model (*i.e.*, 1 point every 5 steps). Following Beaumont (1999), the first 1000 points (the "burn-in") were discarded from the analysis and the 9000 remaining points were used for the convergence test and to approximate the likelihood distributions. For the multiple-loci and the human data sets longer runs were used (see below).

Unless otherwise stated, all statistical analyses were performed using the R language (Ihaka and Gentleman 1996). The likelihood curves were estimated using the program Locfit (Loader 1996) as implemented in the locfit package for R (v. 1.0). The convergence diagnostics used were performed using the coda package (v. 0.4-7) as implemented for R (ported by S. Plummer on the basis of the CODA package by Best *et al.* 1995).

**Simulating according to the model:** To test the method, we simulated data sets according to the model following a coalescent methodology. The two ancestral allele frequency distributions, $x_1$ and $x_2$, of the parental populations were simulated from two independent flat Dirichlet distributions. The allele frequency distributions of the hybrid population $x_h$ were then calculated as $p_1x_1 + p_2x_2$. We simulated the number of founders for the three populations under pure drift using a standard coalescent methodology over the intervals $t_1$, $t_2$, and $t_h$, respectively. The genetic types of the founders of the three populations were then sampled from $x_1$, $x_2$, and $x_h$. For each of the populations, a lineage was chosen randomly and duplicated until the sample size was reached. The output of these simulations was fed into a program implementing our method. Because of the huge amount of calculations involved by MCMC methods we had to limit the parameter combinations that could be analyzed. All simulations were thus performed with $p_1 = 0.3$ and by considering the same sample size for the three populations. However, the effect of sample size was investigated by using two different sample sizes ($n = 50$ and $n = 200$; *i.e.*, 150 and 600 genes from the three populations in total, respectively). Three (scaled) times since admixture were used in the simulations. For simplicity, again, the same value was used for the three $t_i$ (*i.e.*, the three populations were of the same size; see, however, discussion for a test of the effect of dissimilar sizes). We used $t_i = 0.001, 0.01, 0.1$, which for an effective size of 1000 corresponds to 1, 10, and 100 generations of drift, respectively. For each parameter combination 20 independent loci were simulated (*i.e.*, corresponding to 20 independent runs of the coalescent process). We also tested the importance of the number of alleles by using loci with either 2 or 10 alleles. For the 2-allele loci, 10 loci were simulated to reduce the time of analysis. Note that, for real data sets, there are no limitations whatsoever on the sample sizes. They can vary from locus to locus and population to population (see, for instance, the human data set analyzed). Loci with different numbers of alleles can also be used.

To summarize the principle of our approach, we used Bayes' theorem to rewrite the posterior pdf as a function of a prior and a likelihood. The likelihood was estimated at specific values of the parameter space using Griffiths and Tavaré's algorithm and a MCMC was run to obtain samples from the whole distribution. Finally, we used simulated data sets to test the accuracy of the method.

## RESULTS

**Estimation of admixture proportions from single-locus data:** Figure 3 represents the results obtained for the 20 loci of the 10-allele simulations. It shows the effect of the sample size and $t_i$ on the estimation of the admixture parameter $p_1$. The results are also summarized in Table 1 while those of the 2-allele simulations are summarized in Table 2. The numbers given in both tables represent the averages of the medians of each of the pdf's of the independent loci and the width of the 95% ETPI across the 20 loci.

For all sample sizes and numbers of alleles, the pdf's widen as the time since admixture increases. This is because the genetic information about the admixture event is gradually eroded by subsequent genetic drift in the three populations. For $t_i = 0.001$ ($n = 50$, 10 alleles) the average SD across the 20 loci of $p_1$'s posterior pdf's is 0.184 and increases to 0.198 for $t_i = 0.01$ and to 0.222 for $t_i = 0.1$. The 95% ETPI averaged across loci can be rather large and ranges from 0.71 to 0.81 as $t_i$ goes from 0.001 to 0.1 (for the $n = 50$, 10-allele case, Table 1). This indicates that the number of values that can be regarded as unlikely is in fact limited when only 1 locus is used (~20–30% of the $p_1$ values).

The effect of increasing sample size can be seen by comparing the left and right sides of Figure 3 and Tables 1 and 2. For $t_i = 0.001$ the average width of the 95% ETPI decreases from 0.71 to 0.56 and the average SD from 0.184 to 0.144 when the sample size goes from 50 to 200. For $n = 200$ the 95% ETPI reaches a value ~0.70 and the average SD becomes 0.184 only for $t_i$ somewhere between 0.01 and 0.1 (Table 1). In other words, the precision is higher for $n = 200$ than for $n = 50$, even when drift is 10–100 times as large. Note that the effect of sample size seems particularly strong for small $t_i$ values (95% ETPI of 0.56 *vs.* 0.70 for $t_i = 0.001$, as opposed to 0.77 *vs.* 0.81 for $t_i = 0.1$). It is thus worth increasing the sample size only if $t_i$ is <0.01. This means that, as drift increases, the amount of information that can be extracted about $p_1$ is quite limited even with large samples. In such cases, the only solution is to increase the number of loci (see below).

The most dramatic factor affecting the estimation of $p_1$ seems to be the number of alleles (Figure 4 and Table 2). Two-allele loci seem to provide little information on the admixture proportion even for very small values of

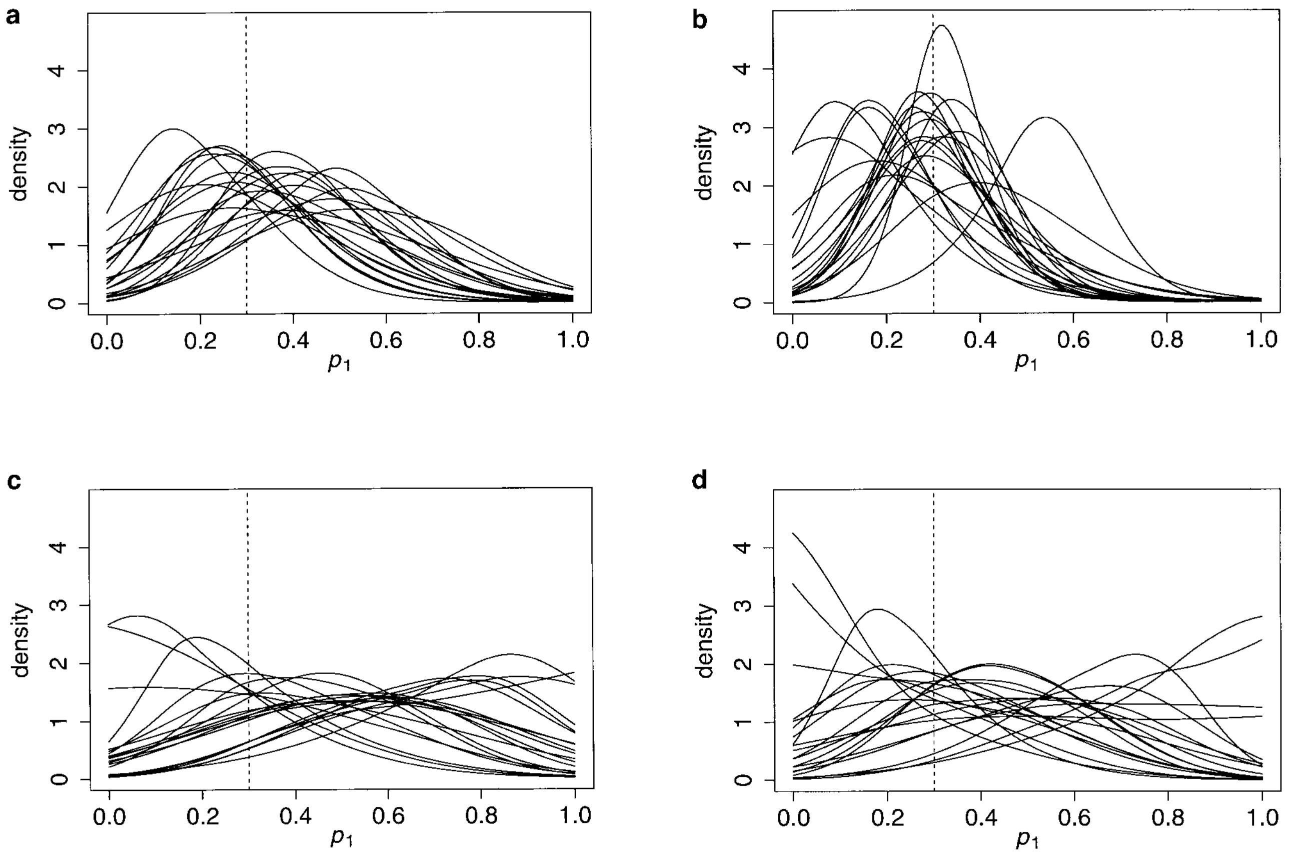

FIGURE 3.—Posterior pdf's of $p_1$ for the 10-allele case. Allelic distributions for 20 independent loci were simulated and analyzed using our method. Each curve is the posterior pdf obtained for 1 locus. The vertical dashed lines represent the values of the parameter with which the data were simulated ($p_1 = 0.3$). The parameter combinations presented here are (a) $n = 50$, $t_i = 0.001$; (b) $n = 200$, $t_i = 0.001$; (c) $n = 50$, $t_i = 0.1$; (d) $n = 200$, $t_i = 0.1$.

$t_i$. It is clearly preferable to have a single 10-allele locus after 100 times more generations of drift than one biallelic locus. With 2-allele loci it is practically impossible to exclude any value of $p_1$ as can be seen from the 95% ETPIs (Table 2), which cover nearly 95% of the possible values of $p_1$ even with $n = 200$ (*i.e.*, as one would expect if there were no data).

As should be clear from Figures 3 and 4, single point estimates such as those provided in Tables 1 and 2 should be used with caution. Even though these values indicate that the method is reliable (the estimates of $p_1$ are very close to the real value for both $t_i = 0.001$ and 0.01 and differ only moderately for $t_i = 0.1$), some single-locus pdf's can point to very different values (Figure 3). For instance, when drift is important ($t_i = 0.1$), as many as 11 of the 20 pdf's had a median >0.5, 6 of which were >0.6 and 2 of which were >0.7 (for $n = 50$). For $n = 200$ there were, respectively, six, three, and two

**TABLE 1**

**Summary statistics of the pdf's for $p_1$, $t_1$, $t_2$, and $t_h$ for the 10-allele case**

| | | $n = 50$ | | | | $n = 200$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $p_1$ | $t_1$ | $t_2$ | $t_h$ | $p_1$ | $t_1$ | $t_2$ | $t_h$ |
| $t_i = 0.001$ | Median | 0.37 | 0.041 | 0.031 | 0.020 | 0.29 | 0.022 | 0.013 | 0.008 |
| | Width 95% | 0.71 | 0.177 | 0.142 | 0.097 | 0.56 | 0.099 | 0.069 | 0.049 |
| $t_i = 0.01$ | Median | 0.40 | 0.046 | 0.035 | 0.026 | 0.28 | 0.036 | 0.025 | 0.017 |
| | Width 95% | 0.75 | 0.183 | 0.150 | 0.114 | 0.63 | 0.135 | 0.102 | 0.081 |
| $t_i = 0.1$ | Median | 0.50 | 0.101 | 0.120 | 0.126 | 0.44 | 0.132 | 0.095 | 0.097 |
| | Width 95% | 0.81 | 0.337 | 0.415 | 0.337 | 0.78 | 0.396 | 0.282 | 0.233 |

For each parameter, we provide the mean across the 20 loci of the single-locus medians and width of the 95% ETPI.

**TABLE 2**

**Summary statistics of the pdf's for $p_1$, $t_1$, $t_2$, and $t_h$ for the two-allele case**

| | | $n$ = 50 | | | | $n$ = 200 | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $p_1$ | $t_1$ | $t_2$ | $t_h$ | $p_1$ | $t_1$ | $t_2$ | $t_h$ |
| $t_i$ = 0.001 | Median | 0.49 | 0.697 | 0.697 | 0.633 | 0.49 | 0.677 | 0.663 | 0.631 |
| | Width 95% | 0.94 | 3.719 | 3.631 | 3.552 | 0.94 | 3.601 | 3.570 | 3.582 |
| $t_i$ = 0.01 | Median | 0.49 | 0.657 | 0.651 | 0.633 | 0.50 | 0.655 | 0.685 | 0.613 |
| | Width 95% | 0.95 | 3.564 | 3.535 | 3.153 | 0.95 | 3.539 | 3.657 | 3.485 |
| $t_i$ = 0.1 | Median | 0.50 | 0.670 | 0.733 | 0.692 | 0.49 | 0.723 | 0.668 | 0.633 |
| | Width 95% | 0.95 | 3.649 | 3.661 | 3.525 | 0.94 | 3.596 | 3.574 | 3.512 |

For each parameter, we give the mean across the 10 loci of the single-locus medians and width of the 95% ETPI.

loci. Clearly, the whole distribution or the 95% ETPI should be used in place of the point estimates for $p_1$.

When drift increases, it is possible for at least one of the populations to become fixed for one allele. In such cases the absence of polymorphism means that the corresponding population had either a very small size or a very large $T$. As a result large values of $t_i$ become equally likely and the MCMC cannot reach equilibrium for the corresponding $t_i$. In practice this is easily overcome by introducing a prior on the distribution of the corresponding $t_i$. We come back to this point in the analysis of the human data set. We observed this effect in the two-allele case for a few loci (1 locus for $n$ = 200 and 7 loci for $n$ = 50). As a consequence, the averages presented take into account only the parameters for which a posterior pdf was available (*i.e.*, between 7 and 10 loci depending on the parameter combination).

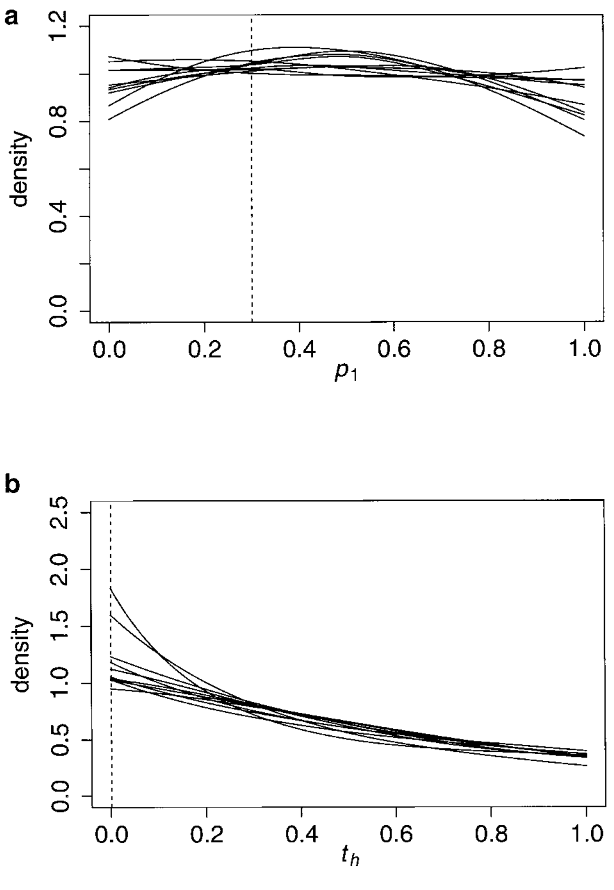

FIGURE 4.—Posterior pdf's of $p_1$ and $t_h$ for the two-allele case. Ten independent loci were used in the two-allele case (see text). (a) $p_i$'s pdf for $n$ = 200, $t_i$ = 0.001; (b) $t_h$'s pdf for $n$ = 200, $t_i$ = 0.001.

**Estimation of time since admixture from single-locus data:** Figure 5 shows the effect of drift on the estimation of $t_h$. As for $p_1$, the pdf's 95% ETPIs increase as $t_i$ increases, reaching values ∼0.3–0.4 for $t_i$ = 0.1 (in the 10-allele case, Table 1) and even 3.5 for the 2-allele cases (Table 2). In the 10-allele cases, large samples ($n$ = 200) provide more information than smaller ones ($n$ = 50) even when drift is 10 times as large. However, we do not observe a greater effect of the sample size for small $t_i$, which is similar to that observed for $p_1$. Note that in the 2-allele cases, where the amount of information is very limited, increasing the sample size has virtually no effect (Table 2) and we therefore focus on the 10-allele cases.

Another difference from $p_1$ pdf's is that the median is a rather poor point estimator of $t_i$ for small values of $t_i$ whereas it is reasonable for $t_i$ = 0.1 (Table 1). It is possible that because the pdf's of the $t_i$ are highly skewed toward zero, the maximum-likelihood estimate (MLE) should be preferred. However, regardless of the choice of a point estimator, the distributions are very wide: the 95% ETPIs are of the same order of magnitude for all $t_i$'s and are therefore more than two orders of magnitude larger than the real $t_i$ value for $t_i$ = 0.001. The simplest solution is probably to follow the full-likelihood approach and consider the whole distribution rather than point estimates. Indeed, the pdf's obtained for $t_i$ = 0.001 and $t_i$ = 0.1 are clearly different (Figure 5) even though summary statistics miss the differences.

Even if one uses the whole distribution one may wonder why these 95% intervals are so large and similar for $t_i$ = 0.001 and 0.01. A possible reason is that, for small

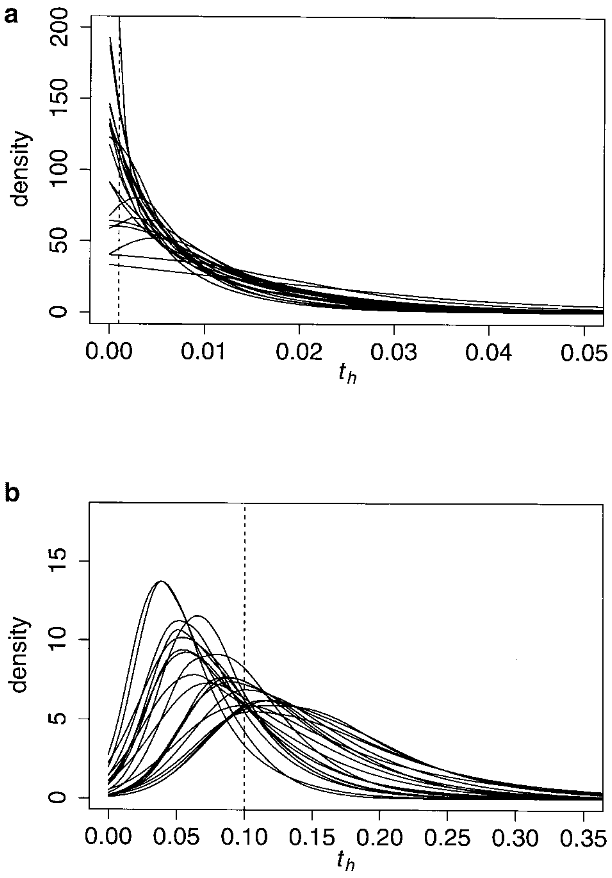

FIGURE 5.—Posterior pdfs of $t_h$ for the 10-allele case. Each curve is the posterior pdf obtained for one locus. The following parameter combinations were used: (a) $n = 200$, $t_i = 0.001$; (b) $n = 200$, $t_i = 0.1$.

$t_i$ values, the effect of sampling is no longer negligible as compared to drift. Indeed, in the 10-allele cases, when $n = 200$ the width of the 95% ETPI is much more reduced for small $t_i$'s than when $n = 50$ (Table 1). Thus, increasing the sample size does provide information on the $t_i$'s even if it does not have a great effect on $p_1$. Note that pdf's for $t_1$ are usually larger than those for $t_2$, simply because $p_1 < 0.5$ (*i.e.*, there is more genetic information on parental population 2). Even though reasonable amounts of information can be extracted from single-locus data, it appears that this is true when the admixture event is recent and both the sample size and number of alleles are large.

**Estimation of parameter values from multilocus data:** Multilocus estimation was performed for the 10-allele case using the data from 5 loci together. This was done for one sample size ($n = 50$) and three values of $t_i$, namely 0.001, 0.01, and 0.1. We used the data from the first 5 independent loci analyzed for the parameter combination (*i.e.*, 5 of the 20 loci represented in Figure 3, a and c, respectively) to compare the single- and multiple-locus results. The likelihood for multiple loci is estimated using Griffiths and Tavaré's algorithm by multiplying the likelihoods for individual loci at each step of the MCMC. In other words, loci are assumed to be independent.

Figure 6a shows the three pdf's obtained from the five-loci data together (represented by the solid lines). For comparison, the five single-locus pdf's obtained for $t_i = 0.001$ are represented by dashed lines. The increase in information is such that it is clearly better to use five loci than one locus even when the drift is >100 times as large. Indeed, the 95% ETPI of the five single-locus pdf's varies between 0.58 and 0.74 for $p_1$ whereas it is 0.23, 0.27, and 0.47 when $t_i = 0.001$, 0.01, and 0.1, respectively. In other words, the uncertainty on the real value of $p_1$ is divided by 2 to 3 depending on whether the amount of drift is equal or 10 times larger. Even when drift is 100 times larger it is still better to have five loci than one locus. To put this into perspective, in a population whose effective size is 1000, admixture will be better estimated with five loci after 100 generations of drift than it would have been with one locus such as mitochondrial DNA just one generation after the admixture event. The effect on $t_i$ is even larger with a reduction of the 95% ETPI by a factor 3 to 5. Figure 6a also shows another solid line, which was obtained using the information from five biallelic loci together for $t_i = 0.001$. Clearly, the pdf is not distinguishable from the pdf's obtained for single loci having 10 alleles. This indicates that biallelic loci such as allozyme loci can provide reasonable amounts of information on admixture events when they are used jointly. If, as appears here, five biallelic loci are approximately equivalent to one 10-allele locus, then studies using 40 allozymes such as those produced in the last decades might be comparable to studies currently using 5–10 microsatellites. This comparison is certainly very rough, but shows that precise estimates of admixture proportions can be estimated with very easily obtained genetic markers.

Figure 6b shows the apparently flat distributions obtained with single-locus data for $t_i = 0.001$. The distribution for five loci (solid line) shows an improvement but apart from pointing toward low values of $t_i$ the pdf is still flat. As was said earlier, information on the amount of drift is limited because of the inherent stochastic behavior of the coalescent and of other sources of variation. Note that the appearance of flatness is increased by the scale used to represent the six curves (see Figure 2b where the $t_i$'s are represented on another scale).

It might be thought that a convenient way of combining the information across loci would be to multiply the posterior pdf's across loci (instead of running the multiple loci data) and then renormalizing. However, these pdf's are marginals and the correct procedure would be to multiply the full pdf (*i.e.*, across all parameters) and then take the marginal. Using the full pdf from the $n$ independent loci is impractical because of the very high dimensional density estimation that would be involved. Therefore, it is necessary to run the MCMC simulations with all loci simultaneously. We found that

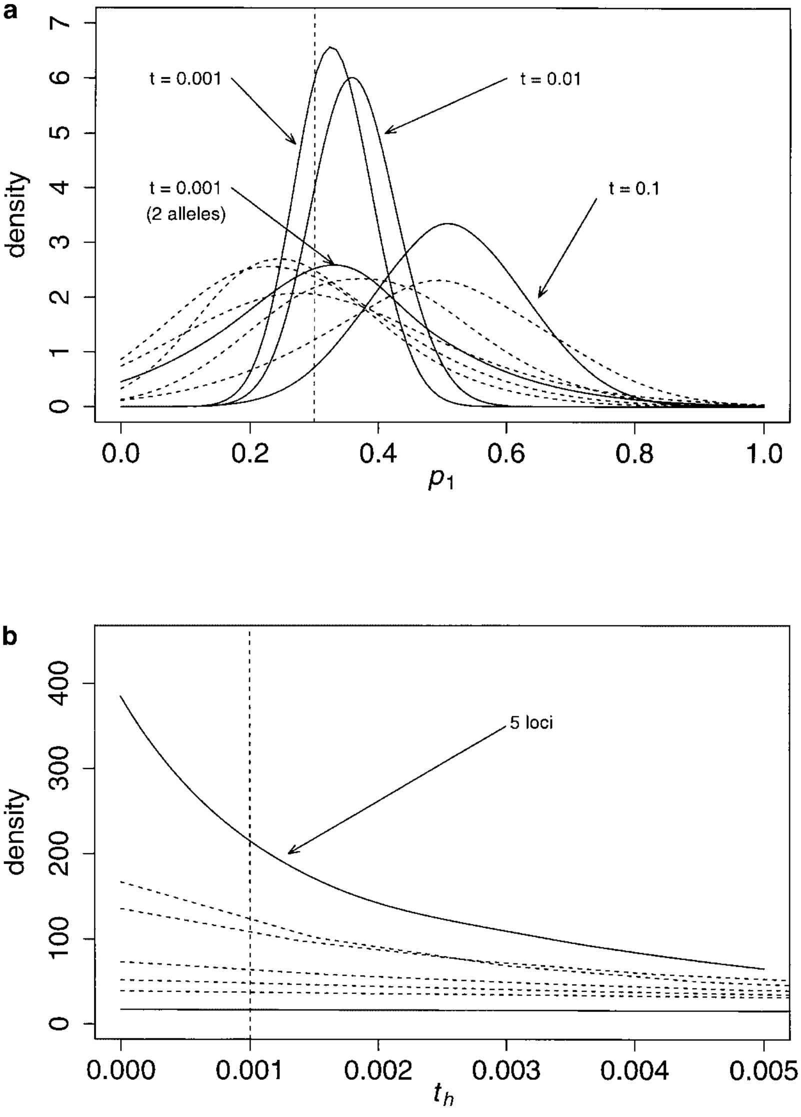

Figure 6.—Multiple-locus analysis: multiple-*vs.* single-locus pdf's. The solid lines represent the pdf's obtained with the five loci. The dashed lines represent the five pdf's obtained for the five independent loci for $t_i = 0.001$.

with multiple loci, the MCMC runs take longer to converge than expected on the basis of the number of loci. For instance, we ran the five-loci data for 350,000 instead of 250,000 steps. Even though multiple-loci data take longer to analyze, the data produced justify this extra analysis time. Also, when only single-locus data are available the whole pdf should be used.

**Effect of dissimilar sizes:** Even though the present method does not require the populations to have the same size, all previous data sets were simulated under this assumption (the three $t_i$ were always equal, see methods). It is necessary to test the method when the parental and hybrid populations have been subject to dissimilar amounts of drift and assess the effect on the estimation of the parameters. To do this we simulated data sets where the two parental populations were always of the same size and the hybrid was either 10–100 times smaller ($t_1 = t_2 = 0.001$, $t_h = 0.01$ or 0.1) or larger ($t_h = 0.001$ and $t_1 = t_2 = 0.01$ or 0.1). For each of the four parameter combinations, 3 independent loci were simulated and analyzed. Figure 7 shows some of these results for $p_1$ and $t_h$. Even though the number of loci analyzed is limited, a number of features are apparent. First, the precision on $p_1$ (Figure 7, a and c) seems to be mostly dependent on the population that has drifted most (*i.e.*, the largest value of $t_i$) whether it is the hybrid (Figure 7a) or parental (Figure 7c) population. Indeed the 95% ETPI averaged across the 3 loci is 0.68 for $t_h = 0.01$ and 0.81 for $t_h = 0.1$. These values are within the values observed for the 20 loci when the three $t_i$'s were equal to 0.01 and 0.1, respectively (Table 1). Second, Figure 7, b and d shows that when there is a 100-fold ratio between $t_h$ and both $t_1$ and $t_2$ the posterior pdf's become clearly different. When the ratio is 10 then the difference in the posterior pdf's is not as obvious and would certainly require multilocus or large sample data to be visible. This is because the pdf's on $t_i$ are wide even for small $t_i$'s as was shown before (Figure 5 and Table 1). Third, a comparison of Figure 7, a and c indicates that the pdf's for $p_1$ are thinner when the

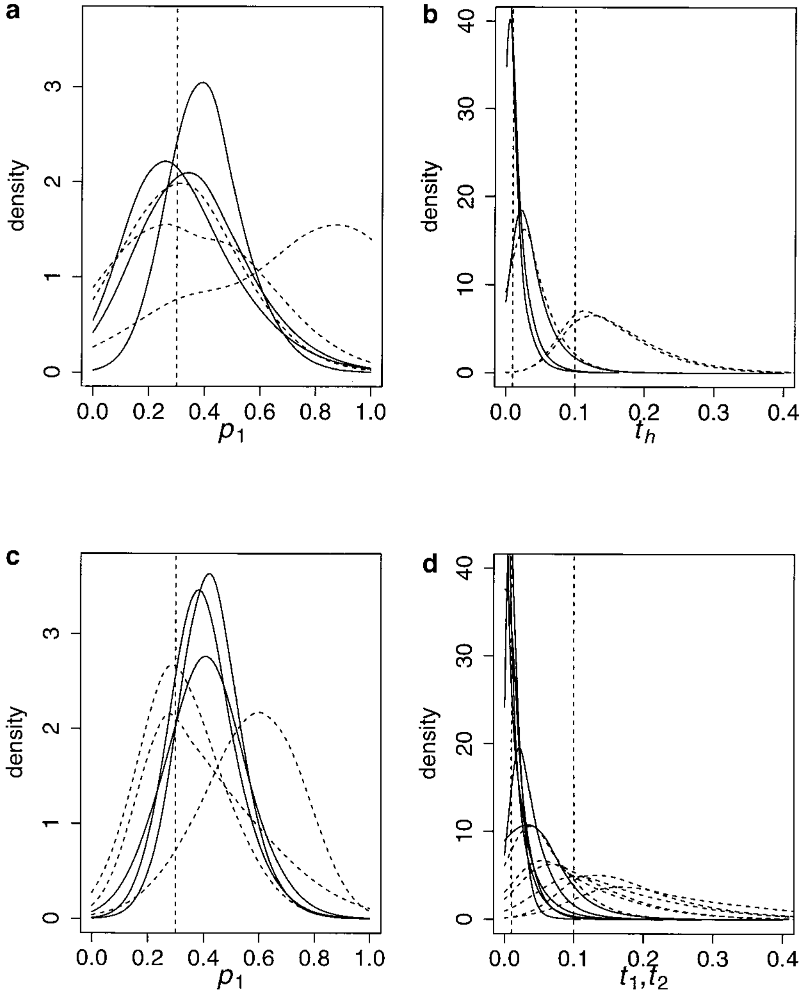

FIGURE 7.—Posterior pdf's when the hybrid and parental populations have different sizes. *Large hybrid* (c and d) cases correspond to simulations where $t_h = 0.001$ ($t_1$ and $t_2$ vary), while *Small hybrid* (a and b) cases correspond to simulations where $t_1 = t_2 = 0.001$ (and $t_h$ varies). (a) The three solid lines represent the pdf's of $p_1$ for three independent loci for which $t_h = 0.01$ while the three dashed lines correspond to $t_h = 0.1$; (b) the three solid lines are the pdf's of $t_h$ for three independent loci simulated with $t_h = 0.01$ while the dashed lines correspond to $t_h = 0.1$; (c) the three pdf's of $p_1$ for $t_1 = t_2 = 0.01$ (solid lines) and $t_1 = t_2 = 0.1$ (dashed lines); (d) as in b but showing the three pdf's of $t_1$ and $t_2$ for $t_h = 0.01$ (solid lines) and $t_h = 0.1$ (dashed lines).

hybrid is subject to little drift than when it is the parental populations that are subject to little drift. This is surprising because in our simulations two out of three populations experience large amounts of drift in the "large hybrid" cases instead of only one in the "small hybrid" cases.

## APPLICATION TO A HUMAN DATA SET

We applied the method to a data set published by PARRA *et al.* (1998). They estimated admixture proportions of European and African genes in African-American populations from the United States and from Jamaica using the methods of LONG (1991) and CHAKRABORTY (1975). Nine nuclear loci were used (APO, AT3-ID, GC, FY-null, ICAM-1, LPL, OCA2, RB2300, and Sb19.3, most of which are restriction site polymorphisms; see PARRA *et al.* 1998 for details). All were biallelic with the exception of GC, which was triallelic. The frequencies in the parental populations were obtained by pulling together three European (England, Germany, and Ireland) and three African (one from Central African Republic, two from Nigeria) samples, respectively (PARRA *et al.* 1998).

We applied the method to the Jamaican sample because it is more likely to fit our model than the other samples. The allele frequencies in the three populations are given in Table 2 (average sample size: Europeans, $n = 292$; Africans, $n = 388$; Jamaicans, $n = 186$ chromosomes; Table 3). We ran the data for the nine loci independently first (for 50,000 steps) to check for loci that could have a very different behavior, perhaps indicating selection. Then we ran the data for the nine loci together for 300,000 steps. To check for convergence, we ran the chain six times, from different starting points. We also ran one "long" chain for 600,000 steps. The first 50,000 steps of each chain (100,000 for the long one) were discarded and the analysis was done with the rest of the points after thinning, resulting in 50,000 points per run (100,000 for the long run). Each multilocus run took ~1 week on a Pentium 500 Mhz under Linux.

The single-locus data analysis was performed for all loci; but for two loci (FY-null and ICAM), at least one parental population was fixed for one of the alleles (Table 3). The absence of polymorphism despite the large sample sizes means the data are as likely to have been generated by any large value of $t_i$ and the MCMC

**TABLE 3**

**Summary of the human data set**

| | | APO | ATIII | FY-null | ICAM | LPL | OCA | RB2300 | Sb 19.3 | GC | *n* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Europe | *a* | 294 | 92 | 228 | 302 | 146 | 249 | 94 | 289 | 31 | |
| | *b* | 24 | 230 | 0 | 0 | 168 | 71 | 222 | 29 | 118 | |
| | *c* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 45 | |
| | Sum | 318 | 322 | 228 | 302 | 314 | 320 | 316 | 318 | 194 | 292.4 |
| | HET | 0.14 | 0.41 | 0 | 0 | 0.5 | 0.35 | 0.42 | 0.17 | 0.55 | |
| | $F_{st}$ | 0.04 | 0 | NA | NA | 0.01 | 0.03 | −0.01 | −0.1 | 0.01 | |
| Africa | *a* | 182 | 344 | 0 | 284 | 385 | 40 | 359 | 168 | 302 | |
| | *b* | 212 | 50 | 382 | 100 | 9 | 354 | 29 | 226 | 29 | |
| | *c* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 33 | |
| | Sum | 394 | 394 | 382 | 384 | 394 | 394 | 388 | 394 | 364 | 387.6 |
| | HET | 0.5 | 0.22 | 0 | 0.39 | 0.04 | 0.18 | 0.14 | 0.49 | 0.3 | |
| | $F_{st}$ | 0.01 | 0 | NA | 0.05 | −0.13 | 0.04 | −0.07 | 0.02 | −0.02 | |
| Jamaica | *a* | 95 | 151 | 12 | 138 | 174 | 17 | 160 | 97 | 147 | |
| | *b* | 91 | 35 | 174 | 48 | 12 | 169 | 24 | 89 | 21 | |
| | *c* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 18 | |
| | Sum | 186 | 186 | 186 | 186 | 186 | 186 | 184 | 186 | 186 | 185.8 |
| | HET | 0.5 | 0.31 | 0.12 | 0.38 | 0.12 | 0.17 | 0.23 | 0.5 | 0.35 | |

The absolute frequencies of one, two, or three alleles (*a*, *b*, *c*) are given for each locus. HET is the expected heterozygosity for each locus in each sample. *n* is the average sample size. $F_{st}$ measures the differentiation between the different samples of the same continent and was estimated as $F_{st} = \{H_T - (\Sigma H_i)/n_s\}/H_T$, where $n_s$ is the number of samples, $H_i$ is the expected heterozygosity within sample *i*, and $H_T$ is the total heterozygosity. This estimator is not unbiased but gives an idea of the amount of differentiation between samples within continents.

will not reach equilibrium for the corresponding $t_i$'s, which move to larger and larger values. Practically, one can introduce a prior on the distribution of $t_i$ during the analysis (a possibility could be to use a flat prior between 0 and some reasonable value such as $t_i = 1$ or 10) and then use the marginals of the other parameters of interest. A simpler solution is to use directly the marginals obtained from the run. Note that this situation disappears when all loci are analyzed together because large $t_i$'s become unlikely. The seven remaining loci showed similar results, with GC, the three-alleles locus, showing thinner and slightly shifted pdf's with regard to the others (not shown).

When all loci are used together the estimates of admixture proportions in the Jamaican sample are very similar to those obtained by Parra *et al.* (1998) using Long's (1991) and Chakraborty's (1975) methods, pointing to an approximate value of $p_1 \sim 7\%$ (Figure 8a; see also McKeigue *et al.* 2000). However, a look at the standard deviations obtained with the three methods shows very different results, *i.e.*, 0.2% (Chakraborty), 1.2% (Long), and 3.0% (our method). Our method seems to indicate a much greater uncertainty on the true value of $p_1$ than the two others.

By using a Bayesian approach, we obtain estimates that integrate across all possible gene frequency distributions in the parental populations. The genealogical approach allows us to take explicitly into account both drift in the three populations and the sampling process. As a consequence all factors that contribute to variance in the estimates are taken into account. Not all of these factors of variation are taken into account by the two other methods. This explains why our SD is larger than theirs. This also means that these methods underestimate the true variance and therefore provide the user with a misleading impression of precision. We are currently testing different methods of admixture estimation (including that of Long) on simulated data sets and find that our method usually performs best (lower mean square error and more accurate interval estimation; L. Chikhi, R. A. Nichols, M. W. Bruford and M. A. Beaumont, unpublished results).

Therefore, our results, while supporting the point estimates given by Parra *et al.* (1998), suggest that the true value of admixture may be within wider bounds (95% between 1.9 and 14.1%) than suggested by the use of Long's and Chakraborty's methods. McKeigue *et al.* (2000) developed a Bayesian approach to estimate individual admixture proportions and applied it to the same data set, estimating that the 95% ETPI for the Jamaican population was of 6%, that is, larger than that of Long but smaller than ours. Note that McKeigue *et al.* (2000) used 10 loci instead of the 9 we used, which makes the comparison difficult.

Figure 8 shows the pdf's obtained for $p_1$ and $t_h$ in the European, African, and Jamaican populations, respectively. The most striking result is the fact that $t_h$'s pdf indicates much smaller values for the Jamaican (95% ETPI: 0.00032–0.05243) as compared to both the African (95% ETPI: 0.00068–0.08560) and particularly Eu-

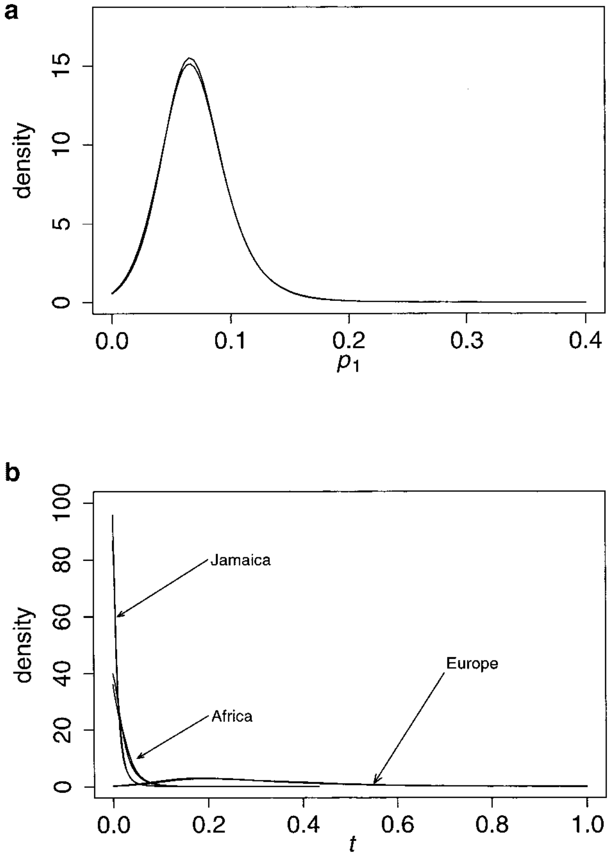

FIGURE 8.—Human data set: admixture in the Jamaican population. For each parameter, the pdf's obtained using both the long run and the combined sample of the six runs are shown. In most cases the curves are nearly indistinguishable, indicating that equilibrium is reached. (a) pdf of $p_1$ (European contribution to the Jamaican population); (b) pdf's of the $t_i$ for the Jamaican, African, and European populations.

ropean (95% ETPI: 0.04917–0.67681) populations. Given the inverse relationship between $t_i$ and the size of the populations, this result is the opposite to what one would expect. This can be interpreted in different ways. First, one could observe that the distributions of the $t_i$ overlap and therefore may not necessarily indicate different effective population sizes. This interpretation is not satisfactory because the overlap is very limited at least between the Jamaican and the European populations. This can be tested by looking at the joint distribution of the three pairs of $t_i$'s. We find indeed that $t_1 > t_h$ ($P = 0.0035$) and $t_1 > t_2$ ($P = 0.0134$) whereas there is no significant difference between $t_h$ and $t_2$. Also, we showed with simulated data sets that clear differences in the $t_i$'s pdf's appear only when the differences are large (see Figure 7, b and d).

Another interpretation is that the data were not generated according to the model. A number of assumptions of the model are certainly not met. One could argue that gene flow from European to Jamaican has taken place during the last 200 years and this may explain why the Jamaican population seems to have the largest effective size. However, the fact that three other methods used give similar values for $p_1$ indicates that this particular assumption should not be problematic. Indeed, the two methods used by PARRA *et al.* (1998) and that of MCKEIGUE *et al.* (2000) do not make any assumption on the number of admixture events and the admixture level could have been reached by constant gene flow as well.

In admixture studies, the choice of the parental populations is often crucial. In most cases the exact parental populations cannot be identified with certainty (in fact it may not even be clear whether a "hybrid" population is really admixed). In the present case, the Jamaican population is admixed and the parental populations are known to be European and African. Any pair of samples from both continents would be as good as any other if the level of population differentiation within continents were low. That is unlikely to be the case, and PARRA *et al.* (1998) were aware of this problem. To circumvent it, they used a collection of samples from different areas from both Europe and Africa and assumed that the differentiation with other samples would be negligible. This assumption was based on the fact that the different samples they used within each continent were not highly differentiated. Indeed, the $F_{st}$ estimates we find are negligible (Table 3). However, our analysis indicates that the real ancestral parental populations may have been misrepresented by present-day parental population samples. This is indirectly confirmed by MCKEIGUE *et al.* (2000) who proposed a test to detect misspecification of ancestry-specific allele frequencies. They applied it to the data of PARRA *et al.* (1998) for four of the Afro-American samples (but not to the Jamaican sample) and found significant results for AT3-ID, OCA2, and GC. MCKEIGUE *et al.* (2000) were not able to distinguish whether it is the African-specific frequencies, the European-specific frequencies, or both that are misspecified. *A priori* Africa is most likely to be misrepresented since it is the continent where the greatest amount of genetic differentiation is observed among human populations, and only samples from Nigeria and the Central African Republic were used. Also, as noted by PARRA *et al.* (1998) Angola contributed as much as the Bight of Biaffra (currently Nigeria and Cameroon) into the North American mainland (25% each, see Curtin 1969 in PARRA *et al.* 1998). Considering the geographic distance between Angola and Nigeria it should be expected that the samples used by PARRA *et al.* (1998) may not represent the original variability of the ancestral populations. A similar argument could be made for the European samples, which are all Northern European, even though the amount of differentiation is more limited than in Africa.

*A posteriori* the data point to a larger misrepresentation of European gene frequencies than African. A likely explanation is that, if only some slave owners contributed disproportionately to the Jamaican gene pool, the present-day European sample would appear to have undergone a greater degree of drift from the ancestral population. Indeed, taking a sample representing England, Ireland, and Germany may represent *more* variability than there really was in the more limited number of European ancestors of Jamaicans.

In conclusion, our method indicates that the European samples used, and perhaps the African samples as well, are unlikely to be representative of the parental populations of the Jamaican population. The effect on the final admixture estimate is difficult to predict. One way to test this is by using the information that is currently available on the geographic origin of the African slaves and European slave owners who settled in Jamaica. Samples should then be obtained from these areas in proportion to their known contributions. Parra *et al.* (1998) analyzed their data in this way, using as many samples as possible. The robustness of these estimates should be checked by excluding the samples of one or more of the areas alternatively. Although tedious, this last step appears very necessary, given the uncertainty of admixture estimates and their importance in epidemiological studies, to cite one example.

## DISCUSSION AND CONCLUSION

A number of methods have been developed to estimate admixture proportions since Bernstein's (1931) seminal paper (see review by Chakraborty 1986). Most of them usually neglect stochastic effects apart from the sampling of the hybrid population. Exceptions include the early work of Thompson (1973), who introduces drift in the estimation of the population frequencies, or Long (1991), who takes into account sampling error in all populations but drift only in the hybrid population. Recently, Bertorelle and Excoffier (1998) introduced a number of improvements by (i) explicitly parameterizing the history of the parental populations prior to the admixture event and (ii) using molecular information (*i.e.*, genetic differences between alleles and not only allele frequency information). Recently also, McKeigue *et al.* (2000) developed a Bayesian method that allows the estimation of admixture for each individual and the distribution of individual admixture in the population. This method does not take a genealogical approach, does not consider drift, and is currently limited to biallelic loci. However, it could be extended to multiple-allele loci easily (McKeigue *et al.* 2000) and could also take into account some stochasticity in the ancestry-specific allele frequencies within the present-day admixed population (P. M. McKeigue, personal communication).

The method presented here takes into account most sources of variation (sampling and drift in all populations and uncertainty over the parental allelic frequencies) that affect the estimation of the parameters. It also allows for the populations to have different sizes and therefore experience different amounts of drift (Thompson 1973 also allows for different effective sizes). Also, it is the first method that provides an estimation of the time (scaled by the population size) since the admixture event.

It is important to note, though, that two important sources of variation were not taken into account by our method: gene flow and mutations. Only Bertorelle and Excoffier's method considers the latter. The introduction of mutations to our model is theoretically possible (Griffiths and Tavaré 1994; Nielsen 1997; Beaumont 1999) but would slow the estimation of the likelihood enormously. In fact, Stephens and Donnelly (2000) recently showed that when mutations are added, the choice of the IS function can be critical. In particular they give a new IS function, which, when the mutation rate is high, can be typically orders of magnitude quicker and more accurate than Griffiths and Tavaré's. Clearly, the lack of mutations in our model can be seen as a limitation. Indeed, when any of the $t_i$ is large, mutations may not be negligible for some markers *if* the population size is large (indeed, $t_i = T/N_i$). However, for small populations, mutations will be negligible even for large $t_i$'s. Therefore, the method should be used with caution when the admixture event occurred over a time scale comparable to $1/\mu$, where $\mu$ is the mutation rate of the marker used. Our aim in this article was to test the effect of a number of important factors (*i.e.*, the number of alleles, loci, and varying sample and effective population sizes). The maximum amount of drift simulated was 0.1, which is much smaller than the expected fixation time of 2. Note that other available methods take into account neither mutations (apart from Bertorelle and Excoffier's) nor drift in all populations. Also, the advantage of considering pure drift is that we make no assumption on the mutation model that generated the variation observed in the ancestral parental populations. The method can therefore be applied to any type of marker.

It is clear that gene flow will also affect the estimates of admixture. Clearly, our method should be used only when there are good reasons to believe that gene flow was limited in comparison to the original admixture event (see Bertorelle and Excoffier 1998). Practically, one can argue, as noted by Chakraborty and others, that the admixture proportion estimated by most methods is in fact the result of the cumulative effect of gene flow across generations. It is probable that this will be the same with our method but since this was not tested thoroughly, admixture estimates should be used with care when they could be the result of continuous gene flow. Other methods that estimate gene flow should then be used instead (*e.g.*, Beerli and Felsenstein 1999).

Another source of error in the estimation of admixture proportions is the problem of nonrepresentative sampling. As for gene flow, effort should be made to use our method (and actually any method) only if there is good evidence that the parental populations are identified. We suggested in the analysis of the Jamaican data set that the robustness of admixture estimates should be checked by using different combinations of parental populations, using historical and/or biological information to identify them.

Having identified some of the weaknesses that may impair our method, it should be clear, though, that when the population conforms to expectation (*i.e.*, according to our simplified version of the world, Figure 1) our method provides reliable results both for $p_1$ and the $t_i$ when large sample sizes and multiple-loci data are used. At the same time our study shows that, even when populations do evolve according to the model, single-locus data, such as those obtained with mitochondrial DNA, should be used only with extreme caution. In any case the whole distribution should be given rather than point estimates since we have seen how they can be weak representations of the data.

Likelihood-based methods such as those developed by Griffiths and Tavaré or Felsenstein's group are computer intensive and it is often difficult to test them accurately on a wide set of parameters as is done for classical coalescent-based methods, although it is easier to do now than when the methods were developed. This means that (i) comparison with other methods is difficult and (ii) most likelihood and MCMC methods published so far were not thoroughly tested (see Stephens and Donnelly 2000 for examples). In our study, we tested for convergence of the MCMC and compared the precision of the IS algorithm with the analytical treatment (O'Ryan *et al.* 1998; see appendix). Indeed, in the drift case, unlike the methods considered by Stephens and Donnelly (2000), the likelihood can be analytically estimated for comparison with estimates from IS. Also, we compared our method with those implemented in Bertorelle and Excoffier's program Admix1 (including Roberts and Hiorns 1965; Long 1991 implemented according to Chakraborty *et al.* 1992; Bertorelle and Excoffier 1998) and found that our method is usually more precise (L. Chikhi, R. A. Nichols, M. W. Bruford and M. A. Beaumont, unpublished results).

Note that when compared to the large amount of time often required to collect the data (sampling and molecular analysis), the time required to do the runs should not be a serious limitation, particularly if a reliable estimate is required (for instance, in epidemiology or conservation biology). Practically, because of the time involved, we also strongly suggest that single-locus pdf curves should always be obtained before any multiple-loci analysis is performed (as we did for the human data set). A program called LEA (likelihood-based estimation of admixture) for performing the calculations is freely available from M.A.B. (m.a.beaumont@reading.ac.uk or at http://www.rubic.rdg.ac.uk/∼mab/).

We thank R. Nichols, E. Parra, P. McKeigue, and M. Slatkin for useful comments on earlier versions of the manuscript. We also thank D. Charlesworth and two anonymous reviewers who provided comments that greatly helped us improve the article. We are also grateful to E. Parra and M. Shriver for sending us the human data set. L.C. was funded by the Ministry of Agriculture, Fisheries and Food (MAFF contract ref. OC913, awarded to M.W.B.) and by a small National Environment Research Council grant (ref. GR9/04474).

## LITERATURE CITED

Beaumont, M. A., 1999 Detecting population expansion and decline using microsatellites. Genetics **153:** 2013–2029.

Beaumont, M. A., and M. W. Bruford, 1999 Microsatellites in conservation genetics, pp. 165–182 in *Microsatellites Evolution and Applications*, edited by D. B. Goldstein and C. Schlötterer. Oxford University Press, Oxford.

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.

Bernstein, F., 1931 Die geographische Verteilung der Blutgruppen und ihre anthropologische Bedeutung, pp. 227–243 in *Comitato Italiano per lo Studio dei Problemi della Popolazione*. Istituto Poligrafico dello Stato, Roma.

Bertorelle, G., and L. Excoffier, 1998 Inferring admixture proportions from molecular data. Mol. Biol. Evol. **15(10):** 1298–1311.

Best, N. G., M. K. Cowles and S. K. Vines, 1995 *CODA Manual Version 0.30*. MRC Biostatistics Unit, Cambridge, UK.

Brooks, S. P., and A. Gelman, 1998 General methods for monitoring convergence of iterative simulations. J. Comput. Graph. Stat. **7(4):** 434–455.

Chakraborty, R., 1975 Estimation of race admixture: a new method. Am. J. Phys. Anthropol. **42:** 507–511.

Chakraborty, R., 1986 Gene admixture in human populations: models and predictions. Yearb. Phys. Anthropol. **29:** 1–43.

Chakraborty, R., and K. M. Weiss, 1988 Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. Proc. Natl. Acad. Sci. USA **85:** 9119–9123.

Chakraborty, R., M. I. Kamboh, M. Nwankwo and E. Ferrel, 1992 Caucasian genes in American Blacks: new data. Am. J. Hum. Genet. **50:** 145–155.

Ciofi, C., M. A. Beaumont, I. R. Swingland and M. W. Bruford, 1999 Genetic divergence and units for conservation in the Komodo dragon *Varanus komodoensis*. Proc. R. Soc. Lond. Ser. B **266:** 2269–2274.

Edwards, A. W. F., 1972 *Likelihood*. Cambridge University Press, Cambridge, UK.

Felsenstein, J., 1992 Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. Genet. Res. **59:** 139–147.

Felsenstein, J., M. K. Kuhner, J. Yamato and P. Beerli, 1999 Likelihoods on coalescents: a Monte Carlo sampling approach to inferring parameters from population samples of molecular data. IMS Lect. Notes Monogr. Ser. **33:** 163–185.

Gelman, A., J. B. Carlin, H. S. Stern and D. B. Rubin, 1995 *Bayesian Data Analysis*. Chapman & Hall, London.

Goodman, S. J., N. H. Barton, G. Swanson, K. Abernethy and J. M. Pemberton, 1999 Introgression through rare hybridization: a genetic study of a hybrid zone between red and sika deer (genus Cervus) in Argyll, Scotland. Genetics **152:** 355–371.

Griffiths, R. C., and S. Tavaré, 1994 Simulating probability distributions in the coalescent. Theor. Popul. Biol. **46:** 131–159.

Hastings, W. K., 1970 Monte Carlo sampling methods using Markov chains and their applications. Biometrika **57:** 97–109.

Hewitt, G., 2000 The genetic legacy of the quaternary ice ages. Nature **405:** 907–913.

Hudson, R. R., 1990 Gene genealogies and the coalescent process, pp. 1–44 in *Oxford Surveys in Evolutionary Biology*, edited by D. J. Futuyma and J. D. Antonovics. Oxford University Press, Oxford.

Ihaka, R., and R. Gentleman, 1996 R: a language for data analysis and graphics. J. Comput. Graph. Stat. **5:** 299–314.

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

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

Kuhner, M., J. Yamoto and J. Felsenstein, 1995 Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. Genetics **140:** 1421–1430.

Loader, C. R., 1996 Local likelihood density estimation. Ann. Stat. **24:** 1602–1618.

Long, J. C., 1991 The genetic structure of admixed populations. Genetics **127:** 417–428.

McKeigue, P. M., J. R. Carpenter, E. J. Parra and M. D. Shriver, 2000 Estimation of admixture and detection of linkage in admixed populations by a Bayesian approach: application to African-American populations. Ann. Hum. Genet. **64:** 171–186.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller, 1953 Equations of state calculations by fast computing machines. J. Chem. Phys. **21:** 1087–1092.

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

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

O'Ryan, C., E. H. Harley, M. W. Bruford, M. A. Beaumont, R. K. Wayne *et al.*, 1998 Microsatellite analysis of genetic diversity in fragmented South African buffalo populations. Anim. Conserv. **1:** 85–94.

Parra, E. J., A. Marcini, J. Akey, J. Martinson, M. A. Batzer *et al.*, 1998 Estimating African American admixture proportions by use of population-specific alleles. Am. J. Hum. Genet. **63:** 1839–1851.

Raftery, A. E., and S. M. Lewis, 1996 Implementing MCMC, pp. 115–130 in *Markov Chain Monte Carlo in Practice*, edited by W. R. Gilks, S. Richardson and D. J. Spiegelhalter. Chapman & Hall, London.

Ripley, B. D., 1987 *Stochastic Simulation.* Wiley, New York.

Roberts, G. O., 1996 Markov chain concepts related to sampling algorithms, pp. 45–54 in *Markov Chain Monte Carlo in Practice*, edited by W. R. Gilks, S. Richardson and D. J. Spiegelhalter. Chapman & Hall, London.

Roberts, D. F., and R. W. Hiorns, 1965 Methods of analysis of the genetic composition of a hybrid population. Hum. Biol. **37:** 38–43.

Rogers, A. R., and H. Harpending, 1992 Population growth makes waves in the distribution of pairwise genetic differences. Mol. Biol. Evol. **9:** 552–569.

Slatkin, M., 1996 Gene genealogies within mutant classes. Genetics **143:** 579–587.

Slatkin, M., and R. R. Hudson, 1991 Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics **129:** 555–562.

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

Taberlet, P., L. Fumagalli, A. G. Wust-Saucy and J.-P. Cosson, 1998 Comparative phylogeography and postglacial colonisation of Europe. Mol. Ecol. **7:** 453–464.

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

Thompson, E. A., 1973 The Icelandic admixture problem. Ann. Hum. Genet. Lond. **37:** 69–80.

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

Communicating editor: D. Charlesworth

## APPENDIX

We compare the results obtained in estimating the likelihood either using the analytical expression derived in (2) or using the algorithm of Griffiths and Tavaré (1994) with 500 iterations. We applied both methods to the first locus (APO) from the human data set analyzed in this article. In Figure A1 we plot the likelihoods calculated from one method against those obtained for the other, in the initial 200 steps of the Metropolis-Hastings simulations. The correlation between the two series is 0.999 (the solid line is the line $y = x$). The analytical method was ~8 times slower for APO and 21 times slower for the triallelic locus GC (not shown). Simulations performed for the O'Ryan *et al.* (1998) article (M. Beaumont, unpublished data) indicated that, as the number of alleles increases, this ratio increases, making the analytical treatment less and less useful.

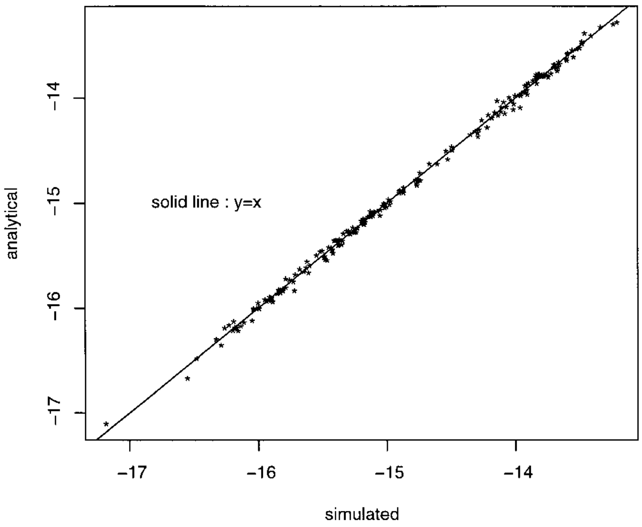

Figure A1.—Analytical *vs.* simulated estimates.