Maximum-Likelihood Estimation of Admixture Proportions From Genetic Data

Jinliang Wang[1]

Institute of Zoology, Zoological Society of London, London NW1 4RY, United Kingdom

Manuscript received December 12, 2002
Accepted for publication February 13, 2003

ABSTRACT

For an admixed population, an important question is how much genetic contribution comes from each parental population. Several methods have been developed to estimate such admixture proportions, using data on genetic markers sampled from parental and admixed populations. In this study, I propose a likelihood method to estimate jointly the admixture proportions, the genetic drift that occurred to the admixed population and each parental population during the period between the hybridization and sampling events, and the genetic drift in each ancestral population within the interval between their split and hybridization. The results from extensive simulations using various combinations of relevant parameter values show that in general much more accurate and precise estimates of admixture proportions are obtained from the likelihood method than from previous methods. The likelihood method also yields reasonable estimates of genetic drift that occurred to each population, which translate into relative effective sizes (N_e) or absolute average N_e's if the times when the relevant events (such as population split, admixture, and sampling) occurred are known. The proposed likelihood method also has features such as relatively low computational requirement compared with previous ones, flexibility for admixture models, and marker types. In particular, it allows for missing data from a contributing parental population. The method is applied to a human data set and a wolflike canids data set, and the results obtained are discussed in comparison with those from other estimators and from previous studies.

HYBRIDIZATION and the formation of admixed populations are common phenomena widely observed in various species, including humans. These frequently occur when different populations, having been isolated and differentiated over a period of time, overcome the barrier of isolation and come into contact due to various causes, such as range expansion. Many current human populations, such as those in South America (Chakraborty 1986), are actually admixed populations contributed by different ethnic groups.

An important question about an admixed population is the size of the proportional contribution from each parental population. Estimating such admixture proportions helps to clarify the historical background of admixture and is becoming useful in genetic epidemiological investigations (Chakraborty and Weiss 1986, 1988) and in assessing the risk of diseases in human populations. In conservation biology, knowledge of admixture proportions helps in making informed management of endangered species in the wild.

Since the early work of Bernstein (1931), many statistical methods have been developed to estimate admixture proportions from genetic data. All methods, except that of Bertorelle and Excoffier (1998), are based directly on the same principle that allele frequencies of the admixed population should be linear combinations of those of the contributing parental populations at the time when admixture occurs. Because those frequencies are generally unknown, but are inferred from samples taken from current parental and admixed populations, the estimation could be influenced potentially by several factors. First, estimation errors can come from sampling, since sample sizes are generally not large enough in practice. Second, except for the extreme case of sampling immediately after admixture, genetic drift occurs and changes the allele frequencies inevitably in all parental and admixed populations during the period between admixture and sampling events. If this is not accounted for, further estimation errors arise because the allele frequencies estimated from the samples refer to the current populations rather than to those when the admixture event occurred. Third, the parental populations can be genetically differentiated to various degrees before they contribute to the formation of the admixed population. The differentiation not only affects the power (precision) of different estimation methods, but also biases likelihood estimates if it is not accounted for (see below). Fourth, mutations that occurred after the admixture event can also change the allele frequencies and thus affect the estimation of admixture proportions.

To make the model and analyses tractable, however, various simplifying assumptions concerning the above factors have been adopted in previous estimation methods. Early methods either ignored all four factors completely (*e.g.*, Roberts and Hiorns 1965) or just took

[1]*Corresponding author:* Institute of Zoology, Regent's Park, London NW1 4RY, United Kingdom. E-mail: jinliang.wang@ioz.ac.uk

into account the sampling error in the admixed population (*e.g.*, Glass and Li 1953; Elston 1971). In her likelihood method that modeled drift and sampling as a Brownian motion diffusion, Thompson (1973) considered both drift and sampling effects in all parental and admixed populations. Long (1991) developed a least-squares method that allows (estimates) both sampling and drift errors in the admixed population, but only sampling error in the parental populations. More recently, Chikhi *et al.* (2001) proposed a coalescence-based likelihood method that takes into account drift since the admixture event and sampling error in all populations, allowing joint estimates of both drift effect and admixture proportions. The method developed by Bertorelle and Excoffier (1998) is the only one that took into account mutations and the genetic differentiation among parental populations before the admixture event.

The available admixture estimation methods can be broadly classified into two categories, moment and likelihood estimators. Moment estimators are generally simple to calculate, but usually have low statistical power because they use only the first two moments of a distribution (allele frequency or coalescence time) and have difficulty in weighting information (*e.g.*, among loci and sample sizes) and in accounting for the effects of drift and sampling. Likelihood estimators are more powerful if the model is appropriately specified, but can be computationally demanding. For this reason, they have not been thoroughly tested on wide ranges of parameters. Furthermore, the differentiation between parental populations is ignored in all previous likelihood estimators, which could result in bias and low precision. The purposes of this study are (1) to develop a new maximum-likelihood method to estimate admixture proportions, taking into account the effects of sampling and genetic drift in all populations and the differentiation between parental populations before the admixture event; (2) to compare the precision and accuracy of the new and previous methods, using extensive simulations; and (3) to investigate the robustness of the new and previous methods in some more realistic situations where one or more assumptions in the admixture model are violated. A human data set and a wolflike canids data set are also analyzed using different methods and the results are compared and discussed. I hope this study will be helpful in understanding the factors influencing admixture estimation and, for empiricists, in both designing admixture experiments and selecting methods for analyzing the acquired data in practice.

METHODS

The genetic model: Several genetic models on admixture are proposed and utilized in previous studies, the most recent and comprehensive one being that of Bertorelle and Excoffier (1998). This model is adopted by this study and diagrammed in Figure 1. It assumes that an ancestral population, P_0, splits into two parental populations, P_1 and P_2, which then evolve independently for ξ generations. At that point, a hybrid population, P_h, is instantaneously created by combining genes of proportions p_1 and $p_2 = 1 - p_1$ taken at random from parental populations P_1 and P_2, respectively. For a period of ψ generations between the admixture event and the current time when samples are taken, the three populations are isolated and do not exchange genes among them. I also assume, as is implicit in all previous admixture models, that neither direct nor indirect selection is associated with the markers surveyed, and the markers are from diploid and autosomal loci. Mutations at each marker locus since the admixture event are assumed to be negligible in each population. Therefore, the genetic structure of the current populations is shaped mainly by admixture and drift.

In this model, eight parameters are involved, which are the admixture proportion p_1, the two periods of time (in generations) ξ and ψ, and the average effective sizes of the parental populations P_1 and P_2 during period ξ (denoted by n_1 and n_2) and of the parental and hybrid populations during period ψ (denoted by N_1, N_2, and N_h). Rescaling time by the effective size of each population reduces the number of parameters to only six, which are p_1, $t_i = \xi/(2n_i)$ (for $i = 1, 2$) and $T_j = \psi/(2N_j)$ (for $j = 1, 2, h$). The parameter of particular interest is the admixture proportion p_1. However, the other parameters (t_i, T_j) are also important because they give information about the extent of genetic drift that occurred to each population. The data available for estimating these parameters are DNA sequences or genotypes at some marker loci assayed for three samples, one taken at random from each population at the same present time.

It should be noted that, although Bertorelle and Excoffier's (1998) method assumed the same model as shown above, it is a moment estimator and estimates the single parameter p_1 only. In addition, it assumes an equal effective size of all populations ($n_1 = n_2 = N_1 = N_2 = N_h$). All other methods ignored the differentiation between parental populations when the admixture occurred (*e.g.*, Thompson 1973; Long 1991; Chikhi *et al.* 2001). Such an assumption has little effect on moment estimators, but could result in biased estimates of p_1 for the likelihood methods assuming independent uniform priors for the allele frequency distributions of populations P_1 and P_2 when the admixture event occurs (see below).

Figure 1 shows the basic model, which can be extended to the case of three or more parental populations contributing to the admixed population in various moment methods (*e.g.*, Roberts and Hiorns 1965; Dupanloup and Bertorelle 2001). Previous likelihood methods, in contrast, adhere to the simple case of two parental populations only, at least in their published

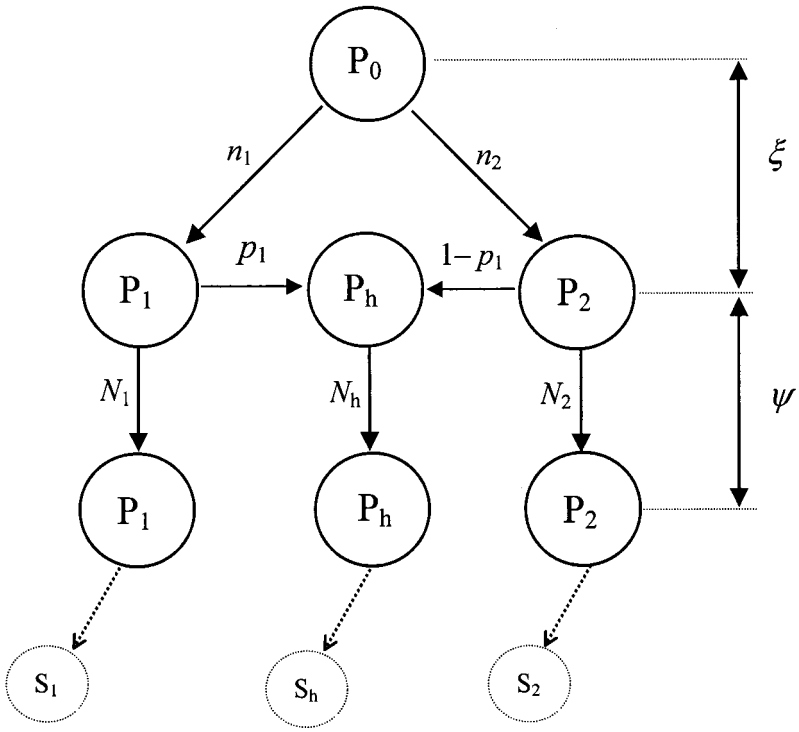

FIGURE 1.—The admixture model. It is assumed that an ancestral population, P_0, is split into two parental populations, P_1 and P_2 (with effective sizes n_1 and n_2, respectively), which evolve independently for ξ generations before they contribute genes of proportions p_1 and $1 - p_1$ to form the hybrid population, P_h. After the admixture event, P_1, P_2, and P_h with effective sizes N_1, N_2, and N_h, respectively, evolve independently for ψ generations before a sample (S_j, $j = 1, 2, h$) is taken from each of them.

versions. Herein I propose a likelihood method that can cope with any number of parental populations. The effects of violating the assumptions of the simple model, such as constant migration (gene flow) from parental to hybrid populations after the initial admixture event, are also investigated through simulations.

The likelihood method: For the time being, I assume that a single biallelic locus is genotyped for each sample taken at random from each of the three present populations. Among the S_j sampled genes (sample size), the count of a particular allele is c_j for the sample obtained from population j ($j = 1, 2, h$). Denote the allelic counts by a vector $\mathbf{C} = (c_1\ c_2\ c_h)$, the set of parameters being estimated by $\Omega = \{p_1, t_1, t_2, T_1, T_2, T_h\}$, the allele frequency of the ancestor population P_0 by w, and the allele frequencies of the parental and hybrid populations by x_j when admixture occurs and by y_j when samples are taken ($j = 1, 2, h$). The likelihood of the parameters, given data $\mathbf{C}$, can be derived as

$$\begin{aligned}\Pr(\mathbf{C}|\Omega) = &\int_w\int_{x_1}\int_{x_2}\int_{y_1}\int_{y_2}\int_{y_h}\Pr(\mathbf{C}|y_1, y_2, y_h)\\ &\times \Pr(y_1, y_2, y_h|p_1, T_1, T_2, T_h, x_1, x_2)\\ &\times \Pr(x_1, x_2|t_1, t_2, w)\Pr(w)\,dy_h dy_2 dy_1 dx_2 dx_1 dw. \quad (1)\end{aligned}$$

The probability of obtaining allelic count c_j in a sample of S_j genes, given population allele frequency y_j ($j = 1, 2, h$), is binomial and is given by

$$\Pr(c_j|y_j) = \frac{S_j!}{c_j!(S_j - c_j)!}y_j^{c_j}(1 - y_j)^{S_j - c_j}. \quad (2)$$

Because the sampling events are independent among populations, we have

$$\Pr(\mathbf{C}|y_1, y_2, y_h) = \prod_j \Pr(c_j|y_j), \quad (3)$$

where $\Pr(c_j|y_j)$ for $j = 1, 2$, or h is calculated by (2).

The initial allele frequency of the hybrid population is a linear combination of those of the two parental populations, $x_h = p_1x_1 + (1 - p_1)x_2$. The probability of y_j given x_j is determined by T_j only; therefore, we have

$$\Pr(y_1, y_2, y_h|p_1, T_1, T_2, T_h, x_1, x_2) = \prod_j \Pr(y_j|T_j, x_j). \quad (4)$$

The probability $\Pr(y_j|T_j, x_j)$ can be calculated by the diffusion approximation (CROW and KIMURA 1970, pp. 382–386). However, the probability density, expressed as an infinite series involving the hypergeometric function, is difficult to calculate quickly and accurately when T_j becomes small because of the slow convergence of the series. I use instead the transition matrix method with some computing tricks to reduce computational intensity (WANG 2001) for calculating $\Pr(y_j|T_j, x_j)$. With the diffusion approximation, the choice of the scaled effective population size (Λ_j) and scaled time [Γ_j, so that $T_j = \Gamma_j/(2\Lambda_j)$] for the transition matrix method is not important as long as Λ_j is not very small. For $T_j = 0.005$ as an example, one generation with $\Lambda_j = 100$ and 10 generations with $\Lambda_j = 1000$ will make little difference in calculating $\Pr(y_j|T_j, x_j)$. The computational load, however, increases dramatically with increasing values of Λ_j.

Similarly, we have

$$\Pr(x_1, x_2|t_1, t_2, w) = \prod_i \Pr(x_i|t_i, w), \quad (5)$$

where $\Pr(x_i|t_i, w)$ (for $i = 1, 2$) is calculated using the same method as $\Pr(y_j|T_j, x_j)$.

Finally, $\Pr(w)$ is generally unknown. Without prior information, I assume that any starting allele frequency of the ancestor population, w, is equally likely. Such a uniform distribution for w is chosen because no additional parameters need to be specified.

Even though each term on the right side of (1) can be calculated as shown above, the evaluation of this likelihood function is still problematic in computation due to the multiple integration. A discrete approximation can achieve satisfactory results with computation dramatically reduced. The likelihood function with the discrete treatment is

$$\Pr(\mathbf{C}|\Omega) = \sum_{w,x_1,x_2,y_1,y_2,y_h} \Pr(w) \prod_{j=1}^{h} \Pr(c_j|y_j) \times \prod_{j=1}^{h} \Pr(y_j|T_j, x_j) \prod_{i=1}^{2} \Pr(x_i|t_i, w), \quad (6)$$

where allele frequencies w, x_j, and y_j now take equally spaced discrete values between 0 and 1. Obviously the larger the number of the discrete values one uses in computing (6), the more accurate the approximation is but the more computation it requires. For w, I use a scaled effective size λ_0, so that w has $2\lambda_0 + 1$ equally spaced possible values between 0 and 1. For x_i ($i = 1$, 2), the scaled effective size (λ_i) is λ_0 if $t_i > 1/(2\lambda_0)$ and is $1/(2t_i)$ otherwise. For y_j ($j = 1$, 2, h), the scaled effective size (Λ_j) is $1/(2T_j)$ if $T_j < 0.05$ and is $4/(2T_j)$ otherwise. Because of the large number of replicates in the simulations described below, I use in general $\lambda_0 = 50$ in (6) except where sample size is large. Simulation results (below) show that the estimates of admixture proportions and drift change little with λ_0 when $\lambda_0 \geq 50$. To be conservative in application to real data sets, λ_0 can take much larger values. For the analyses of the human data set and wolflike canids data set, I use $\lambda_0 = 500$ and $\lambda_0 = 250$, respectively.

To further reduce computation, (6) is calculated as follows. First, the joint probability of x_1 and x_2 given t_1 and t_2 is

$$\Pr(x_1, x_2|t_1, t_2) = \sum_{w} \Pr(w) \prod_{i=1}^{2} \Pr(x_i|t_i, w). \quad (7)$$

For each possible value of w, the distribution of x_i given t_i, $\Pr(x_i|t_i, w)$, can be calculated by the transition matrix. Because x_1 and x_2 are independent given w, the joint probability is simply $\Pr(x_1, x_2|t_1, t_2, w) = \Pr(x_1|t_1, w) \times \Pr(x_2|t_2, w)$. Summing the joint probabilities weighed by $\Pr(w)$ over all possible w values, we obtain the left side of (7), which is independent of nuisance parameters w and y_j.

Second, define a composite quantity as

$$\gamma(c_j|x_j, T_j) = \sum_{y_j} \Pr(y_j|x_j, T_j) \Pr(c_j|y_j) \quad (8)$$

for $j = 1$, 2, and h and calculate it in a way similar to (7) for each possible value of x_j. Note that $x_h = p_1 x_1 + (1 - p_1) x_2$ for the admixed population.

Third, the likelihood is calculated by

$$\Pr(\mathbf{C}|\Omega) = \sum_{x_1,x_2} \Pr(x_1, x_2|t_1, t_2) \prod_{j=1}^{h} \gamma(c_j|x_j, T_j), \quad (9)$$

reducing the integration over six dimensions to effectively the summation over two dimensions. The algorithm reduces computation dramatically especially when combined with Powell's quadratic convergence method (Press *et al.* 1996) for searching the maximum likelihood. The quantities in (7) and (8) can be calculated once but stored and used many times in likelihood evaluation as long as the values of the parameters involved stay unchanged.

Another problem in computation is to search for the maximum likelihood. With several parameters in the likelihood function to be estimated simultaneously, more than one maximum could exist on the likelihood surface. To avoid the local maxima and find the global maximum is an acknowledged problem in computing science. One can adopt the simulated annealing method, which uses the Metropolis algorithm (Metropolis *et al.* 1953) to find the global maximum likelihood (Press *et al.* 1996). This method is, however, very time consuming and not suitable for simulation studies where many replicates are necessary. Alternatively, I use Powell's quadratic convergence method (Press *et al.* 1996) with several different starting points to maximize the likelihood function and pick the largest likelihood as the global maximum. The larger the number of starting points one uses, the more confident one is about the global maximum likelihood obtained. In the simulations I use 10 random starting points except where specified.

For this high-dimensional ($3d$ parameters for d parental populations) likelihood model, finding the confidence interval (C.I.) is also a problem. Here I use the profile log-likelihood as an approximation (Azzalini 1996). Consider p_1 as an example and denote the parameter subset $\Omega_p = \{t_1, t_2, T_1, T_2, T_h\}$. If $\hat{\Omega}_p$ denotes the value of Ω_p, which maximizes the log-likelihood l for a given value of p_1, we define $l^*(p_1) = l(p_1, \hat{\Omega}_p)$ as the profile (maximized) log-likelihood function of p_1. In a number of ways, the profile log-likelihood behaves like a proper log-likelihood and can be used for likelihood-ratio tests or for finding the C.I. For our example, the maximum of l^* is achieved at p_1, the usual maximum-likelihood estimator (MLE) of p_1, and the 95% C.I. is obtained by finding the values of p_1 that result in a decrease of l^* by 2 below the maximum $l^*(\hat{p}_1)$. The 95% C.I.'s for other parameters are found similarly.

For a multiallelic locus, I follow my previous approximation treatment (Wang 2001), which converts a k-allelic locus into k biallelic "loci." The dependence of such converted loci is accounted for in calculating the likeli-

hood (for details see Wang 2001) by using an appropriate weight. For several marker loci, the overall likelihood is their product for each locus, if these loci are statistically independent. The pertinence for the treatment of multiallelic loci is verified by checking the accuracy and confidence intervals of the estimates, using simulated data.

Dominant markers can also be used either separately or in conjunction with codominant markers in the estimation. For biallelic dominant markers such as restriction fragment length polymorphisms (RFLPs) and assuming Hardy-Weinberg equilibrium, (2) can be replaced by

$$\Pr(c_{j\mathrm{R}}|y_j) = \frac{(1/2S_j)!}{c_{j\mathrm{R}}!(1/2S_j - c_{j\mathrm{R}})!}(y_j^2)^{c_{j\mathrm{R}}}(1 - y_j^2)^{1/2S_j - c_{j\mathrm{R}}}, \quad (10)$$

where $c_{j\mathrm{R}}$ is the observed number of individuals with the recessive phenotype in the jth sample and y_j now refers to the recessive allele frequency in population j.

The likelihood function for two parental populations (Equation 1, 6, or 9) can be extended straightforwardly to allow for three or more parental populations contributing to the hybrid population. The assumption is that they all contribute at once to the hybridization (Dupanloup and Bertorelle 2001). However, as is shown below, this assumption is not important for estimating admixture proportions for the present likelihood method. With d contributing parental populations, $3d$ parameters are to be estimated jointly from the likelihood method. These are $d - 1$ admixture proportions, d scaled times between population split and admixture events (t_i, $i = 1 \sim d$), and $d + 1$ scaled times between population admixture and sampling events (T_j, $j = 1 \sim d$, h).

The likelihood method above can also be extended to the situation where no sample is available from a parental population contributing to the hybrid population. In such a situation, the likelihood method can still use the incomplete samples for admixture estimation. If there is no sample from the dth parental population, the likelihood can be calculated by (9) by simply setting $\Pr(c_d|y_d) \equiv T_d$. Missing data for some marker loci from a parental or hybrid population can be treated similarly. The likelihood method therefore does not require that each sample must be taken from the hybrid population and each of all parental populations and that each sample must be genotyped for the same set of loci.

Comparison of methods: The likelihood method described above is compared in performance with three previous methods using simulations. These methods have recently been compared by Bertorelle and Excoffier (1998).

Roberts and Hiorns (1965) proposed a least-squares method for estimating the single parameter of p_1, ignoring the potential effects of genetic drift, sampling, and the differentiation between parental populations on the estimation. Long (1991) developed another least-squares method that takes both drift and sampling errors in the hybrid population but only sampling error (not drift) in the parental populations into account. The method was later elaborated by Chakraborty *et al.* (1992), who provided a closed-form expression of Long's estimator in the case of parental allele frequencies being known without drift or sampling error. In this most widely implemented version of Long's method, which I use herein in the comparison, both p_1 and T_h can be estimated. This method does not allow for three or more parental populations.

Bertorelle and Excoffier (1998) proposed two moment estimators of p_1 based on the average coalescent times for a pair of alleles taken from within and between populations, and one (their m_Y) was found to be much better than the other on the basis of simulation studies. Recently this estimator has been extended to the case of more than two parental populations (Dupanloup and Bertorelle 2001). This estimator, fundamentally different from the other methods, which are all allele-frequency based, uses information about the molecular difference as well as the frequencies of the alleles. Mutations are also taken into account in the estimator. It carries, however, additional assumptions compared with frequency-based methods. Some of the assumptions were stated explicitly, such as mutation models specific to the markers (the infinite-site model for DNA sequences or RFLPs and single-stepwise mutation model for microsatellites), but others were implicitly made, such as no recombination within a DNA sequence.

Finally, it is notable that a coalescence-based likelihood method was recently proposed by Chikhi *et al.* (2001), which can estimate p_1 and T_j jointly. However, the possible correlation in allele frequencies between parental populations when the admixture event occurs is not taken into account. Because of the high computational demand of this method, it is very difficult to compare its performance with that of other methods in large-scale simulation studies.

Hereafter, the estimator of Roberts and Hiorns (1965) is designated as RH, that of Long (1991) and Chakraborty *et al.* (1992) as LC, that of Bertorelle and Excoffier (1998) and Dupanloup and Bertorelle (2001) as BD (their m_R estimator), and that of the new likelihood estimator (Wang) as W. The performance of each estimator is indicated by the bias (deviation of mean estimates from the true parameter value used in generating the simulated data) and the root mean square error (RMSE, which is the square root of the sum of variance and squared bias) obtained from a large number of replicate runs for a given set of parameters in simulations.

Simulations: Monte Carlo simulations were run to generate data sets for comparative analyses among different estimation methods. Following the coalescence approach (Hudson 1990) and the genetic model of admixture

shown in Figure 1 (except for an additional parameter n_0, the effective size of population P_0), the genealogies of $d + 1$ samples (one from the hybrid and one from each of the d parental populations) of S genes were reconstructed until the most recent common ancestor of all $(d + 1)S$ sampled genes was found. Poisson-distributed mutations were then introduced in the reconstructed tree, assuming the infinite-site model (ISM) and the single stepwise mutation model (SMM) for the simulation of DNA sequences and microsatellites, respectively. The mean coalescence time (scaled by mutation rate) was estimated by the mean number of pairwise differences for DNA sequences or by the mean squared difference between numbers of repeats for microsatellite data and was inserted into the molecular estimator to obtain admixture estimates. For the frequency-based methods, different alleles were identified and their counts were made from the original DNA sequence or microsatellite data. The allele counts or frequencies were then used in the estimation.

The numbers of replicates for different sets of parameters vary in simulations, depending on the variation of the estimates. For those sets of parameter values resulting in more variable estimates, therefore, more replicates were run to obtain reliable summary statistics. In some replicates, especially in the case of small mutation rates, it may be impossible to compute an estimate of p_1 from the BD, LC, or RH estimator because it is undefined (*e.g.*, the denominator is zero). In addition, an estimate of p_1 ($\hat{p}_1$) from the BD, LC, or RH estimator is occasionally too large or small. In simulations, a replicate is discarded and replaced by another whenever one of the BD, LC, and RH estimators is undefined or gives an estimate of $\hat{p}_1 > 100$ or $\hat{p}_1 < -100$. In contrast, the new likelihood method can always yield estimates of p_1, t_i, and T_j in the reasonable ranges as long as the marker is polymorphic. Because of the discard of replicates in favor of BD, LC, and RH estimators, the small value of λ_0, and the small number of initial points used in likelihood estimation to reduce the computation due to the large number of replicates, the performance of the likelihood method shown by simulations could be slightly conservative.

For comparison with the molecular estimator of admixture, only molecular markers are included in simulations. These markers are microsatellites and DNA sequences. The mutation rate for a microsatellite locus, or the global mutation rate for a whole DNA sequence, is denoted by U. Extensive simulations were run to compare the performance of different estimators under various combinations of p_1, t_i, T_j, S, U, and L (number of loci). Because of the large number of parameters involved, it is not possible to consider all parameter combinations in simulations. The basic set of parameter values for most of the simulations is chosen as $n_i = N_j = 5000$ ($i = 0, 1, 2$; $j = 1, 2, h$), $\xi = 5000$, $\psi = 100$, $p_1 = 0.2$, $L = 1$, and $U = 10^{-4}$ for a DNA sequence. The parameters scaled by population size are therefore $t_i = 0.01$, $T_j = 0.5$, and $\theta = 4NU = 2$. The effect of each parameter on the admixture estimation is investigated by changing the parameter in question over a wide range of values in simulations.

The basic simulation procedure described above was also extended to allow for the violations of the assumptions regarding the ideal mutation models or admixture process, and the robustness of different estimators was tested. More details are described in RESULTS.

RESULTS

Checking the likelihood method by simulations: To reduce the computational demand for the likelihood function, I made some approximations such as the discrete treatment, the multiallelic conversion, and the diffusion approximation using the transition matrix. How well do they work? Herein they are checked by simulated data sets generated using the particular parameter combination $n_0 = n_1 = 5000$, $n_2 = 25{,}000$, $N_1 = 5000$, $N_2 = 25{,}000$, $N_h = 500$, $\xi = 5000$, $\psi = 100$ (so that the scaled drift parameters are $t_1 = 0.5$, $t_2 = 0.1$, $T_1 = 0.01$, $T_2 = 0.002$, $T_h = 0.1$), $p_1 = 0.2$, $L = 10$, and $U = 10^{-4}$ for a DNA sequence. Using this parameter set, 500 simulated data sets are generated. From each data set, the likelihood estimates of p_1, t_i, and T_j are obtained using various values of λ_0 (the parameter for the discrete treatment and for scaling the effective size used in the transition matrix). The correlation coefficients between the estimates obtained with different values of λ_0 are calculated for each parameter. Except for T_1, which is poorly estimated, the estimates are highly correlated, with correlation coefficients being >0.93 once $\lambda_0 \geq 100$. Almost identical estimates of each parameter are obtained using $\lambda_0 = 250$ and $\lambda_0 = 500$. The computational load increases quickly with λ_0, however. In the following comparative analyses among methods on simulated data, therefore, I generally use $\lambda_0 = 50$ in the likelihood method to reduce computation. Its performance (precision and accuracy) indicated by the simulation results can be slightly conservative as a result.

The 95% C.I.s for each parameter were obtained using the profile log-likelihood from 1000 data sets simulated with the above parameter values. For p_1, 94% of the estimated 95% C.I.s cover the true p_1 value of 0.2 used in simulations. The means (SDs) of the MLE and lower and upper limits of the 95% C.I. are 0.204 (0.085), 0.073 (0.060), and 0.363 (0.096), respectively. In contrast, the 95% C.I.s for p_1 from moment estimators obtained by bootstrapping over loci are too narrow; only ~80–87% of the estimated 95% C.I.s cover the true value $p_1 = 0.2$. This is expected because the number of loci is small, which is unfortunately the usual case in practice. With a small number of resampling units (loci herein), bootstrapping underestimates the variance of estimates and gives too narrow a distribution of the

estimates. Similar results of coverage of 95% C.I.s are obtained for T_h. The likelihood estimates for other parameters are slightly biased and much noisier, and their estimated 95% C.I.s can be slightly either too narrow or too broad. It seems that the 95% C.I.s from the likelihood method using the profile log-likelihood are appropriate if the parameter in question is estimated without bias. The results also justify the treatment of multiallelic loci adopted in the likelihood method.

Simulation results for the basic admixture model: *The effects of population properties:* The effect of the time of isolation between ancestral populations, t_i, on the admixture proportion estimation of different estimators is shown in Figure 2, A and B. When the isolation is very short, $\hat{p}_1$ is biased toward 0.5 for all the four estimators. With an increasing t_i, however, all estimators except LC become essentially unbiased. The LC estimator is always biased for the entire range of t_i values. The bias is presumably due to the rare alleles present in the samples, to which the LC estimator is very sensitive (see below).

The RMSE of $\hat{p}_1$ decreases with increasing t_i. The RMSE of the three frequency-based methods declines rapidly with t_i initially, but tends to attenuate when t_i is ~1. In contrast, the RMSE of the molecular estimator decreases almost linearly with t_i (both axes in logarithm scale) for the whole range of t_i. As a result, the three frequency-based methods are much better than the molecular estimator when t_i is small, but are similar to the molecular estimator when t_i is large. Among the four methods in comparison, the likelihood method has consistently the smallest RMSE.

The above results for the effect of t_i are expected. Obviously, any method for estimating p_1 relies on the genetic differences between ancestral populations. With a small t_i, the ancestral populations are little differentiated genetically, leaving all estimators little power for p_1 estimation. Although the large bias of the estimators for the case of small t_i is caused mainly by the lack of relevant information from the samples, the estimators themselves seem also to have some inherent effects. The magnitude of decrease in bias by increasing marker information is dramatically different among the four estimators. Using the same parameters as in Figure 2, A and B, except $L = 10$ for the case of $t_i = 0.03125$, the means (RMSEs) of $\hat{p}_1$ are 0.32 (0.36), 0.35 (0.19), 0.34 (0.18), and 0.25 (0.14) for BD, LC, RH, and W estimators, respectively. Compared with the single-locus ($L = 1$) results shown in Figure 2, A and B, use of multiple loci decreases the RMSE of all four estimators, but decreases the bias of the W estimator only.

When t_i is ~ 1, the distribution of allele frequency in the ancestral population tends to be flat (Crow and Kimura 1970), and therefore the RMSE of any frequency-based estimator changes little with further increase in t_i. In contrast, the molecular difference between ancestral populations increases indefinitely with t_i and therefore the RMSE of the molecular estimator keeps decreasing with t_i.

The current samples have some information about t_i (see Table 1), the genetic drift of the ancestral populations during the period between population split and admixture events. If the genetic differentiation among ancestral populations is not accounted for in the likelihood model by assuming independent uniform priors for the allele frequency distributions of different parental populations, then the p_1 estimates are biased away from 0.5 (toward 0 if $p_1 < 0.5$ and toward 1 if $p_1 > 0.5$). For the case of $t_i = 0.5$ in Figure 2, A and B, for example, the mean (RMSE) of likelihood estimates of p_1 is 0.208 (0.101) if the relation between ancestral populations is accounted for and is 0.122 (0.155) if ignored. The corresponding values for the case of $t_i = 1$ are 0.204 (0.085) and 0.117 (0.121), respectively. Similar results are obtained for other parameter combinations. The likelihood estimates of p_1 are not only biased but also noisier if differentiation among ancestral populations is neglected compared with those with the differentiation taken into account. This is because a uniform prior is not without information, and independent uniform priors for parental populations strongly dictate that the populations have diverged long enough (say, $t_i > 2$) that their allele frequencies are uncorrelated. The impact of the false priors becomes important with a decreasing amount of information from data. In the extreme case of the same sample size and sample allele frequency for the parental and hybrid populations, the adoption of independent priors will tend to give a p_1 estimate of either 0 or 1, while allowing for the differentiation between parental populations can yield a p_1 estimate of any value in the range [0, 1]. Although neglecting the differentiation between parental populations reduces the number of parameters (from $3d$ to $2d$) being estimated and thus computational load dramatically, it is not justified because of the large bias and noise introduced by assuming independence between parental population allele frequencies.

The drift occurring in each population after admixture has a negative effect on the estimation of admixture proportions. Figure 2, C and D, shows the effect of T_j on the mean and RMSE of p_1 estimates from different estimators. In general, LC has the largest bias while BD has the highest RMSE. When $T_j > 0.1$, the mean of p_1 estimates from the LC estimator becomes smaller than zero and is therefore not shown in Figure 2C. The likelihood estimates of p_1 have consistently the lowest RMSE and the smallest bias as well when $T_j < 0.1$.

In contrast to the genetic drift and mutation that occurred before the admixture event (t_i), which increases the power of the estimators, the drift and mutation that happened to the parental and hybrid populations after the admixture (T_j) reduces the power for estimating p_1. This is because the current samples become less and less informative about the admixture

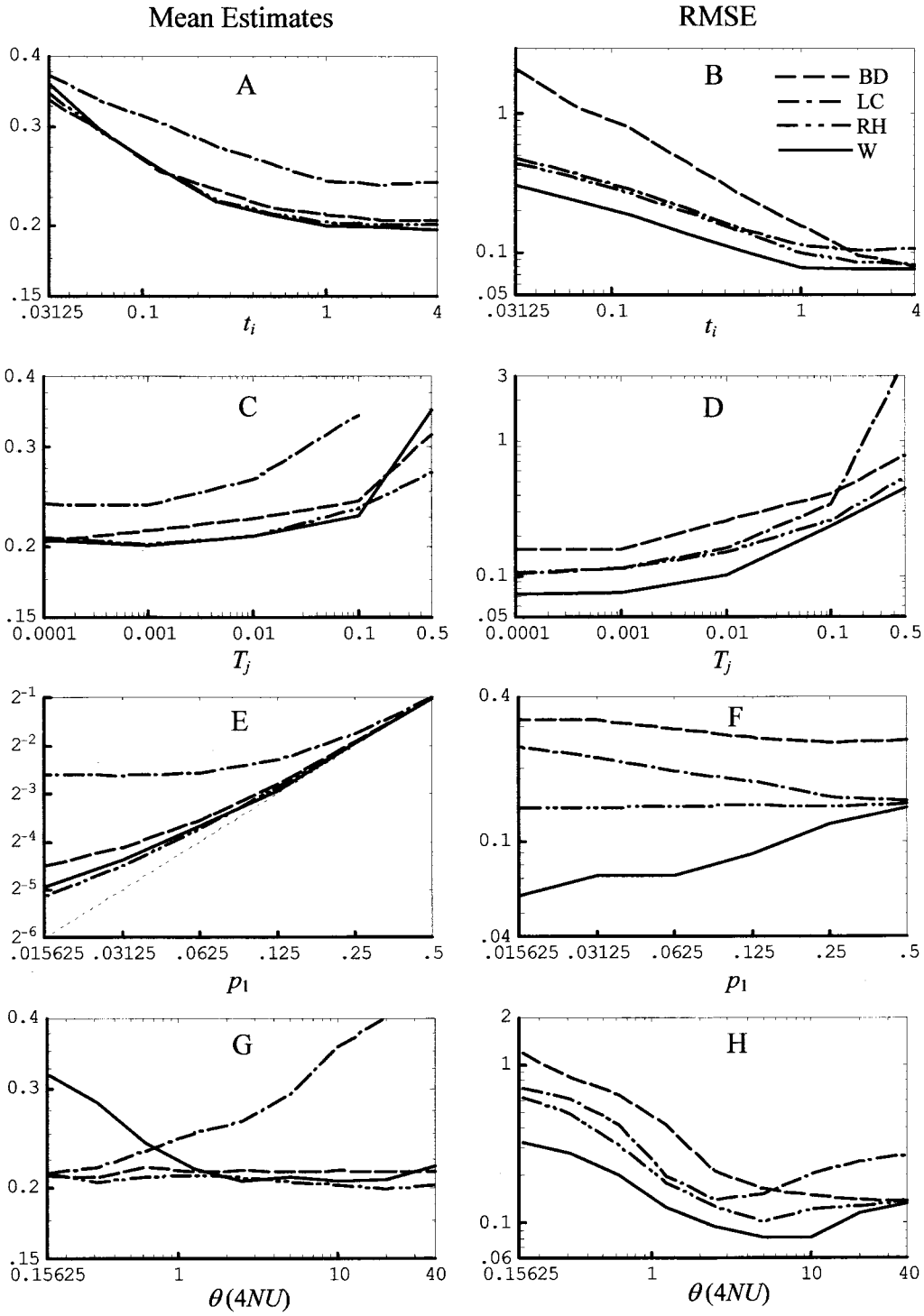

FIGURE 2.—The effects of population properties on the mean and RMSE of admixture proportion estimates ($\hat{p}_1$) from different estimators (indicated by different lines shown at the top right corner of the graph). Both *x*- and *y*-axes are in logarithm scale. All the simulation results are obtained using parameters $n_i = N_j = 5000$ ($i = 0, 1, 2$; $j = 1, 2, h$), $S_j = 50$, and $L = 1$ for a DNA sequence. (A and B) The effects of divergence time (t_i, $t_1 = t_2$) of parental populations, with results obtained from 5000 (for $t_i < 0.1$), 2000 (for $0.1 \le t_i < 1$), or 1000 (for $t_i \ge 1$) replicates using various values of ξ and parameters $\psi = 100$, $p_1 = 0.2$, and $U = 0.0001$. (C and D) The effects of genetic drift after the admixture event (T_j, $T_1 = T_2 = T_h$) on the mean and RMSE of $\hat{p}_1$, with results obtained from 500 (for $T_j < 0.01$), 1000 (for $0.01 \le T_j < 0.1$), or 5000 (for $T_j \ge 0.1$) replicates using various values of ψ and parameters $\xi = 5000$, $p_1 = 0.2$, and $U = 0.0001$. (E and F) The effects of the true admixture proportion (p_1) on the mean and RMSE of $\hat{p}_1$, with results obtained from 1000 replicates using various values of p_1 and the parameters $\xi = 5000$, $\psi = 100$, and $U = 0.0001$. (G and H) The effects of the mutation rate ($\theta = 4NU$) of a DNA sequence on the mean and RMSE of $\hat{p}_1$, with results obtained from 5000 (for $\theta < 1$), 500 (for $1 \le \theta < 10$), or 1000 (for $\theta \ge 10$) replicates using various values of θ and the parameters $\xi = 5000$, $\psi = 100$, and $p_1 = 0.2$.

event with an increasing T_j. With a large value of T_j, the admixture effect is obscured by drift and mutation, and the genetic structure of the current populations is shaped mainly by drift and mutation rather than by admixture. It is encouraging, however, that p_1 can be estimated reasonably well even when T_j is as large as 0.1, which means that the admixture event occurred $0.2N$ generations ago in the past. With larger values of T_j, more marker information is required to obtain reliable p_1 estimates. With the same parameters as in Figure 2, C and D, in the case of $T_j = 0.5$ except that 10 loci are used in the estimation, for example, the means (RMSEs) of estimates of p_1 (true value being 0.2) are 0.34 (0.19), 0.52 (0.40), 0.28 (0.20), and 0.31 (0.16) for the BD, LC, RH, and W estimators, respectively. Compared to the results using a single locus [0.32 (0.78), −0.04 (3.77), 0.27 (0.53), and 0.35 (0.45) for the BD, LC, RH, and W estimators, respectively], use

of multiple loci does not reduce bias much but decreases RMSEs dramatically.

Figure 2, E and F, compares the means and RMSEs of $\hat{p}_1$ estimated from different estimators for various true values of p_1. The p_1 estimates are biased toward 0.5, and the magnitude of the bias decreases with the true value of p_1 increasing toward 0.5 for all estimators. The bias of the LC estimator is especially high when p_1 is small. While the RMSE of the BD or RH estimator is almost constant, that of the LC estimator decreases, and that of the W estimator increases with an increasing true value of p_1. Part of the decline of RMSE with the true p_1 value for the LC estimator comes from the decrease in bias. In general, the likelihood method has the least RMSE for all possible values of p_1. The differences among estimators are strikingly large especially when p_1 is small. With $p_1 = 0.015625$, for example, the RMSEs of the likelihood estimates are only 42, 24, and 18% of the RMSEs of the RH, LC, and BD estimators, respectively.

For the case of true p_1 values >0.5, Figure 2, E and F, applies with the true and estimated p_1 values being replaced by 1 − these values. Overall, therefore, the difference in performance (accuracy, RMSE) among the four estimators increases with the true p_1 value deviating from 0.5.

Mutations have more complex effects on admixture proportion estimation. The effect of U (or $\theta = 4NU$) on the mean and RMSE of p_1 estimates from different estimators is shown in Figure 2, G and H. A single DNA sequence with different global mutation rates (U) was used in the estimation. The average number of distinct alleles observed in the samples (50 sequences from each current population) ranges from 3 for $\theta = 0.15625$ to 90 for $\theta = 40$. In the latter case, the vast majority of the alleles are observed in a single sample only; very few alleles are shared among two or more samples.

When the true value of θ is very small, the likelihood estimates of p_1 are biased toward 0.5, while the estimates from other estimators are almost unbiased. This is because likelihood estimates of p_1 are constrained, by nature, to the proper range [0, 1], while those from the other estimators can be either smaller than zero or larger than one. With $p_1 = 0.2$ and a very small mutation rate such that a single sequence has little information about the admixture event, estimates of p_1 from any estimator are extremely variable, and a considerable proportion of them from the BD, LC, or RH estimator can be negative. When $\theta = 0.15625$, for example, the proportions of negative estimates of p_1 are 15, 11, and 22%, and the smallest estimates are −26, −26, and −10.7, from the BD, LC, and RH estimators, respectively.

While the bias for the likelihood method decreases with θ and can be removed by using multiple loci (sequences; see Figure 3, A and B) even for small θ, that for the LC method increases rapidly with θ and remains high even if multiple loci are used. The bias for the LC method is caused by rare alleles, because at a given mutation rate, the bias decreases rapidly with increasing sample size (see Figure 3, C and D).

The changes of RMSEs of the three frequency-based estimators with θ are complicated. The RMSEs first decrease with increasing true θ values, but when θ is sufficiently large RMSEs begin increasing with θ. This is because while mutations that occurred before the admixture event increase the marker information for p_1 estimation, those that occurred after the admixture event obscure the event. With all frequency-based methods available for estimating admixture, the effects of mutations on the change in allele frequency are ignored. In contrast, the molecular estimator (Bertorelle and Excoffier 1998) takes mutations before and after the admixture event into account and therefore its RMSE decreases monotonically with θ. When θ is very large, therefore, BD has a performance similar to the best frequency-based method.

At the same global mutation rate, a microsatellite locus under the stepwise mutation model is less informative than a DNA sequence for estimating admixture proportions. This is true for both molecular and frequency-based estimators. Simulation results (not shown) indicate that the molecular estimator has in general a much larger RMSE than frequency-based estimators have when a few microsatellite loci are used. Only when scores of microsatellite loci are used in the estimation does the molecular estimator give a similar or slightly better performance than that of frequency-based estimators. These results are in agreement with Bertorelle and Excoffier (1998).

The effects of samples: The effect of the number of marker loci (L) assayed in the samples on the performance of different estimators for p_1 estimation is shown in Figure 3, A and B. Note that the mutation rate used here is $\theta = 0.3125$, resulting in four distinct alleles per locus on average for the set of parameters used in these simulations. The bias of the likelihood estimates for the single-locus case is due to the fact that these estimates are lower bounded by zero and quickly disappears with increasing L. For all four estimators, RMSEs decrease almost linearly (in log scale) with L. However, the RMSE of the less powerful method decreases faster with L and therefore all methods tend to have a similar RMSE when many loci are used.

Figure 3, C and D, shows the effect of sample size (S, number of alleles sampled from each population) on the means and RMSEs of $\hat{p}_1$ from different estimators. The bias for the LC method decreases with S presumably because the probability of rare alleles decreases with increasing sample size. Increasing sample size can reduce RMSEs for all estimators, but with diminishing returns. For a given total value of LS, therefore, in practice it is generally more rewarding to genotype more loci than to sample more genes.

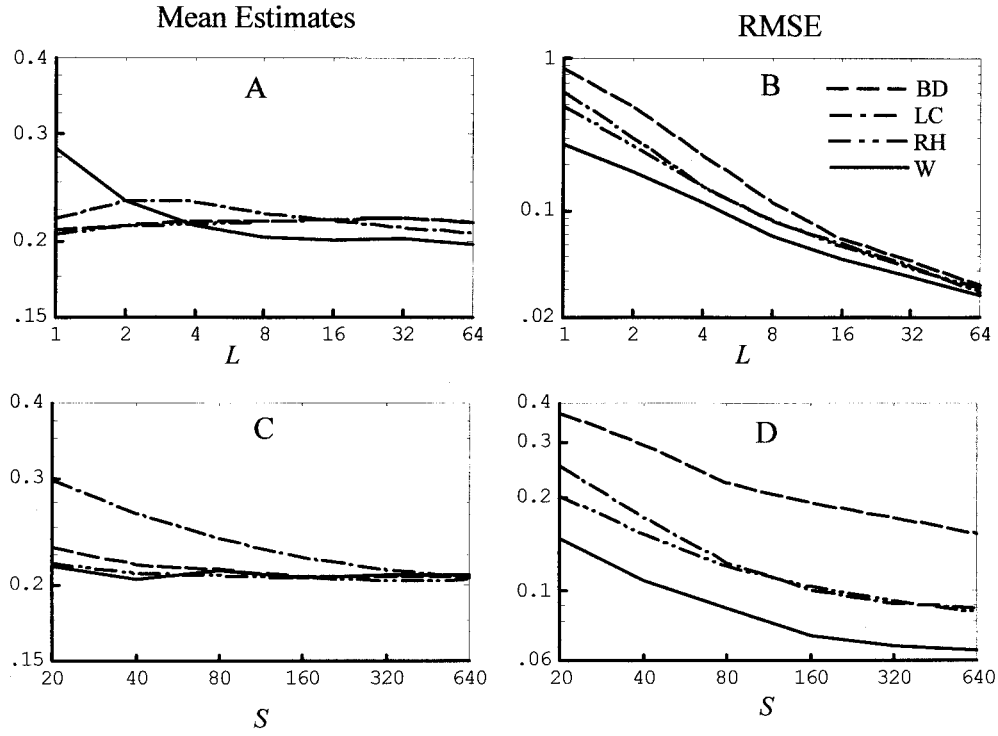

FIGURE 3.—The effects of sample properties on the mean and RMSE of admixture proportion estimates (p_1) from different estimators. Both *x*- and *y*-axes are in logarithm scale. All the simulation results are obtained using parameters $n_i = N_j = 5000$ ($i = 0, 1, 2$; $j = 1, 2, h$), $\xi = 5000$, $\psi = 100$, $p_1 = 0.2$, and DNA sequences. (A and B) The effect of the number of loci (L), with simulation results obtained from 5000 (for $L < 4$), 2000 (for $4 \leq L < 16$), or 500 (for $L \geq 16$) replicates, using various values of L and the parameters $S_j = 50$ and $\theta = 0.3125$. (C and D) The effect of sample size (S, the number of genes), with simulation results obtained from 5000 (for $S < 40$), 2000 (for $40 \leq S < 160$), or 500 (for $S \geq 160$) replicates, using various values of S and the parameters $U = 0.0001$ and $L = 1$ for a DNA sequence.

Estimating other parameters: The current samples also contain information about the drift that has occurred to the populations between the ancestral population split and admixture events (t_i, $i = 1, 2$) and between the admixture and sampling events (T_j, $j = 1, 2, h$). Regarding t_i and T_j as time (in generations) scaled by effective population size, the inverse of them can be viewed as effective population sizes scaled by time. Therefore, if the time ξ or ψ is known from other sources of information, then we can obtain estimates of the average effective sizes of the ancestral or current populations. If the time ξ or ψ is unknown, we can still get an estimate of relative effective sizes. One of the advantages of the likelihood method is that it can estimate these parameters jointly with p_1. The admixture time or population sizes are also important in understanding the admixture events.

Table 1 lists the means and RMSEs of maximum-likelihood estimates of t_i and T_j and LC estimates of T_h. These summary statistics are calculated using estimates of the parameters relative to their true values used in simulations. In general, both t_i and T_j can be estimated by the likelihood method, at least to the correct order for the range of the true parameter values covering several orders. Compared with the admixture proportion, however, t_i and T_j are difficult to estimate. The RMSEs of t_i and T_j estimates are generally of similar magnitudes to the mean estimates. It seems that large sample size and many marker loci are required to obtain reliable estimates of t_i and T_j.

Compared with T_2 and T_h, T_1 is poorly estimated as indicated by larger biases and RMSEs in general. This is understood because $p_1 < 0.5$, and therefore there is less information in the samples about T_1 than about T_2 and T_h. In contrast, p_1 seems to have no obvious effect on the estimation of t_1 and t_2. The LC estimator can also give an estimate of T_h, but its performance is rather poor compared with the likelihood estimator, except for the case of strong drift ($T_h = 0.1$).

Simulation results for the extended admixture model: The basic model described in Figure 1 is very stringent and is unlikely to be realistic in several respects in practice. In this section I use simulations to investigate the effects of violations to some of the assumptions made in the basic model on the estimation of admixture proportions. Different estimators are compared in their robustness.

Constant migration: The basic model assumes that the hybrid population is created instantaneously by mixing genes of proportions p_1 and $1 - p_1$ from parental populations 1 and 2, respectively. In reality, a hybrid population could be formed as a result of more or less constant migration from the parental populations over a long time. The LC and RH estimators do not make assumptions about the admixing process and should estimate the cumulative contributions of parental populations to the hybrid population up to the time of sampling. The molecular estimator was derived under the explicit assumption of instantaneous admixture (BERTORELLE and EXCOFFIER 1998), and the likelihood estimator needs to estimate T_j, which is difficult to define without such an assumption. It is therefore not obvious whether the

TABLE 1

Means and RMSEs (in parentheses) of t_i and T_j estimates relative to their true values

		Maximum-likelihood method					LC estimator:
t_i	T_j	$\hat{T}_1$	$\hat{T}_2$	$\hat{T}_h$	$\hat{t}_1$	$\hat{t}_2$	$\hat{T}_h$
1	0.001	6.36 (9.83)	1.73 (3.70)	2.02 (2.98)	0.37 (0.65)	0.36 (0.65)	38.5 (38.7)
	0.01	0.75 (0.97)	0.75 (0.80)	0.82 (0.44)	0.33 (0.68)	0.35 (0.67)	3.57 (2.69)
	0.1	0.55 (4.36)	0.41 (0.70)	0.51 (0.57)	0.36 (0.67)	0.26 (0.76)	0.35 (0.67)
0.1	0.001	8.73 (10.34)	2.57 (3.54)	2.60 (3.44)	0.86 (0.65)	1.10 (0.69)	37.8 (39.2)
0.01		3.03 (5.90)	1.08 (1.94)	1.78 (3.30)	1.29 (1.14)	1.49 (1.49)	34.0 (40.5)
0.001		1.34 (2.94)	1.32 (3.07)	2.60 (4.17)	6.24 (9.10)	4.23 (7.54)	64.8 (162.1)

The parameters used in the simulations are $n_i = N_j = 5000$ ($i = 0, 1, 2$; $j = 1, 2, h$), $U = 0.0001$ for a single DNA sequence ($L = 1$), $p_1 = 0.2$, and $S = 200$. Different true values of t_i ($t_1 = t_2$, first column) and T_j ($T_1 = T_2 = T_h$, second column) are used in conjunction with the above parameters in simulations. For each set of parameters, 500 replicates are run and the mean and RMSE (in parentheses) of the estimates relative to their true values of t_i and T_j from the likelihood and T_h from the LC estimator are listed.

molecular and likelihood estimators are still applicable to estimating admixture proportions if the assumption is violated.

Let us consider the situation that the current hybrid population is formed by an initial admixture event followed by constant migration and hybridization. Initially, a hybrid population is created by admixing genes of proportions m_0 and $1 - m_0$ from parental populations 1 and 2, respectively. Thereafter, it receives genes, at each generation, of proportions of m_1 and m_2 from parental populations 1 and 2, respectively. After ψ generations when samples are drawn, the cumulative proportion of genes in the hybrid population that come from parental population 1 is $p_1 = (1 - m_1 - m_2)^{\psi}(m_0 - m_1/(m_1 + m_2)) + m_1/(m_1 + m_2)$.

Figure 4 depicts the changes in mean square error (MSE) of $\hat{p}_1$ from each estimator with the constant rate (m_2) of migration from parental population 2 to the hybrid population during the period between initial admixture and sampling events. The m_2 and m_1 values in the simulations are chosen (using the above equation for p_1) so that, for fixed values of $m_0 = 0.05$ and $\psi = 100$, the cumulative contribution of population 1 is always $p_1 = 0.2$. With an increasing m_2 (and thus m_1, dotted line in Figure 4), therefore, the admixture is more and more determined by recent migration.

Clearly, all estimators are robust in the face of constant migration. The estimates of p_1 are actually more accurate slightly (data not shown) and have a smaller MSE with larger constant migration rates for each estimator. This is understood because migration after the initial admixture event actually reduces the effects of genetic drift and mutation in estimating p_1. The more recent the migration, the more informative of the current samples about cumulative admixture because of the less obscuring effects of drift and mutation.

The means of likelihood estimates of T_j decrease with increasing m_1 and m_2. This is expected because with increasing constant migration rates, the admixture is increasingly determined by recent migration. The estimated T_h decreases slightly with m_1 and m_2 (Figure 4), while T_1 and T_2 have the same trend but more noise (not shown). When there is migration, therefore, T_j is usually underestimated and should be treated with caution in practice.

Mutation models: While frequency-based methods apply to any kind of markers, the molecular estimator

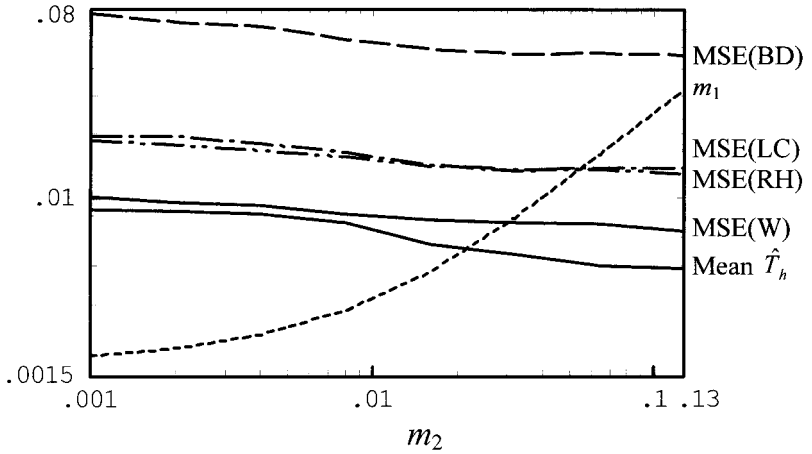

FIGURE 4.—The effects of constant migration rates (m_1, m_2) on the mean square error (MSE) of admixture proportion estimates (p_1) from four estimators, and the mean of estimates of T_h from the likelihood method. Both *x*- and *y*-axes are in logarithm scale. The simulation results are obtained from 1000 replicates, using various values of m_2 and m_1 (indicated by the thin dotted line) so that the cumulative contribution to the hybrid population from parental population 1 is always 0.2 (p_1), when the initial contribution from parental population 1 is set at $m_0 = 0.05$. The parameters used are $n_i = N_j = 5000$ ($i = 0, 1, 2$; $j = 1, 2, h$), $\xi = 5000$, $\psi = 100$, $S_j = 50$, $U = 0.0001$, and $L = 1$ for a DNA sequence.

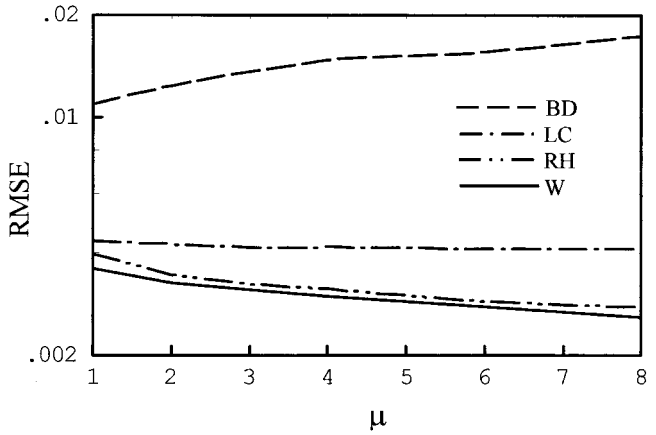

FIGURE 5.—The effects of mutation models on the RMSE of admixture proportion estimates (p_1). Both *x*- and *y*-axes are in logarithm scale. It is assumed that 90% of mutations at a microsatellite locus follow the standard SMM, while the number of repeats (k) being added or deleted in a single mutation follows a Poisson distribution with mean μ truncated for $k < 2$ for the remaining 10% of mutations. The simulation results are obtained from 10,000 replicates, using various values of μ (on the *x*-axis) and the parameters $n_i = N_j = 5000$ ($i = 0, 1, 2$; $j = 1, 2, h$), $\xi = 5000$, $\psi = 100$, $p_1 = 0.2$, $S_j = 50$, $U = 0.00025$, and $L = 10$ for microsatellites.

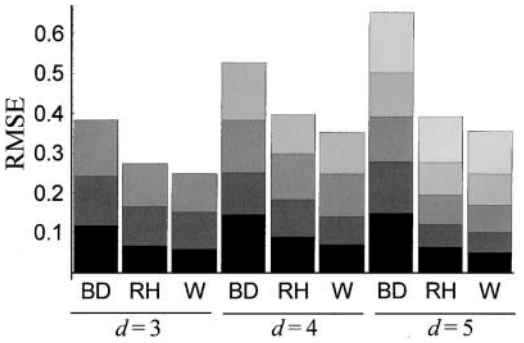

FIGURE 6.—Comparison of RMSE of admixture proportion estimates from different estimators when three or more parental populations are involved in the hybridization. The true values of admixture proportions are $p_1 = 0.1$, $p_2 = 0.2$, $p_3 = 0.7$ for the case of $d = 3$; $p_1 = 0.1$, $p_2 = 0.2$, $p_3 = 0.3$, $p_4 = 0.4$ for the case of $d = 4$; and $p_1 = 0.05$, $p_2 = 0.1$, $p_3 = 0.15$, $p_4 = 0.2$, $p_5 = 0.5$ for the case of $d = 5$, where d is the number of parental populations. For $d = 3$, 4, and 5, results were obtained from 100, 50, and 50 replicates, respectively, all using parameters $n_i = N_j = 5000$ ($i = 0, 1 \sim d$; $j = 1 \sim d$, h), $\xi = 10{,}000$, $\psi = 100$, $S_j = 50$, $U = 0.0001$, and $L = 1$ for a DNA sequence. For each bar, the height of the ith segment (counted from the bottom) corresponds to the RMSE of $\hat{p}_i$.

is suitable only for molecular markers whose coalescent times can be estimated. Although the molecular estimator has taken into account mutations that are ignored in frequency-based methods, it carries with it additional assumptions about mutation models. In particular, it assumes the infinite-sites mutation model for DNA sequences and the SMM for microsatellites (BERTORELLE and EXCOFFIER 1998). In reality, however, there might be mutational "hotspots" for DNA sequences, and multiple tandem repeats could be involved in a single mutational event for microsatellites (SHRIVER *et al.* 1993; DI RIENZO *et al.* 1994). It is therefore important to compare the robustness of different estimators to the deviation of the mutation model from the ideal one assumed.

Here I consider a mutation model for microsatellites in which the number of repeats (k) being added or deleted in a single mutation is one (the standard SMM) for the majority of mutations and follows a Poisson distribution with mean μ truncated for $k < 2$ for the remaining mutations. The addition and deletion of repeats in a mutation are assumed to be equally likely. Figure 5 shows the changes of RMSEs of p_1 estimates as a function of μ when 90% of mutations follow the standard SMM and 10% of mutations have k values drawn from the truncated Poisson distribution with mean μ. Other parameters being fixed, the RMSE of BD increases and those of RH and W decrease with increasing deviation of the mutation model from the standard SMM. This is expected for the molecular method, which relies on the estimation of coalescent time. When mutations do not follow the SMM, the mean coalescent time is no longer proportional to the average squared difference in allele size. However, the BD estimator is surprisingly quite robust to violations of the SMM, with mean p_1 estimates essentially unchanged (data not shown) and RMSE increased only slightly with increasing μ. An increase of μ results in greater polymorphism (number of alleles) at each locus, which leads to the decrease in RMSE for the RH and W estimators. The impact of μ on the LC estimator is a little complicated. A larger μ gives more alleles but also a higher frequency of rare alleles in the samples. Therefore, the standard deviation of p_1 estimates decreases but the bias increases (data not shown) with μ, resulting in RMSE being almost constant.

More than two parental populations: Except for LC, the other three estimators can cope with any number of parental populations contributing to the admixture. When three or more parental populations are involved, the BD estimator assumes that they all contribute at once to the creation of the hybrid population. Similar to the situation of constant migration investigated above, however, this assumption is not necessary for the estimation of admixture proportions. All estimators allow different parental populations to contribute genes to the hybrid population at variable times. Figure 6 compares the RMSEs of admixture proportions estimated from BD, RH, and W estimators when three, four, and five parental populations have contributed to the hybrid. Because of the heavy computation of the likelihood method with increasing d, only a small number of replicates are run and only three initial points are used in searching for the maximum likelihood for each replicate. The performance of the likelihood method shown in Figure 6 can be therefore conservative. It is, however,

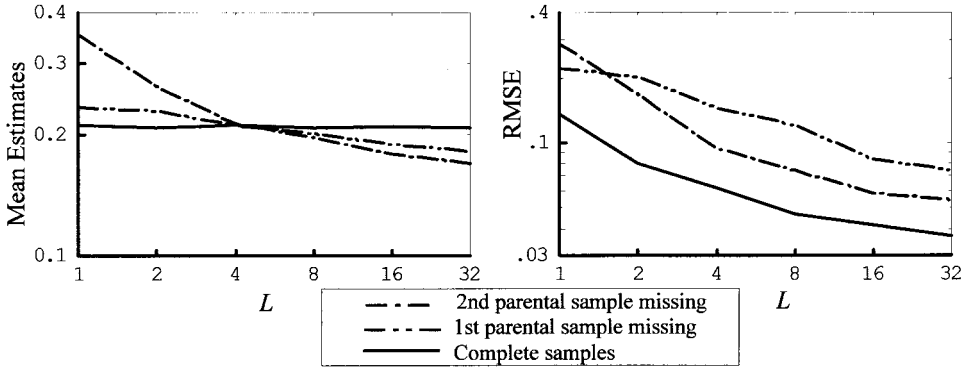

FIGURE 7.—Means and RMSEs of likelihood estimates of p_1 as a function of the number of marker loci (L) used in the estimation. Estimates obtained from complete samples or only two samples (one from a single parental population and the other from the hybrid population) are compared. Both x- and y-axes are in logarithm scale. The results were obtained from 2000 (for $L < 4$) or 500 (for $L \geq 4$) replicates, all using parameters $n_i = N_j = 5000$ ($i = 0, 1, 2$; $j = 1, 2, h$), $\xi = 5000$, $\psi = 100$, $p_1 = 0.2$, $S_j = 200$, and $U = 0.00025$ for microsatellites.

still much better than the other two methods. All methods are essentially unbiased for estimating admixture proportions irrespective of d (data not shown).

Unknown parental populations: Existing methods for estimating admixture from markers rely on correctly identifying all parental populations contributing to a hybrid population and drawing a representative sample from each of them. In practice, however, these conditions are not always met. A parental population may be unidentified (from other sources of information) to the investigator as a contributor, may be known to be extinct, or may no longer exist as a pure breeding population. In all three cases, no sample representative of the parental population is possible. Even if a parental population is known to exist, a sample from it might still be unavailable in some cases. The likelihood method can deal with such a situation of incomplete samples and estimate admixture proportions from samples taken from some of the parental populations.

Figure 7 shows the means and RMSEs of $\hat{p}_1$ estimated by the likelihood method when a sample from one of the two parental populations is unavailable (missing), in comparison with those obtained using complete samples. With limited information (few loci) available from a single contributing parental population, $\hat{p}_1$ is biased toward 0.5, especially when the missing sample is from the parental population that contributed more to the hybrid population. With increasing number of loci (L), however, $\hat{p}_1$ is increasingly biased away from 0.5 slightly. Missing a parental sample also results in decreased precision of $\hat{p}_1$, especially when the missing sample is from the parental population of less contribution to the admixture (data not shown). Overall (in terms of RMSE), a sample from the less-contributed parental population is more crucial in admixture estimation, except when little marker information is available (L is very small in Figure 7). For more than two parental populations, similar results are obtained. In general, admixture proportions can be estimated with acceptable accuracy and precision from incomplete samples by the likelihood method, provided the amount of information (determined by sample size, number of loci, diversity of each locus) is reasonably large from available samples.

Applications: *A data set on African-American populations:* A data set published by PARRA *et al.* (1998) and later amended for a few incorrect genotypes and expanded substantially to include more individuals and an extra locus was analyzed by the newly developed likelihood method and several other methods. The data set contains a sample from each of 11 African-American populations (10 from the United States, 1 from Jamaica), a sample from each of four European populations (Irish, Spanish, British, and German), and a sample from each of six African populations (two from Nigeria, one from the Central African Republic, and three from Sierra Leone). Each sample was genotyped for 10 nuclear loci, of which 9 were biallelic and 1 triallelic. These loci were selected because they are African or European population specific or highly differentiated between the two continents and are thus especially informative for admixture analysis. Assuming a dihybrid model (European/African, referred to as parental population 1/2 in the following analysis) and pooling the European samples and African samples as those from parental populations, the admixture proportions for each of the 11 African-American populations can be estimated.

The European ancestral contributions to each of the 11 African-American populations estimated by BD, LC, RH, and W estimators are listed in Table 2. Different estimators give almost identical point estimates. This is not surprising because the marker frequencies are highly differentiated between parental populations and are thus extremely informative about admixture. Simulations show that different estimators are increasingly discrepant with a decreasing amount of marker information. The level of European contribution varies greatly among these African-American populations, from ~7% in Jamaica to 23% in New Orleans. These results are also similar to previous reports (PARRA *et al.* 1998; MCKEIGUE *et al.* 2000).

The new likelihood method could also obtain information about genetic drift occurring in different popu-

TABLE 2

European ancestral proportions of 11 African-American populations estimated from different methods

Admixed population	Estimation methods			
	BD	LC	RH	W
Maywood, Illinois	18.5 (15.1, 21.6)	18.7 (15.4, 21.6)	18.2 (14.9, 21.2)	19.0 (15.8, 22.3)
Detroit	17.0 (12.6, 26.5)	15.9 (12.6, 22.9)	18.0 (12.8, 26.9)	16.1 (10.8, 22.4)
New York	21.2 (18.0, 23.6)	20.7 (17.7, 23.0)	21.3 (18.1, 23.6)	21.2 (18.1, 24.2)
Philadelphia 1	13.8 (11.2, 15.9)	13.5 (11.0, 15.5)	13.8 (11.2, 15.9)	14.0 (11.6, 16.4)
Philadelphia 2	14.7 (10.8, 17.8)	14.3 (10.0, 17.7)	14.5 (10.6, 17.4)	15.0 (11.3, 19.0)
Pittsburgh	21.0 (19.0, 23.6)	20.4 (17.9, 23.0)	21.0 (18.4, 24.8)	20.8 (17.2, 24.5)
Baltimore	15.2 (11.8, 20.2)	14.6 (11.2, 19.8)	16.0 (12.2, 20.7)	15.0 (10.5, 20.1)
Charleston, South Carolina	12.2 (10.1, 15.7)	11.1 (9.3, 14.0)	12.7 (10.5, 15.6)	11.4 (8.4, 14.5)
New Orleans	22.8 (19.7, 27.6)	22.2 (19.1, 26.4)	23.4 (20.1, 27.5)	21.9 (18.3, 26.2)
Houston	16.4 (12.6, 18.4)	16.3 (13.2, 18.2)	16.1 (12.3, 18.5)	16.6 (13.6, 20.0)
Jamaica	7.5 (3.2, 12.4)	6.0 (1.7, 10.1)	7.9 (3.5, 12.8)	6.5 (3.4, 10.5)

Estimates of admixture proportions are in percentages, and their 95% confidence intervals (%) are in parentheses. The 95% confidence intervals from moment estimators (BD, LC, and RH) were obtained from 1000 bootstrapping samples (over loci), and those from the likelihood estimator were obtained from profile log-likelihood curves.

lations from the data. The MLEs and 95% confidence intervals of t_i and T_j for each of the 11 African-American populations are listed in Table 3. Since we use the same parental population samples together with each admixed population sample in the 11 analyses, t_i are expected to be the same across analyses, whose point estimates turn out to be 0.46 and 0.35 for European and African ancestral populations, respectively (see Figure 8 for the relative profile log-likelihood curves). The direct interpretation is that in the period between the split of African and European populations and the formation of African-American populations, the average effective size of the African population is ~31% larger than that of the European population. However, a more appropriate explanation is difficult. First, t_1 for the European or t_2 for the African population is not clearly defined herein. For the European population, for example, only four countries are represented in the sample and there could be migration between these and other European countries. Therefore, t_1 is vague as to referring to the whole European population or only the part in the four sampled countries. Second, if there is migration between African and European populations after their split, then it is even more difficult to interpret t_i.

Simulations indicate that it is much more difficult to estimate drift than admixture proportions. Large sample sizes as well as many markers are required for estimating T_j with reasonable confidence. For this data set,

TABLE 3

Maximum-likelihood estimates of genetic drift of 11 African-American populations and their parental (European and African) populations

Admixed population	T_1	T_2	T_h
Maywood, Illinois	0.000517 (—, 0.014700)	— (—, 0.005570)	— (—, 0.006215)
Detroit	0.000180 (—, 0.014453)	0.001456 (—, 0.006817)	0.002877 (—, 0.027837)
New York	0.000494 (—, 0.014369)	0.000001 (—, 0.006300)	0.001161 (—, 0.007580)
Philadelphia 1	— (—, 0.013739)	— (—, 0.005004)	— (—, 0.004710)
Philadelphia 2	0.000736 (—, 0.015064)	0.001300 (—, 0.006545)	0.002153 (—, 0.015191)
Pittsburgh	— (—, 0.013738)	— (—, 0.005519)	— (—, 0.005651)
Baltimore	0.000370 (—, 0.014644)	0.001417 (—, 0.006783)	0.004639 (—, 0.023769)
Charleston, South Carolina	0.000240 (—, 0.014399)	— (—, 0.004867)	— (—, 0.005267)
New Orleans	0.000002 (—, 0.014128)	0.001163 (—, 0.006213)	— (—, 0.009843)
Houston	0.000002 (—, 0.013776)	— (—, 0.005499)	— (—, 0.005708)
Jamaica	0.000433 (—, 0.014728)	0.001569 (—, 0.006665)	0.001913 (—, 0.018244)

Subscripts 1, 2, and h represent the parental European and African populations and the admixed African-American populations, respectively. The estimates (95% confidence intervals) of t_1 and t_2 are 0.46 (0.16, 0.93) and 0.35 (0.11, 0.76), respectively, for any of the 11 admixed populations. — indicates that the value is <0.000001. The estimated 95% confidence intervals are in parentheses.

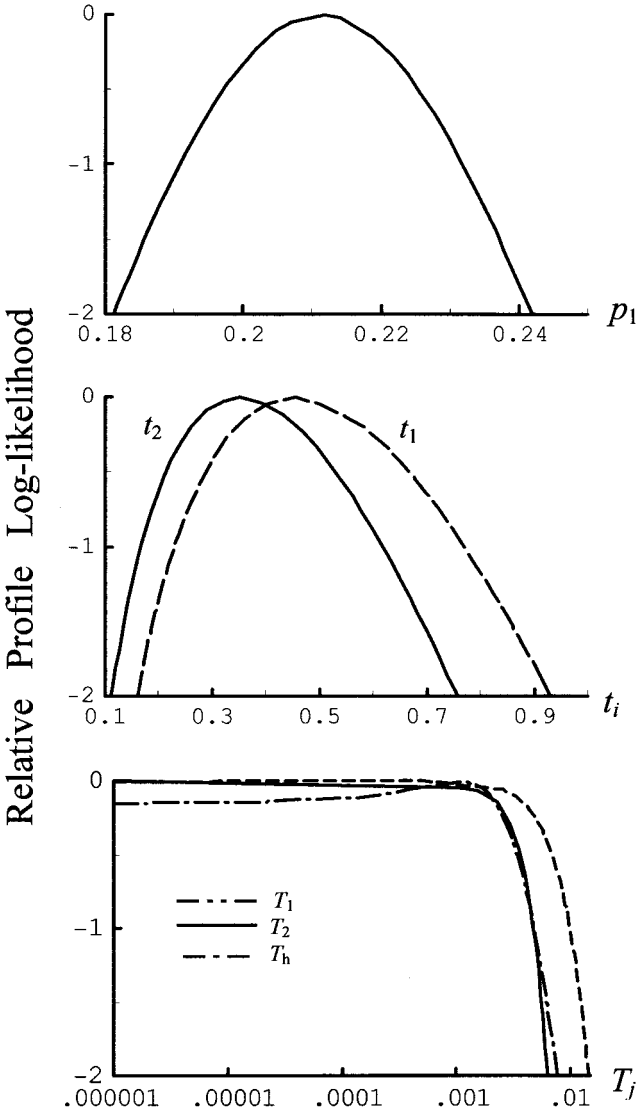

FIGURE 8.—Relative profile log-likelihood *vs.* p_1, t_i, and T_j obtained from the analysis on the data from the admixed African-American population in New York. The European genetic contribution to the admixed population is p_1; the genetic drift that occurred to the European and African populations between population split and admixture events is denoted by t_1 and t_2, respectively; and the genetic drift that occurred to the European, African, and admixed populations between admixture and sampling events is denoted by T_1, T_2, and T_h, respectively.

the parental populations could be very large and thus the drift in them very weak. Evidence suggests that the American-African populations were formed 150 years ago (PARRA *et al.* 1998). Since then, the average effective population size of either the European or the African population is at least in the order of millions. Therefore, the effect of drift could be swamped by that of sampling, because the sample size is only ~84–236 individuals. Despite the inaccuracy, the estimates of drift listed in Table 3 still give some interesting information. First, the drift in the European population is estimated to be much worse than that in the other populations, as indicated by much higher +95% confidence limits. This is consistent with simulation results (Table 1), because the European population contributed less than one-quarter to the admixed populations. Second, the estimated drift in the African population is in general weaker than that in the American-African populations ($T_2 < T_h$). This is intuitively plausible because the admixed populations must be much smaller than the parental populations. Third, overall the estimates of drift (T_1, T_2, and T_h) have the highest precision from the analysis of the New York African population, which happens to have the largest sample size. For this analysis, Figure 8 shows the relative profile log-likelihood curves as a function of p_1, t_i, and T_j. As can be seen, the log-likelihood changes very slowly when T_j becomes small. Taking the generation interval as ~30 years, then 150 years correspond to five generations. The point estimates of drift from the analysis of the New York African population then translate to the effective sizes of the African, European, and New York African populations during the past 150 years being ~2,500,000, 1,063,738, and 2153, respectively. Of course these estimates could be quite inaccurate. Continuous integration over time of African and European genes into the admixed populations would bias the estimates of drift.

A data set on North American wolflike canids populations: ROY *et al.* (1994) analyzed 10 microsatellite loci in seven gray wolf populations and six coyote populations in North America. Hybridization between the two species was identified in two gray wolf populations and two coyote populations. The data set was analyzed by BERTORELLE and EXCOFFIER (1998) for admixture proportions. Here I apply my likelihood method to the same data for estimating admixture and genetic drift jointly. Following BERTORELLE and EXCOFFIER (1998), I pool the five nonhybridized gray wolf samples and the four nonhybridized coyote samples to act as the first and second parental population samples and pool the two hybridized gray wolf samples and the two hybridized coyote samples as the gray-wolf-like and coyote-like admixed samples, respectively. Then I apply different methods to estimate the genetic contribution of gray wolf (p_1) to each of the two admixed populations.

The genetic contributions of gray wolf to the hybrid populations estimated from different methods are compared in Table 4. In contrast to the human data set (Table 2), the admixture estimates are quite different between methods, and the confidence intervals obtained from these methods are very broad. This is expected because the markers used in these wolflike canids populations are neither species specific nor highly differentiated between the two species. The point estimates from the three moment estimators are close to the results of BERTORELLE and EXCOFFIER (1998), but the 95% confidence intervals are generally broader because of the difference in bootstrapping (over loci and over

TABLE 4

Estimates of genetic contribution (%) of gray wolf populations to hybrid populations

	Estimation methods			
Admixed population	BD	LC	RH	W
Gray wolf hybrid	47.8 (22.0, 77.2)	65.2 (15.7, 89.8)	50.9 (31.6, 75.1)	36.4 (14.5, 57.7)
Coyote hybrid	17.9 (−77.7, 53.3)	12.2 (2.0, 19.8)	10.5 (−2.0, 21.8)	4.1 (−15.9)

Estimates of admixture proportions are in percentages, and their 95% confidence intervals (%) are in parentheses. The 95% confidence intervals from moment estimators (BD, LC, and RH) were obtained from 1000 bootstrapping samples (over loci), and those from the likelihood estimator were obtained from profile log-likelihood curves.

alleles for this and the previous study, respectively). These admixture proportion estimates are roughly in agreement with Roy *et al.* (1994), who showed, by both the multidimensional scaling method and Nei's (1978) unbiased genetic distance, that the coyote hybrid populations were indistinguishable from the pure coyote populations and the gray wolf hybrid populations were distinct from both parental populations.

The likelihood estimates of genetic drift are listed in Table 5. In contrast to the human data set (Table 3), the drift in hybrid and parental populations during the period between hybridization and sampling (T_j) is much better estimated. Most of the estimates of T_j have narrow 95% confidence intervals. This is presumably because these wolflike populations may have much smaller effective sizes or/and a larger divergence time (ψ) than that of the human populations. The analyses on both gray wolf and coyote hybrid populations consistently show that the drift in the gray wolf parental population (T_1) is much smaller than that in the coyote parental population (T_2). Taking generation intervals (G) as ~3 and 2 years for gray wolves (Mech and Seal 1987) and coyotes (Nowak 1991), respectively, the results indicate an average N_e of the gray wolf population being about three to nine times larger than that of the coyote population. Further, if the period (ψG) between hybridization and sampling is 90 years (Roy *et al.* 1994; Vila *et al.* 1999), then we obtain from Table 5 that the average effective population sizes of gray wolves and coyotes (during that period) in North America are ~3350–23,440 and 1025–1993, respectively.

The results above are supported by previous studies. Both mtDNA (*e.g.*, Wayne *et al.* 1992) and microsatellite (Roy *et al.* 1994) analyses revealed a much higher level of genetic differentiation among gray wolf populations than among coyote populations in North America. It is hypothesized that the coyote populations in North America probably expanded their range from a much narrower geographic distribution in the past few hundred years (*e.g.*, Voigt and Berg 1987), while gray wolves have existed throughout much of North America for most of the late Pleistocene and have likely survived the most recent ice age in two or more separate refugia (Nowak 1991). The current gray wolf population size is believed to be <60,000 individuals in North America (Carbyn 1987). Assuming $N_e/N = 0.1$ (Frankham 1995), the current N_e would be at most 6000, which is well within the range of my likelihood estimates. The upper limit of the recent average N_e can be slightly larger than this number, considering that the substantial decrease in geographical range and population size occurred to North American wolves in the last few centuries (Mech 1970; Carbyn 1987).

The estimates of genetic drift in hybrid gray wolf and

TABLE 5

Maximum-likelihood estimates of genetic drift of two hybridized wolf-like populations

Population	t_1	t_2	T_1	T_2	T_h
Gray wolf hybrid					
Point estimate	0.10149	0.00098	0.00448	0.02196	0.02911
−95% C.I.	0.06965	0.00065	—	0.01608	0.02263
+95% C.I.	0.13651	0.00102	0.01415	0.02715	0.03639
Coyote hybrid					
Point estimate	0.10704	0.04100	0.00064	0.01129	0.02825
−95% C.I.	0.07443	0.02007	—	0.00485	0.01940
+95% C.I.	0.14886	0.06867	0.01295	0.01744	0.03755

Subscripts 1, 2, and h represent the parental gray wolf and coyote populations and the hybrid populations, respectively. The point estimates and 95% confidence interval limits are listed for t_1, t_2, T_1, T_2, and T_h for each admixed population. — indicates that the value is <0.00001.

coyote populations are generally larger than those in parental populations, suggesting a smaller N_e of hybrid *vs.* parental populations.

Although the gray wolf population has a larger N_e than the coyote population in North America has had in the recent past (the previous 100 years), coyotes could be much more abundant in the more remote past as indicated by a much smaller estimate of t_2 than of t_1 (Table 5). Assuming the period (ξG) between the split of gray wolf and coyote lineages and the admixture event in North America as ~1 million years (Nowak 1979; Kurtén and Anderson 1980) and the period (ψG) between the admixture event and sampling as 90 years (Nowak 1979), then a comparison between t_i and T_i ($i = 1, 2$) in Table 5 indicates that both coyote and gray wolf populations in North America have contracted dramatically recently. However, these results about genetic drift should be treated with caution, because some of the estimates are not good enough. Vilà *et al.* (1999) estimated, from surveys of mtDNA diversity from worldwide samples, that the historical N_e's of the global coyote and gray wolf populations are in millions. They also found dramatic decreases in genetic diversity in both species, possibly because of the substantial decrease in geographical range and population size of North American wolves during the last few centuries (Mech 1970; Carbyn 1987) and the decrease in coyote numbers since the last glacial maximum ~18,000 years ago (Nowak 1979).

DISCUSSION

On the basis of the admixture model proposed by Bertorelle and Excoffier (1998), I developed a maximum-likelihood method that takes into account the genetic differentiation between parental populations and the effects of genetic drift and sampling in each population. The method can be used to estimate admixture proportions and the genetic drift that occurred to different populations over different periods of time simultaneously. Compared with previous moment estimators, the likelihood method is more powerful and has higher precision and accuracy for admixture estimation over various parameter combinations, as verified by extensive simulations. This is not surprising because the likelihood method utilizes most of the information available in the data and takes all relevant factors (except for mutations) into consideration. The molecular estimator developed by Bertorelle and Excoffier (1998) is novel in several respects compared with all frequency-based methods. It allows for mutations and considers not only the frequency differences but also the levels of divergence of different alleles. However, it relies on the estimated average coalescence times within and between populations, which are known to suffer from large variances due to the stochasticity of the genealogical process (Tajima 1983). Extensive simulations in this and a previous study (Bertorelle and Excoffier 1998) show that this estimator has in general the largest RMSE, except for the case of high differentiation between parental populations and very high mutation rates for markers. Essentially, the molecular estimator requires strong mutation against weak drift to outperform frequency-based methods. A limitation to the available likelihood methods, including the current one, is the ignorance of mutations and molecular divergence among alleles. It would be rewarding to explore the molecular approach further in the future, perhaps using a coalescence-based likelihood method.

In contrast to previous likelihood methods (*e.g.*, Thompson 1973; Chikhi *et al.* 2001), the present one considers explicitly the extent of genetic differentiation between parental populations. This is important because falsely assuming independent parental population allele frequency distributions would bias the estimation of admixture proportions, as is exemplified by simulations, especially when marker information is scarce and the parental populations are not completely differentiated. For the allele frequency of the ancestral population (P_0) before population split, w, I assumed a uniform prior to avoid introducing new nuisance parameters. When we do have some information about w, however, we can use a more appropriate prior to reflect our knowledge. If we assume that P_0 is at drift and mutation equilibrium, then w would be in β- or Dirichlet distributions for bi- or multiallelic loci (Wright 1951). The exact distribution depends on parameters such as the effective size of P_0 (n_0) and the mutation rate (u). The uniform prior is actually a special case of β-distribution with $2un_0 = 1$. One may argue that a "U-shaped" distribution ($2un_0 < 1$) might be more plausible for neutral markers under drift and mutation balance. However, the number of nuisance parameters introduced by the β prior is L if L loci (of different mutation rates) are used in estimation, which quickly makes the likelihood computation unmanageable. To simplify the model, one may have to assume a constant $2un_0$ (the same prior) across loci, whose impact on admixture estimation needs to be evaluated. On the other hand, even though a uniform prior is known to be inappropriate, it does not necessarily lead to serious estimation errors provided the amount of data available is not very small. The current extensive simulation results show that the likelihood method using a uniform prior recovers the true parameter values over wide ranges. However, it would be interesting to explore further how much gain we can get from using a more appropriate prior.

The current likelihood method is also computationally simple, making its use in extensive simulations to systematically investigate the performance and statistical properties (such as this study) possible. The computational ease also enables it to be extended readily to more complicated models, including many parental populations, two or more temporal samples taken from a single population to allow better estimates of admixture proportions, and genetic drift. It takes a PIII PC

~20–24 and 12–19 hr to complete the whole likelihood analysis of the human data and the canids data, respectively.

Another advantage of the present likelihood method is its flexibility. For example, it can use dominant markers separately or in conjunction with codominant ones. It can also use cytoplasmic (*e.g.*, mtDNA), sex chromosomal as well as nuclear autosomal markers. However, caution should be exercised to combine information from different kinds of markers, because the admixture might be sex specific. In that case, the admixture proportion is expected to be different for different kinds of markers. A comparison of the estimates obtained from autosomal and cytoplasmic (or sex chromosomal) markers may, however, provide interesting insights into the admixture process (Bertorelle and Excoffier 1998). Assuming a constant ratio of effective population sizes for different kinds of markers [say, N_e (autosomal) = $2N_e$ (cytoplasmic) = $4/3N_e$ (sex linked)], it is straightforward to extend the present likelihood method to analyze data jointly on two or more kinds of markers for separate estimates of sex-specific admixture proportions. The present likelihood method also allows for missing data. Simulations showed that even if no data are available from a single parental population, the method can still yield satisfactory estimates from the incomplete samples. Although some moment estimators (see Chakraborty *et al.* 1992) work as well using a single parental and an admixed sample, they require population-unique markers (those present in one parental population only).

Simulations show that the admixture estimators are surprisingly robust to violations of several assumptions made in the admixture models. A particular concern has been about the assumption that admixture is completed within a short period of time compared with the divergence time (*e.g.*, Bertorelle and Excoffier 1998; Chikhi *et al.* 2001; Dupanloup and Bertorelle 2001). In reality, however, admixture might occur as a more or less continuous gene flow process, lasting from the initial hybridization event to the time point when samples are taken. Simulations indicate that all estimators (including molecular and likelihood) are actually estimating the cumulative contributions from parental to admixed populations and give better estimates in the gene flow model because of the smaller effect of mutation and genetic drift. Gene flow does affect the estimation of drift, as expected, from the likelihood method. The underestimation of T_j is, however, slight, at least in the cases investigated.

In addition to admixture proportions, the likelihood estimators can also estimate genetic drift that occurred to each population. We can therefore obtain estimates of the relative effective sizes of different populations and, when extra information about time is available, the absolute average N_e of each population. Simulations show, however, that drift is much more difficult to estimate than are admixture proportions. Although t_i and T_j are in general correctly estimated (at least to the right order), the precision is low. This means that in practice a large sample size and number of markers are necessary to obtain drift estimates with reasonable confidence. The analyses on both human and wolflike canids data yield encouraging and plausible estimates of genetic drift, which are supported in general by other studies.

A software package, Likelihood Estimation of ADMIXture (LEADMIX), implementing the likelihood method described in this article, is available for free download from http://www.zoo.cam.ac.uk/ioz/software.htm.

I thank Mark Beaumont, Giorgio Bertorelle, Lounès Chikhi, Bill Hill, and two anonymous referees for critical reading and constructive comments on earlier versions of this manuscript.

LITERATURE CITED

Azzalini, A., 1996 *Statistical Inference, Based on the Likelihood*. Chapman & Hall, London.

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

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

Carbyn, L. N., 1987 Gray wolf and red wolf, pp. 358–377 in *Wild Furbearer Management and Conservation in North America*, edited by M. Nowak, J. A. Baker, M. E. Obbard and B. Malloch. Ministry of Natural Resources, Toronto, ON, Canada.

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

Chakraborty, R., and K. M. Weiss, 1986 Frequencies of complex diseases in hybrid populations. Am. J. Phys. Anthropol. **70:** 489–503.

Chakraborty, R., and K. M. Weiss, 1988 Admixture as a toll 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 R. E. Ferrell, 1992 Caucasian genes in American blacks: new data. Am. J. Hum. Genet. **50:** 145–155.

Chikhi, L., M. W. Bruford and M. A. Beaumont, 2001 Estimation of admixture proportions: a likelihood-based approach using Markov chain Monte Carlo. Genetics **158:** 1347–1362.

Crow, J. F., and M. Kimura, 1970 *An Introduction to Population Genetics Theory*. Harper & Row, New York.

Di Rienzo, A., A. C. Peterson, J. C. Garza, A. M. Valdes, M. Slatkin *et al.*, 1994 Mutational processes of simple sequence repeat loci in human populations. Proc. Natl. Acad. Sci. USA **91:** 3166–3170.

Dupanloup, I., and G. Bertorelle, 2001 Inferring admixture proportions from molecular data: extension to any number of parental populations. Mol. Biol. Evol. **18:** 672–675.

Elston, R. C., 1971 The estimation of admixture in racial hybrids. Ann. Hum. Genet. **35:** 9–17.

Frankham, R., 1995 Effective population-size adult-population size ratios in wildlife—a review. Genet. Res. **66:** 95–107.

Glass, B., and C. C. Li, 1953 The dynamics of racial intermixture: an analysis based on the American Negro. Am. J. Hum. Genet. **5:** 1–19.

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, New York.

Kurtén, B., and E. Anderson, 1980 *Pleistocene Mammals of North America*. Columbia University Press, New York.

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. Shiver, 2000 Estimation of admixture and detection of linkage in ad-

mixed populations by a Bayesian approach: application to African-American populations. Ann. Hum. Genet. **64:** 171–186.

Mech, L. D., 1970 *The Wolf: The Ecology and Behavior of an Endangered Species*. University of Minnesota Press, Minneapolis.

Mech, L. D., and U. S. Seal, 1987 Premature reproductive activity in wild wolves. J. Mammal. **68:** 871–873.

Metropolis, N., A. Rosenbluth, M. Rosenbluth, A. Teller and E. Teller, 1953 Equations of state calculations by fast computing machines. J. Chem. Phys. **21:** 1087–1091.

Nei, M., 1978 Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics **89:** 583–590.

Nowak, R. M., 1979 *North American Quaternary Canis*. Museum of Natural History, University of Kansas, Lawrence, KS.

Nowak, R. M., 1991 *Walker's Mammals of the World*, Vol. II , Ed. 5. Johns Hopkins University Press, Baltimore.

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.

Press, W. H., S. A. Teukolsky, W. T. Vetterling and B. P. Flannery, 1996 *Numerical Recipes in Fortran 77*, Ed. 2. Cambridge University Press, Cambridge, UK.

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

Roy, M. S., E. Geffen, D. Smith, E. A. Ostrander and R. K. Wayne, 1994 Patterns of differentiation and hybridization in North American wolflike canids, revealed by analysis of microsatellite loci. Mol. Biol. Evol. **11:** 553–570.

Shriver, M. D., L. Jin, R. Chakraborty and E. Boerwinkle, 1993 VNTR allele frequency distribution under the stepwise mutation model: a computer simulation approach. Genetics **134:** 983–993.

Tajima, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics **105:** 437–460.

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

Vilà, C., I. R. Amorim, J. A. Leonard, D. Posada, J. Castroviejo *et al.*, 1999 Mitochondrial DNA phylogeography and population history of the gray wolf Canis lupus. Mol. Ecol. **8:** 2089–2103.

Voigt, D. R., and W. E. Berg, 1987 Coyote, pp. 345–356 in *Wild Furbearer Management and Conservation in North America*, edited by M. Nowak, J. A. Baker, M. E. Obbard and B. Malloch. Ministry of Natural Resources, Toronto, ON, Canada.

Wang, J., 2001 A pseudo-likelihood method for estimating effective population size from temporally spaced samples. Genet. Res. **78:** 243–257.

Wayne, R. K., N. Lehman, M. W. Allard and R. L. Honeycutt, 1992 Mitochondrial DNA variability of the gray wolf—genetic consequences of population decline and habitat fragmentation. Conserv. Biol. **6:** 559–569.

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

Communicating editor: J. B. Walsh