Bayesian Inference of Recent Migration Rates Using Multilocus Genotypes

Gregory A. Wilson and Bruce Rannala[1]

Department of Medical Genetics, University of Alberta, Edmonton, Alberta T6G 2H7, Canada

Manuscript received May 22, 2002
Accepted for publication December 10, 2002

ABSTRACT

A new Bayesian method that uses individual multilocus genotypes to estimate rates of recent immigration (over the last several generations) among populations is presented. The method also estimates the posterior probability distributions of individual immigrant ancestries, population allele frequencies, population inbreeding coefficients, and other parameters of potential interest. The method is implemented in a computer program that relies on Markov chain Monte Carlo techniques to carry out the estimation of posterior probabilities. The program can be used with allozyme, microsatellite, RFLP, SNP, and other kinds of genotype data. We relax several assumptions of early methods for detecting recent immigrants, using genotype data; most significantly, we allow genotype frequencies to deviate from Hardy-Weinberg equilibrium proportions within populations. The program is demonstrated by applying it to two recently published microsatellite data sets for populations of the plant species *Centaurea corymbosa* and the gray wolf species *Canis lupus*. A computer simulation study suggests that the program can provide highly accurate estimates of migration rates and individual migrant ancestries, given sufficient genetic differentiation among populations and sufficient numbers of marker loci.

In recent decades, indirect estimates of gene flow (reviewed in Slatkin and Barton 1989) have been widely used by biologists, first with allozyme data and more recently with restriction fragment length polymorphisms (RFLPs), DNA sequence data, microsatellite markers, and single-nucleotide polymorphisms (SNPs). Direct estimates of migration rates based on mark-recapture or other methods can be impractical for large populations that exchange small numbers of migrants because the expected number of recaptures is too low; indirect estimates of gene flow using genetic markers are often the only recourse. Commonly used indirect estimators of gene flow, such as $4N_em = 1/F_{ST} - 1$, are derived on the basis of simplified models of population structure that assume constant population sizes, symmetrical migration (at constant rates), and population persistence for periods sufficient to achieve genetic equilibrium (Wright 1931, 1969).

The development of coalescent theory (Kingman 1982; reviewed by Tavare 1984), which traces the ancestral genealogy of a sample rather than modeling changes of gene frequencies in the population as a whole, has allowed less restrictive models to be used in developing indirect estimators of gene flow. The new methods accommodate recent population expansions, nonsymmetrical migration, and other complexities that are typical of real biological populations (Beerli and Felsenstein 1999, 2001; Vitalis and Couvet 2001). However, even coalescent-based methods currently assume that population demography has followed a relatively simple model of either constant size or deterministic expansion (with constant migration rates) for roughly the last $4N_e$ generations, which is the average time until the sampled chromosomes coalesce to a most recent common ancestor (Kingman 1982). For populations with large N_e or species in highly disturbed habitats, this assumption may be unreasonable.

Recently, nonequilibrium approaches have been proposed for identifying migrants (Rannala and Mountain 1997; Pritchard *et al.* 2000) or hybrids between species (Anderson and Thompson 2002) and assigning individuals of unknown population affinity to potential source populations using multilocus genotypes (Paetkau *et al.* 1995; Rannala and Mountain 1997; Cornuet *et al.* 1999; Dawson and Belkhir 2001; Gaggiotti *et al.* 2002). These methods extract information about recent migration (within the last few generations) from transient disequilibrium observed at individual multilocus genotypes of migrants or individuals recently descended from migrants. In comparison with indirect estimators of long-term gene flow, these methods make relatively few assumptions, but are informative only about recent patterns of migration. The two approaches (long-term gene flow and recent migration estimation) are complementary, providing information about migration on different timescales. Previous methods for inferring recent migration have focused on identifying individual migrants and their source populations (Paetkau *et al.* 1995; Rannala and Mountain 1997) or jointly identi-

[1]*Corresponding author:* Department of Medical Genetics, University of Alberta, 8-39 Medical Sciences Bldg., Edmonton, AB T6G 2H7, Canada. E-mail: brannala@ualberta.ca

fying migrants and populations (Pritchard *et al.* 2000). Existing methods do not explicitly estimate migration rates among populations.

In this article, we develop a new Bayesian multilocus genotyping method for estimating rates of recent migration among populations. The method requires fewer assumptions than estimators of long-term gene flow and can be legitimately applied to nonstationary populations that are far from genetic equilibrium. Moreover, the newly proposed method relaxes a key assumption of previous nonequilibrium methods for assigning individuals to populations and identifying migrants—namely that genotypes are in Hardy-Weinberg equilibrium within populations. We allow arbitrary genotype frequency distributions within populations by incorporating a separate inbreeding coefficient for each population. The joint probability distribution of inbreeding coefficients is estimated from the data. Our method also allows for missing genotype data by using data augmentation techniques to integrate over possible genotypes for individuals.

THEORY

Data and model parameters: Consider a collection of I populations of a diploid species, with discrete nonoverlapping generations, and let $\mathbf{m} = \{m_{lq}\}$ be the migration rates between populations, where m_{lq} is the fraction of individuals in population q that are migrants from population l ($\mathbf{m}$ can also be treated as time dependent). Assume that some proportion of an individual's alleles originate via a single migrant ancestor that arrived at the current (or a past) generation (this is justified for low migration rates, see appendix a). The individual itself may also be a migrant, in which case 100% of its genome is of migrant origin. Define $\mathbf{M} = \{M_h\}$, where M_h is the source of migrant ancestry for individual h, and $\mathbf{t} = \{t_h\}$, where t_h is the generation at which a migrant ancestor of individual h arrived (*e.g.*, if $t_h = 0$ the individual has no migrant ancestry, if $t_h = 1$ the individual is itself a migrant, etc.). $\mathbf{M}$ and $\mathbf{t}$ are then unobserved variables describing the ancestry of each individual. To allow population genotype frequencies to deviate from Hardy-Weinberg equilibrium we define $\mathbf{F} = \{F_l\}$, where F_l is the inbreeding coefficient for population l and $-1 \le F_l \le 1$. Let $\mathbf{p} = \{p_{ijl}\}$ be the population frequencies of marker alleles, where p_{ijl} is the frequency of allele i at locus j in population l.

Let $\mathbf{X} = \{X_{hj}\}$ be the multilocus genotypes observed at J marker loci in a random sample of n diploid individuals, where X_{hj} is the genotype of individual h at locus j, and let $\mathbf{S} = \{S_h\}$ identify the population source for each sampled individual, where S_h is the population that individual h was sampled from. The number of individuals sampled from the lth population is n_l. The data (observations) are $\mathbf{X}$ and $\mathbf{S}$. The joint (and marginal) posterior probability distributions of the remaining parameters $\mathbf{M}$, $\mathbf{t}$, $\mathbf{p}$, $\mathbf{m}$, and $\mathbf{F}$ are estimated numerically using Markov chain Monte Carlo (MCMC) methods (Gamerman 1997). The estimated posterior probabilities are used to make inferences about these parameters (including point estimates). The elements of $\mathbf{m}$ are of primary interest, but other parameters, such as $\mathbf{M}$ and $\mathbf{t}$, may also be of interest (as in Rannala and Mountain 1997) and can be estimated similarly.

Likelihood: The likelihood of the data is the probability of the observed genotypes given the model parameters. This is

$$\Pr(\mathbf{X}|\mathbf{S};\, \mathbf{M},\, \mathbf{t},\, \mathbf{F},\, \mathbf{p}) = \prod_{h=1}^{n} \prod_{j=1}^{J} \Pr(X_{hj}|S_h;\, M_h,\, t_h,\, \mathbf{F},\, \mathbf{p}), \qquad (1)$$

where

$$\Pr(X_{hj}|S_h;\, M_h,\, t_h,\, \mathbf{F},\, \mathbf{p}) = \begin{cases} \Phi(X_{hj},\, g) & \text{if } M_h = S_h = g \text{ and } t_h = 0, \\ 0 & \text{if } M_h \neq S_h = g \text{ and } t_h = 0, \\ \Phi(X_{hj},\, r) & \text{if } M_h = r,\ S_h = g, \text{ and } t_h = 1, \\ (1 - \frac{1}{2}^{t_h-2})\Phi(X_{hj},\, g) + (\frac{1}{2}^{t_h-2})\varphi(X_{hj},\, r,\, g) & \\ & \text{if } S_h = g,\ M_h = r, \text{ and } t_h > 1, \end{cases}$$

and

$$\Phi(X_{hj},\, r) = \begin{cases} (1 - F_r)p_{ijr}^2 + F_r p_{ijr} & \text{if } X_{hj}(1) = X_{hj}(2) = i, \\ 2(1 - F_r)p_{ijr}p_{hjr} & \text{if } X_{hj}(1) = i \text{ and } X_{hj}(2) = k \text{ for } i \neq k, \end{cases}$$

and

$$\varphi(X_{hj},\, r,\, g) = \begin{cases} p_{ijr}p_{ijg} & \text{if } X_{hj}(1) = X_{hj}(2) = i, \\ p_{ijr}p_{hjg} + p_{hjr}p_{ijg} & \text{if } X_{hj}(1) = i \text{ and } X_{hj}(2) = k \\ & \quad \text{or } X_{hj}(2) = i \text{ and } X_{hj}(1) = k \\ & \quad \text{for } i \neq k, \end{cases}$$

where $X_{hj}(1)$ denotes the allele present on the maternal chromosome, and $X_{hj}(2)$ denotes the allele present on the paternal chromosome. Note that we define $t_h = 0$ if $M_h = S_h$ (*i.e.*, if the individual has no immigrant ancestry). The likelihood presented in Equation 1 involves a product of individual genotype probabilities across marker loci and individuals because it is assumed that individuals are randomly sampled and the markers are unlinked.

Prior distributions of parameters: To calculate the probability of observing $\mathbf{M}$ and $\mathbf{t}$, given $\mathbf{m}$, we assume that the populations are large enough that there is negligible genetic drift over two, or three, generations (for a justification, see appendix a). The expected proportion of migrants from population l that arrive in the present generation (the generation at which sampling is carried out) is then m_{lq} and the expected proportion of individuals with one migrant ancestor from the previous generation of migration is $2m_{lq}$ (see appendix a). We use only first- and second-generation migrants to estimate m_{lq} in this article, but more distant migrant ancestries could also be used. The probability distribution of $\mathbf{M}$ and $\mathbf{t}$, given $\mathbf{m}$, follows a multinomial distribution,

$$\Pr(\mathbf{M},\, \mathbf{t}|\mathbf{m}) = \prod_{l=1}^{I} n_l! \left(\prod_{t=1}^{2} \prod_{q \neq l}^{I} \left(\frac{[2^{t-1}m_{lq}]^{n_{lqt}}}{n_{lqt}!} \right) \right) \times \prod_{l=1}^{I} \left(\frac{m_{ll}^{n_{ll0}}}{n_{ll0}!} \right), \qquad (2)$$

where

$$m_{ll} = 1 - \sum_{t=1}^{2} \sum_{q \neq l} 2^{t-1} m_{lq},$$

and

$$n_{lqt} = \sum_{h=1}^{n} \Im(M_h, t_h, S_h),$$

and

$$\Im(M_h, t_h, S_h) = \begin{cases} 1 & \text{if } M_h = l, S_h = q, \text{ and } t_h = t, \\ 0 & \text{otherwise.} \end{cases}$$

We use uninformative (uniform) Dirichlet prior densities for **m** and **p** subject to the constraints

$$\sum_{i=1}^{k_{lj}} p_{ijl} = 1, \quad \text{for all } j = 1, 2, \ldots, J \text{ and } l = 1, 2, \ldots, I,$$

where k_{lj} is the total number of alleles at locus j in population l and

$$\sum_{q=1}^{I} m_{ql} = 1, \quad \text{for all } l = 1, 2, \ldots, I.$$

We assume a uniform prior on the interval $(-1, 1)$ for the population inbreeding coefficient of population l, F_l.

Posterior distributions of parameters: The joint posterior probability density of the model parameters, applying Bayes' theorem, is

$$f(\mathbf{m}, \mathbf{M}, \mathbf{t}, \mathbf{F}, \mathbf{p}, |\mathbf{X}, \mathbf{S}) = \frac{\Pr(\mathbf{X}|\mathbf{S}; \mathbf{M}, \mathbf{t}, \mathbf{F}, \mathbf{p}) \times \Pr(\mathbf{M}, \mathbf{t}|\mathbf{m})\, f_{\mathbf{p}}(\mathbf{p})\, f_{\mathbf{m}}(\mathbf{m}) f_{\mathbf{F}}(\mathbf{F})}{\Pr(\mathbf{X}|\mathbf{S})}. \quad (3)$$

The denominator of Equation 3 above involves high-dimensional sums and integrals and it is not practical to evaluate it explicitly for samples of hundreds of individuals. Here, we use MCMC methods to estimate the joint posterior probability density of Equation 3. This requires only that it be possible for the numerator to be evaluated; this can be done using Equations 1 and 2 given above. MCMC can be carried out efficiently, even for large samples. Details of the MCMC algorithm are given in APPENDIX B.

EXAMPLES

Application to data from the plant *Centaurea corymbosa*: The plant species *Centaurea corymbosa* is currently found in only six populations in southern France. In a study by FREVILLE *et al.* (2001), 228 individuals (minimum population sample size of 20) from these six populations were genotyped at six microsatellite loci. This data set provides a useful test for our method, as the genetic differentiation between most populations is large, likely as a result of limited seed and pollen dispersal (FREVILLE *et al.* 2001). While the geographical distances between the populations vary, all occur within a 3-km^2 area. Observed pairwise F_{ST} values between populations ranged between 0.03 and 0.39 (mean F_{ST} = 0.23). An assignment test performed as described in RANNALA and MOUNTAIN (1997) assigned 91.7% of the individuals to their source population and 7.4% to a neighboring population (FREVILLE *et al.* 2001).

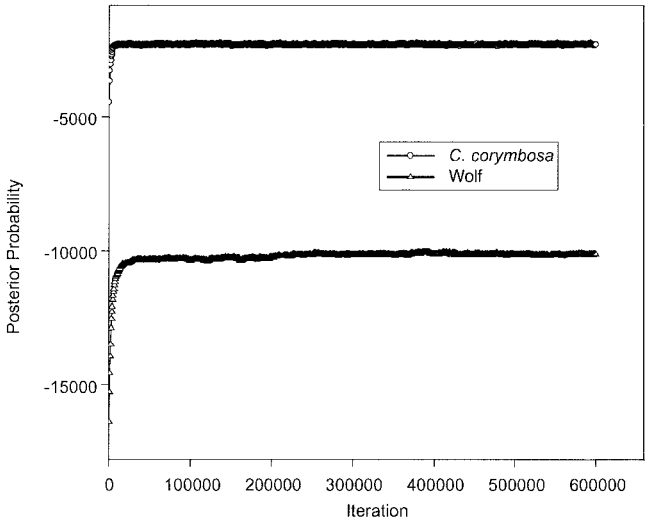

FIGURE 1.—Log-posterior probabilities of the proposed states for the gray wolf and the *C. corymbosa* microsatellite data. Log-posterior probabilities were measured through 600,000 iterations of the MCMC program, sampled every 500 iterations.

To estimate the posterior probability distributions of parameters the MCMC was run for a total of 3×10^6 iterations, discarding the first 10^6 iterations as burn-in (intended to allow the chain to reach stationarity). Samples were collected every 2000 iterations to infer posterior probability distributions of parameters of interest, including the population allele frequencies, migrant proportions, and individual immigrant ancestries. Figure 1 shows the log posterior probability plotted against the iteration number for the *C. corymbosa* data for the first 600,000 iterations. The increase in log probability appears to plateau after only ~500 iterations.

To further examine the convergence of the MCMC algorithm, the posterior probability density of each allele frequency at each locus in each population (grouped in intervals of 0.05) was compared for two independent runs with random initial parameter values, using either 2500 or 3×10^6 iterations. The results are shown in Figure 2, A and B. If the two chains have converged, the relationship between their posterior probabilities should be linear. The high degree of scatter in the plot of 2500 iterations illustrates that the chains have not yet converged (Figure 2A). With 3×10^6 iterations, the relationship is much more linear (Figure 2B). A similar plot of the posterior densities of the inbreeding coefficients in two runs of 3×10^6 iterations also indicates a strong correlation between posterior probabilities estimated from the two independent runs (Figure 3A).

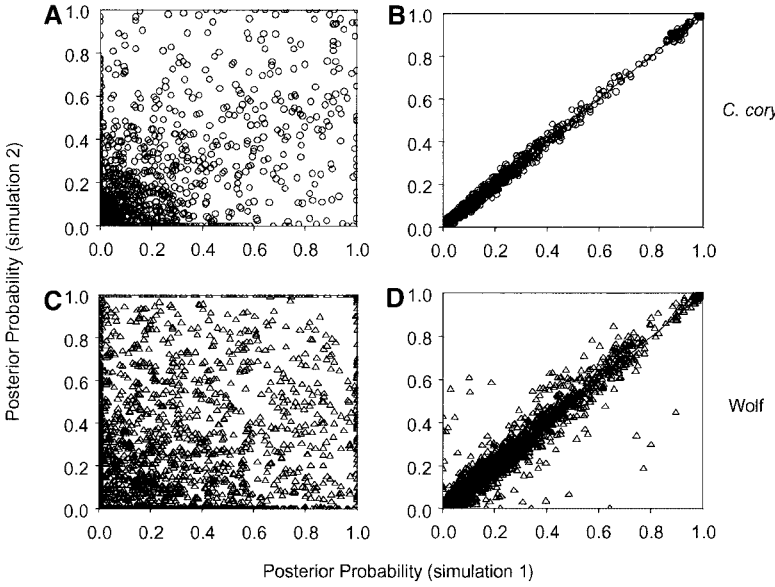

FIGURE 2.—Posterior probability densities of the allele frequencies generated from two separate runs of the program. The runs differed in initial random seed and initial values of m and F. (A and C) The relationship between these runs over the first 2500 iterations, before equilibrium has been reached. (B and D) The relationship between these runs after equilibrium has been reached. The latter runs consist of 3×10^6 iterations, a burn-in of 10^6, and a sampling period of 2000. Allele frequencies are grouped in 0.05 intervals.

The mean posterior probabilities of the immigration rates among populations for the *C. corymbosa* data are shown in Table 1. Most populations have low migrant proportions (when averaged over the posterior probabilities) with the exception of population E1, which appears to have a large expected proportion of migrants ($m = 0.25$) from population E2. There appears to be a source-sink relationship between the two populations because the expected proportion of migrants into population E2 from E1 is much smaller ($m = 0.00$). Figure 4, A and B, presents the posterior densities of the frequencies of two alleles in a population with either a low (Figure 4A) or a high migration rate (Figure 4B). Population sample sizes are nearly identical (38 and 40 individuals, respectively). Both the migration rate and the sample size affect the variance of the posterior probability distribution; higher migration rates and smaller sample sizes both increase the variance. In Figure 4A, the estimated 95% credible set of values for the allele frequency is (0.50, 0.80) while in Figure 4B it is (0.55, 0.95). Migration can also cause the mode of the posterior density of allele frequency to differ from the maximum-likelihood estimate of allele frequency that would be obtained by using the population sample directly and ignoring immigration as is done in many population assignment tests (*e.g.*, PAETKAU *et al.* 1995).

Another property of the populations that can be studied is the posterior probability distribution of the total numbers of nonimmigrants, first-generation immigrants, and second-generation immigrants. Figure 5, A–C, shows

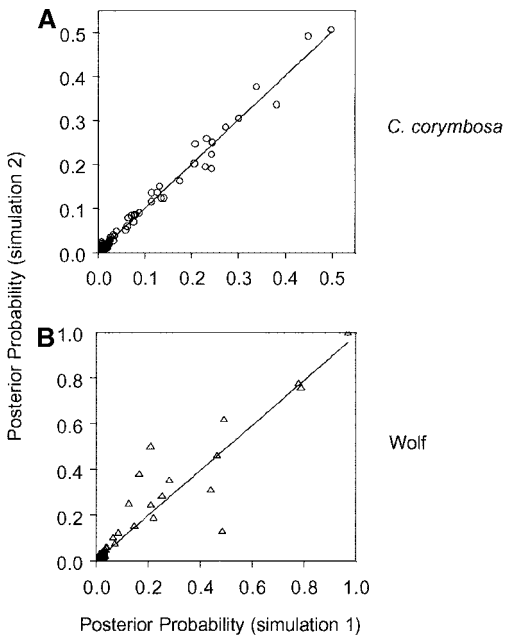

FIGURE 3.—Posterior probability densities of inbreeding coefficients generated from two different runs of the program. Settings are as in Figure 2.

TABLE 1

Migration rates among *C. corymbosa* populations

	E1	E2	A	Pe	Po	Cr
E1	*0.73*	*0.25*	0.00	0.00	0.01	0.00
E2	0.00	*0.99*	0.00	0.00	0.00	0.00
A	0.00	0.00	*0.99*	0.00	0.00	0.00
Pe	0.00	0.00	0.00	*0.99*	0.00	0.00
Po	0.00	0.00	0.00	0.00	*0.99*	0.00
Cr	0.00	0.00	0.00	0.00	0.00	*0.96*

Means of the posterior distributions of **m**, the migration rate into each population, are shown. The populations into which individuals are migrating are listed in the rows, while the origins of the migrants are listed in the columns. Values along the diagonal are the proportions of individuals derived from the source populations each generation. Migration rates ≥0.10 are in italics. Standard deviations for all distributions were <0.05.

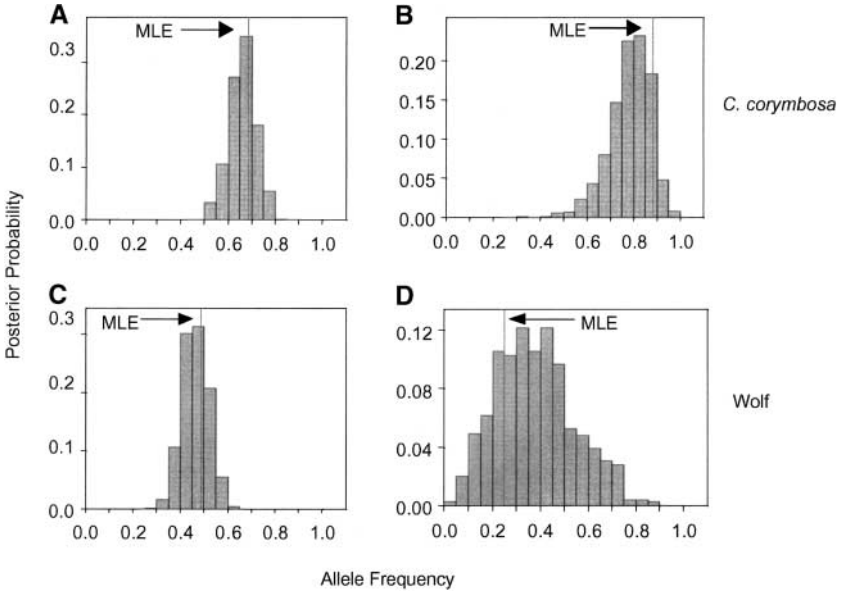

FIGURE 4.—Posterior probability density of a particular allele over all sampled iterations. (A) Allele 174 from locus 13D10 in population Pe. (B) Allele 163 at locus 13B7 in population E1. (C) The frequency distribution of allele 128, locus cxx140, Fort St. John population. (D) The distribution of allele 200, locus cxx204, Great Bear Lake population. The gray line represents the maximum-likelihood estimate for this allele when calculated from individuals sampled from this population. Settings for the MCMC chain are as in Figure 2.

these posterior distributions for *C. corymbosa* population E1. The expected proportions of nonimmigrants and first-generation immigrants overlap, although the variance of the posterior distribution of the proportion of first-generation migrants is lower. The expected proportion of second-generation immigrants is about twice as high and the variance is also larger (this is likely due in part to the fact that assignments of second-generation immigrants are less certain than those of first generation). The 95% credible set for the proportion of first-generation migrants is (0.10, 0.45) *vs.* (0.30, 0.75) for second-generation migrants and (0.00, 0.55) for nonmigrants. The reason that the probability of the proportion of nonimmigrants being above 0.55 is negligible, while the migration rate into this population is ~0.25 (Table 1), is outlined in APPENDIX A. The prior predicts that the expected proportion of first-generation migrants should be m and the proportion of second-generation migrants should be $2m$. As no higher orders of migrants are currently considered in our method, the average proportion of nonmigrants should be $\sim 1 - m - 2m$ under our model, or in this case, 0.25, which falls near the center of our 95% credible set.

Our method can also be used to study the migrant ancestry assignments of individuals, taking account of overall population migration rates and uncertain population allele frequencies. Figure 6A shows the posterior probabilities of nonimmigrant, first-, or second-generation immigrant ancestry for five individuals from population E1 and one individual from population E2. Individual 4-E1 is most likely to be a first-generation immigrant, individuals 11-E1 and 37-E1 are most likely to be second-generation immigrants, and individuals 22-E1 and 31-E1 are roughly equally likely to be either

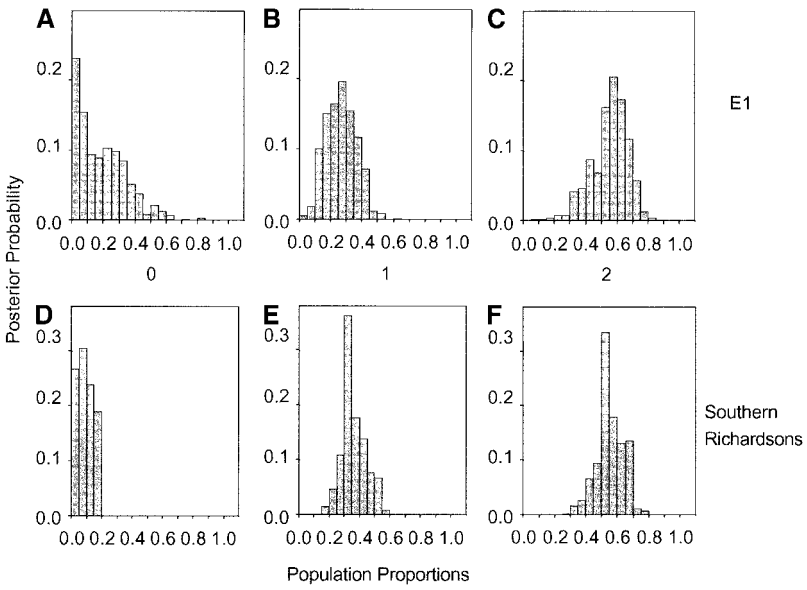

FIGURE 5.—Posterior probability distribution of the proportion of the individuals in a population assigned as nonimmigrants (0), first-generation migrants (1), and second-generation migrants (2) at each sampling iteration. E1 and the Southern Richardsons are populations of *C. corymbosa* and wolf, respectively.

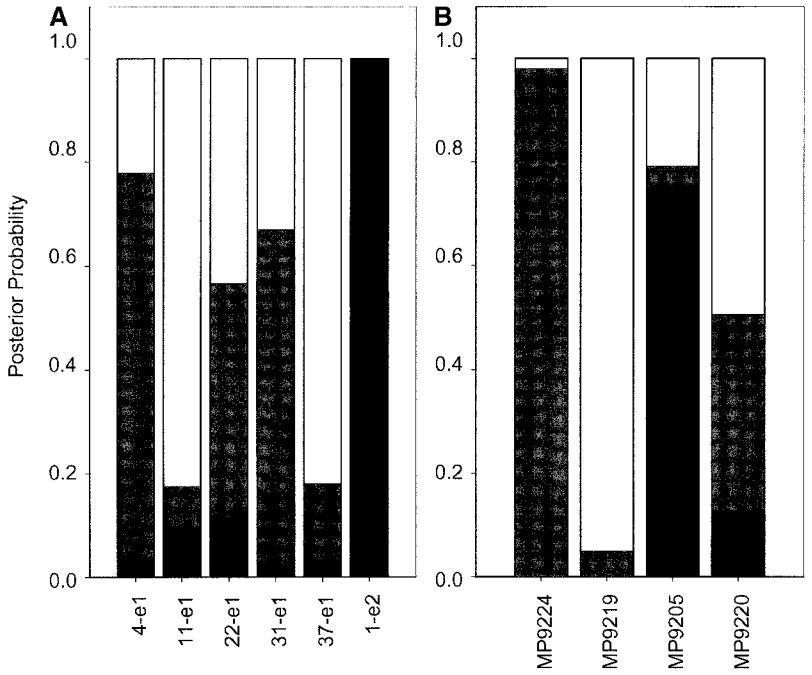

FIGURE 6.—Posterior distribution for the assignment of individuals to ancestry states 0 (■), 1 (▒), and 2 (□) for *C. corymbosa* and wolf. All individuals are from the populations examined in Figure 5, except the last *C. corymbosa* individual, which is from E2.

nonimmigrants or second-generation immigrants. Our method is able to identify second-generation immigrants with a high level of certainty due to the linkage disequilibrium observed in the multilocus genotypes of individuals whose parents have originated in different populations. Individual 1-E2 is most likely to be a nonimmigrant. Excluding population E1, in only 3 of 190 cases did an individual assign with probability >0.05 to a population other than the one it was sampled from, indicating very low levels of migration.

The posterior probability density of the population inbreeding coefficient, F, was concentrated near 0 for most populations, although the standard deviation was large in population E1, which had the greatest amount of immigration; in that case, the estimated mean of the posterior density was $F = 0.027$ but the standard deviation was 0.39. This is likely a result of the lack of information available to the method for estimating F, as most individuals in this population have high posterior probabilities of being first- or second-generation migrants. The remaining populations had much lower standard deviations (<0.08). The population Pe had significant posterior probability associated with relatively large positive values of F (mean of posterior density was $F = 0.123$ with a standard deviation of 0.05), suggesting potential local inbreeding effects.

Application to gray wolf data: In a study of population genetic structure of gray wolves, *Canis lupus*, in the Canadian Northwest, CARMICHAEL *et al.* (2001) genotyped nine microsatellite loci in 491 individuals (minimum sample size of 9 individuals) from nine separate regions. This data set is a valuable test of our method, as the amount of differentiation between populations has a fairly wide range. Some populations are situated fairly close to one another, with no obvious physical barriers to gene flow between them (for example, the Tuktoyaktuk/Inuvik and Paulatuk populations, $F_{ST} = 0.009$), while others are separated by mountain ranges (Kluane National Park), the Arctic Ocean (Banks Island), or large geographic distances (Fort St. John). As such, these samples allow us to determine the effect of differences in genetic differentiation on our method's ability to obtain reliable estimates of migration rates and individual immigrant ancestries.

To estimate the posterior probability distributions of the parameters the MCMC was run for a total of 3×10^6 iterations, discarding the first 10^6 iterations as burn-in. Samples were collected every 2000 iterations to infer posterior probability distributions of parameters. Figure 1B shows the log-posterior probability plotted against iteration number for the gray wolf data. The increase in log-probability appears to plateau after ~10,000 iterations. Figure 2, C and D, shows the correlations (between two independent MCMC runs) of the posterior probability densities of each allele frequency, at each locus, in each population (grouped in intervals of 0.05). The high degree of scatter in the plot of 2500 iterations *vs.* the plot of 3×10^6 iterations (which is highly linear) once again illustrates that the chains have not yet converged at 2500 iterations but have the appearance of convergence after 3×10^6 iterations. A similar plot (Figure 3B) of the posterior densities of the inbreeding coefficients in two runs, each with 3×10^6 iterations, also indicates a strong correlation between posterior probabilities (suggesting the chains have converged).

The means (averaged over posterior probabilities) of the immigration rates between populations for the gray wolf data are shown in Table 2. Four of the populations

TABLE 2

Migration rates among gray wolf populations

	Northern Richardsons	Banks Island	Fort St. John	Great Bear Lake	Kluane National Park	Southern Richardsons	Paulatuk	Tuk/Inuvik	Victoria Island
Northern Richardsons	*0.93*	0.00	0.01	0.00	0.00	0.04	0.00	0.00	0.00
Banks Island	0.00	*0.99*	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Fort St. John	0.00	0.00	*0.99*	0.00	0.00	0.00	0.00	0.00	0.00
Great Bear Lake	0.03	0.01	0.02	*0.68*	0.01	0.09	0.01	*0.14*	0.00
Kluane National Park	0.01	0.00	0.03	0.00	*0.90*	0.04	0.01	0.01	0.00
Southern Richardsons	*0.22*	0.01	0.02	0.01	0.01	*0.71*	0.01	0.02	0.01
Paulatuk	0.00	0.00	0.00	0.00	0.00	0.01	*0.68*	*0.21*	0.00
Tuk/Inuvik	0.03	0.01	0.01	0.00	0.00	*0.23*	0.00	*0.72*	0.00
Victoria Island	0.01	*0.19*	0.02	0.02	0.02	0.02	0.02	0.02	*0.70*

Means of the posterior distributions of **m**, the migration rate into each population, are shown. The populations from which each individual was sampled are listed in the rows, while the populations from which they migrated are listed in the columns. Values along the diagonal are the proportions of individuals derived from the source populations each generation. Migration rates ≥0.10 are in italics. Standard deviations for all distributions were <0.05.

appear quite isolated (Banks Island, Fort St. John, Kluane National Park, and Northern Richardson Mountains). The remaining five populations all have at least one major source of immigrants. There were some notably large mean migration rates between wolf populations. The mean migration rate from the Northern Richardson Mountains to the Southern Richardson Mountains was 0.22; from Tuk/Inuvik to Great Bear Lake, 0.14; from Tuk/Inuvik to Paulatuk, 0.21; and from the Southern Richardson Mountains to Tuk/Inuvik, 0.23. All of these populations are relatively close to one another, occurring on the mainland of the northern Yukon or the Northwest Territories. However, it is worth noting that most of these populations do not have symmetrical migration rates, suggesting that movement of animals between these regions is predominantly unidirectional. For example, while the mean migration rate from the Northern to the Southern Richardson Mountains populations was 0.22, the mean migration rate in the opposite direction was only 0.04. The mean migration rate from Banks Island to Victoria Island was also fairly large at 0.19 while the reverse rate was near zero (see Table 2). These islands are quite close to one another and are joined by ice during the winter months.

Figure 4, C and D, presents the posterior densities of the frequencies of two alleles in populations with either a low immigration rate and a larger sample size or a high immigration rate and a smaller sample size. In these examples, the sample sizes are quite different between the populations (*e.g.*, 41 individuals for Fort St. John and 22 individuals for Great Bear Lake). Immigration causes the mode of the distribution to exceed the maximum-likelihood estimate by a considerable amount (Figure 4D) and the variance of the estimated posterior density of allele frequency is also much larger in the example with a smaller sample size and higher migration rate. In Figure 4C the estimated 95% credible set for the allele frequency is (0.35, 0.60) while in Figure 4D it is (0.10, 0.70).

Figure 5, D–F, shows the posterior probability distributions of the total proportions of nonimmigrants and first- and second-generation immigrants (from any population) for the Southern Richardson Mountains gray wolf population. The mode of the posterior proportion of nonmigrants is much lower than that for the posterior distribution of the proportion of either first- or second-generation migrants. Also, the mode of the posterior distribution of second-generation migrants is roughly twice that of first-generation migrants. The variance of the posterior distributions of first- and second-generation migrant proportions is much greater than that of the nonmigrant proportion. The 95% credible sets for the former are (0.20, 0.50) and (0.40, 0.70), respectively, *vs.* (0.00, 0.20) for the latter.

Figure 6B shows the posterior probabilities of nonimmigrant, first-, or second-generation immigrant ancestry for four individuals from the Southern Richardson

Mountains population. Individual MP9205 is most likely to be a nonimmigrant. Individual MP9224 is most likely to be a first-generation immigrant, individual MP9219 a second-generation immigrant, and individual MP9220 is fairly evenly split between being a first- and second-generation immigrant. The posterior probability density of the population inbreeding coefficient, F, was concentrated near 0 for most populations, with the exception of two populations, Great Bear Lake and Northern Richardson Mountains, which had significant posterior probability associated with negative values of F. F was also approximately uniformly distributed between -1 and $+1$ in the Victoria Island population, likely because most of the individuals in this population were assigned as migrants.

SIMULATION STUDY

Simulation methods: To evaluate the statistical properties of the new method we simulated samples from populations exchanging migrants according to the Wright (1931) island model (at stationarity). The allele frequencies (assuming biallelic loci) in pairs of populations receiving migrants from a common source, with allele frequency q_i at locus i, were simulated from the stationary probability density function (pdf) under the Wright island model. The simulated markers could be SNPs, for example, which are typically biallelic. The pdf of the allele frequency at locus i in population j is

$$f(p_{ij}) = \frac{\Gamma(4Nm)}{\Gamma(4Nmq_i)\Gamma(4Nm[1-q_i])} p_{ij}^{4Nmq_i-1}(1-p_{ij})^{4Nm(1-q_i)-1}. \quad (4)$$

The pdf of the allele frequencies at J unlinked loci in population i is $f(\mathbf{p}_j) = \Pi_i f(p_{ij})$, where the product is over the J loci and $\mathbf{p}_j = \{p_{ij}\}$ is the vector of allele frequencies in population j. The alleles at each locus were therefore simulated as independent and identically distributed with common pdf given by Equation 4. A sample of n individuals was generated from each simulated population according to the multinomial sampling distribution of Equation 2. It was assumed that (recent) migration occurs between the two populations with rates m_{12} and m_{21}. To reduce the number of parameters to be considered in our simulations, we assumed that $m = m_{12} = m_{21}$ and $q_{ij} = q$ for all i, j.

If an individual is a nonmigrant, the genotype is generated by assigning alleles according to the Hardy-Weinberg proportions, conditional on the simulated allele frequencies in the population from which the individual was sampled. A first-generation migrant similarly has its genotype assigned according to Hardy-Weinberg proportions, but conditional on the allele frequencies in the alternative population. A second-generation migrant has its genotype assigned by drawing an allele from each population, respectively, at each locus. To simplify the comparisons, we define the population allele frequencies in terms of F_{ST} by using the standard result for the expected F_{ST} at stationarity under the Wright model, $F_{ST} = 1/(4Nm + 1)$, and solving for $4Nm$ in terms of F_{ST} to obtain $4Nm = 1/F_{ST} - 1$. The right-hand side of this equation was substituted for $4Nm$ in Equation 4. The simulation results are therefore presented in terms of F_{ST}, m, q, and n. To evaluate the statistical performance of the estimator of migration rates under the simulations we focused on two statistics, the mean square error (MSE) and the bias (see Casella and Berger 1990). MSE is a function of both the bias and the variance of the estimator (MSE = bias^2 + variance). A decrease in MSE therefore indicates an improvement in the estimator. To evaluate the statistical accuracy of migrant ancestry assignments we examined the proportion of migrants from each ancestral class (*e.g.*, nonmigrants, first-generation migrants, and second-generation migrants) that were assigned to a given class with maximum posterior probability.

To examine the performance of the model under various conditions, different values were assigned to a number of parameters. The most common allele in a population (q) was assigned a value of either 0.5 or 0.9. The number of individuals sampled from each population (n) was either 20 or 100. Populations were separated by F_{ST} values of 0.01, 0.10, or 0.25. Migration rates between populations (m) were 0.01, 0.05, 0.10, or 0.20. Three different numbers of loci were simulated: 5, 10, and 20. The parameters listed above were used for simulations in all possible combinations, for a total of 144 parameter combinations. Each of these combinations was replicated 10 times. As each simulated data set contained two populations, data were generated for 20 simulated populations for each combination of parameter settings. The MCMC was run with the same settings (number of iterations, etc.) as in each of the examples. As the results with $q = 0.5$ were very similar to those obtained with $q = 0.9$, only the former are examined here.

Simulation results: The results of the simulation study are summarized in Figures 7–10. Figure 7 shows the influence of the number of loci and the migration rate used for the simulations on MSE and bias of the estimated migration rate for a fixed degree of genetic differentiation ($F_{ST} = 0.25$). In the case of 5 loci (Figure 7D), the data have little influence on the estimates, by comparison with the influence of the prior. The prior specifies that m is uniform on the interval (0, 0.33) with mean 0.167. When the actual value of m exceeds the mean of the prior (*e.g.*, when $m = 0.2$), the estimator has a negative bias. When the actual value of m is less than the mean of the prior (*e.g.*, $m \leq 0.1$) the estimator has a positive bias, as expected if the posterior is essentially similar to the prior. With 20 loci, the data have a greater influence than the prior and we see a smaller positive bias for all values of m considered (Figure 7B). In general, MSE decreases with an increase in the num-

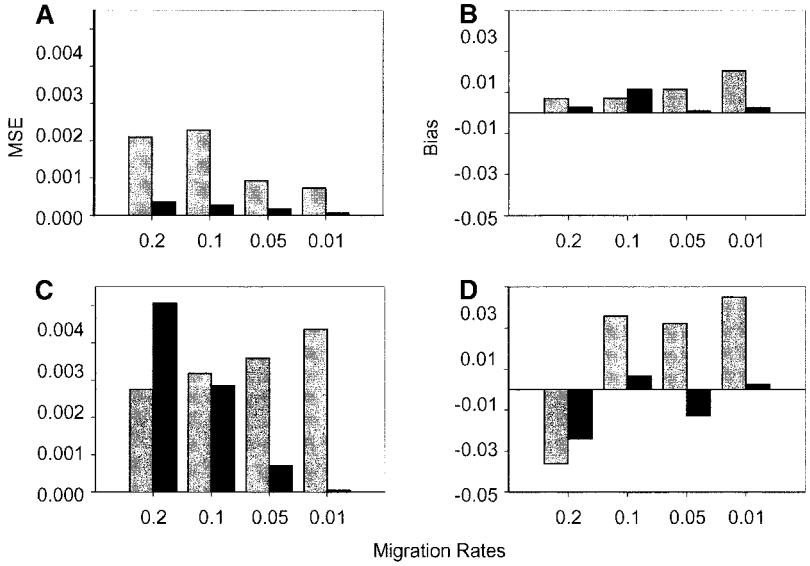

FIGURE 7.—MSE and bias for the migration rate estimate from simulated data. The following parameters were used for data simulation: 5 (C and D) or 20 (A and B) loci, 20 (▒) or 100 (■) individuals per population, and migration rates of 0.2, 0.1, 0.05, or 0.01, when $F_{ST} = 0.25$.

ber of loci sampled (Figure 7, A and C) and with increasing sample size, although sample size appears less important in this case.

It is apparent from our simulation analyses that the effects of sampling either more individuals or more loci are correlated. With a small number of loci, increasing the sample size (from 20 to 100) has little effect on the bias or MSE of the estimated migration rate (Figure 8, A and B), but with a larger number of loci (20 loci), increased sample size dramatically reduces bias and MSE (Figure 8, C and D).

The migration rate and the level of genetic differentiation between populations also influence the mean (and variance) of the maximum posterior probabilities (*i.e.*, the highest posterior probability assignment) of individual migrant ancestries. In the case of a high degree of genetic differentiation between populations ($F_{ST} = 0.25$) and 20 loci, the mean of the maximum posterior probability assignment (across sampled individuals) increases with decreasing migration rate and the variance of the maximum posterior probability (across individuals) decreases (Figure 9, A and B). In the case of low genetic differentiation between populations ($F_{ST} = 0.01$) and 5 loci the migration rate has little influence on the mean or variance of the maximum posterior probability assignments (Figure 9, C and D).

Figure 10 examines the accuracy of the individual migrant ancestry assignments as a function of migration rate, sample size, and number of loci when populations with a high degree of genetic divergence ($F_{ST} = 0.25$) are considered. For each of the categories 0 (nonmigrant), 1 (first-generation migrant), or 2 (second-gener-

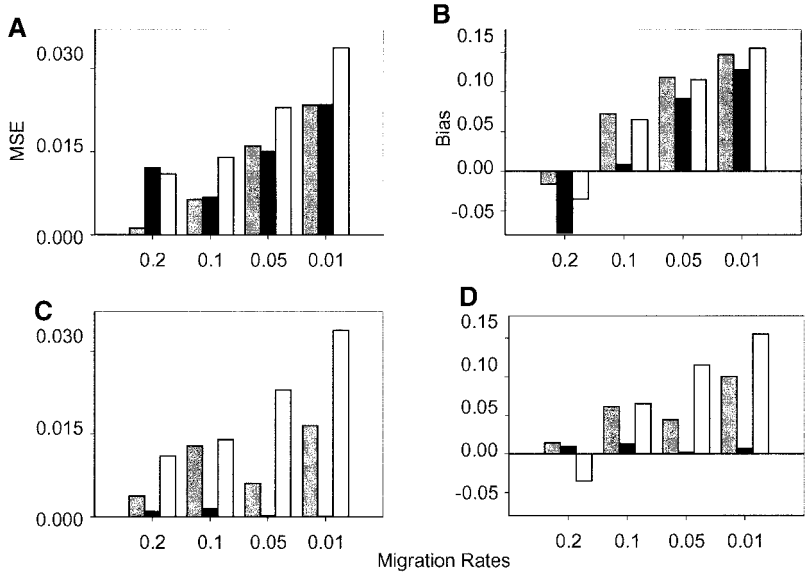

FIGURE 8.—MSE and bias for the migration rate estimate from simulated data. The following parameters were used for data simulation: $F_{ST} = 0.01$ and 5 loci (A and B) or $F_{ST} = 0.10$ and 20 loci (C and D). Simulations were performed with either 20 (▒) or 100 (■) individuals per population and migration rates of 0.2, 0.1, 0.05, or 0.01. MSE and bias for the prior (□) are also given.

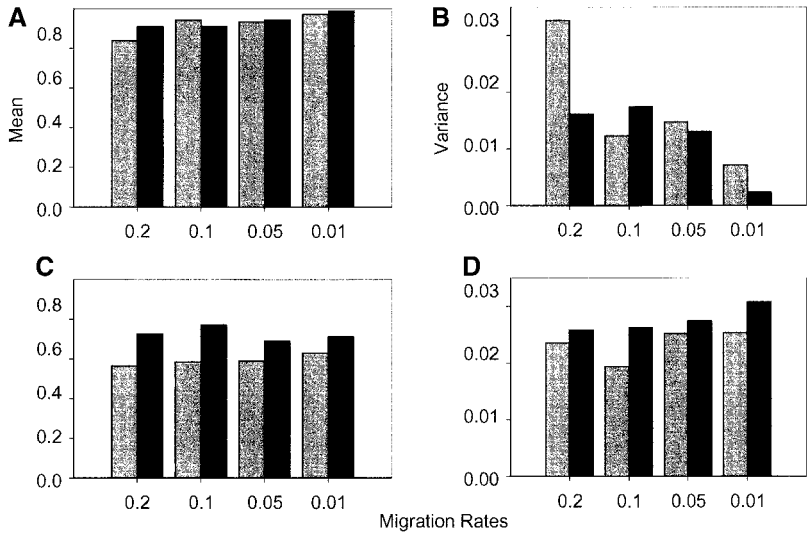

FIGURE 9.—Mean and variance of the maximum posterior probability for each individual migrant ancestry from simulated data. The following parameters were used for data simulation: F_{ST} = 0.25 and 20 loci (A and B) or F_{ST} = 0.01 and 5 loci (C and D). Simulations were performed with either 20 (■) or 100 (■) individuals per population and migration rates of 0.2, 0.1, 0.05, or 0.01.

ation migrant), the total population of individuals actually belonging to that category is represented by the height of the histogram bar. Each histogram bar is then divided into three different shades, representing the proportion of individuals actually belonging to that category that are assigned to each of the three categories. If the assignments were perfectly accurate, each histogram bar would be filled with a single shade (corresponding to the migrant ancestry class represented by that histogram bar).

Of the four cases shown in Figure 10, the cases with either high migration rate (m = 0.2) and large samples of individuals (100) and loci (20) or low migration rate and small samples of individuals (20) and loci (5) provide the most accurate assignments (Figure 10, A and D). Decreasing the number of loci sampled from 20 to 5 has a large effect in decreasing the accuracy of assignments (Figure 10, A and B), but increasing the number of individuals sampled has only a modest effect on accuracy (Figure 10, B and C). Finally, decreasing the migration rate also has a large effect, improving the accuracy of the method even when only 5 loci and 20 individuals are sampled (Figure 10, C and D). At least part of the explanation for this trend is the fact that with lower migration rates population allele frequencies are more accurately estimated (due to the larger proportion of nonmigrants in the sample).

In conclusion, although it is impossible to generalize because of the enormous number of possible parameter combinations that can occur, our simulations suggest that with five or fewer loci and low migration rates very little information is available for inferring migration rates; increasing the number of individuals sampled has a modest effect in improving estimation except in certain cases, such as with low migration and a high degree of genetic differentiation among populations. A higher level of genetic differentiation among populations results in improved accuracy of estimated migration rates and migrant ancestry assignments. Migrant ancestries are most accurate when either a large number of loci and individuals are sampled or migration rates are low.

DISCUSSION

In this article, a new Bayesian method is presented for use with allozyme, microsatellite, RFLP, or SNP multilocus genotype data, which allows one to simultaneously infer recent migration rates, population allele frequencies, population inbreeding coefficients, individual migrant ancestries, and other parameters of potential interest. Our method should be of interest to ecologists assessing the relative importance of specific patterns of population dynamics in nature, the prevalence of male- (or female-) biased dispersal, the importance of geographic barriers to dispersal, and so on.

We have applied our method to two previously published microsatellite data sets for plants (*C. corymbosa*) and mammals (gray wolves) to illustrate its use. We have shown that for each of these data sets reasonably precise information about recent migration patterns can be extracted. In the case of the *C. corymbosa* data, a highly asymmetrical pattern of immigration in one pair of populations (E1 and E2) supports the existence of a source-sink population structure.

Another pattern observed in both example analyses is that a greater proportion of individuals in populations with ongoing migration have more distant migrant ancestry (*e.g.*, second-generation *vs.* first-generation migrant ancestry). This is as expected under the low migration rate approximation presented in APPENDIX A. If migrants beyond the first generation are ignored in an assignment test the result may be biased so that individuals with second-generation migrant ancestry are incorrectly assigned as first-generation migrants. It was also observed that estimated population allele frequencies could deviate considerably from maximum-likeli-

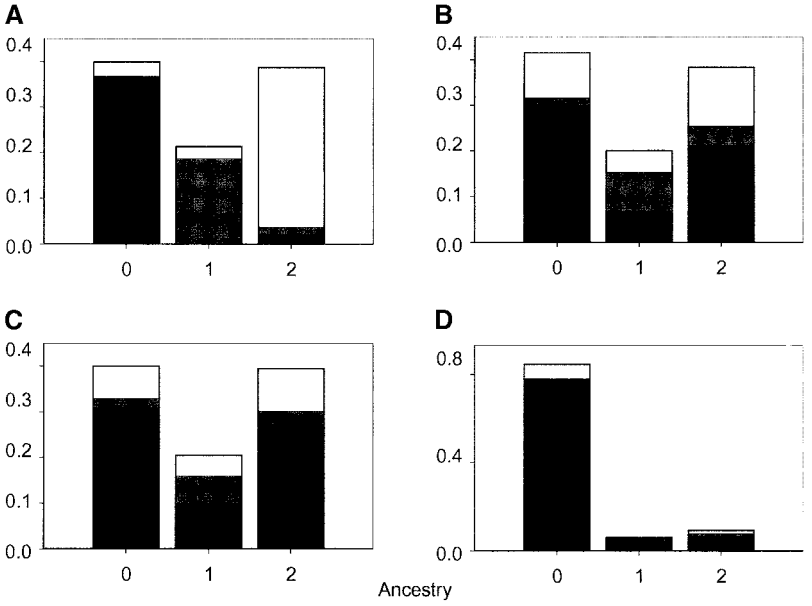

FIGURE 10.—The proportion of individuals with a migrant ancestry of 0 (■), 1 (▒), and 2 (□) (size of the vertical bar) who have their maximum posterior probability in each state (proportion of the bar shaded). Data were simulated from a population F_{ST} of 0.25. Simulations were performed with a migration rate of 0.2 (A–C) or 0.05 (D), 100 (A and B), or 20 (C and D) individuals, and 20 (A) or 5 (B–D) loci.

hood estimates (observed proportions of alleles) in populations experiencing high rates of immigration; it is therefore important to simultaneously estimate individual migrant ancestries and population allele frequencies as we have done in this article. Failing to do so may increase the likelihood that an immigrant individual is incorrectly assigned as a nonimmigrant due to incorrect estimation of the allele frequencies within populations. This study therefore suggests that it may be preferable to estimate migration rates, migrant ancestries, and allele frequencies simultaneously in population assignment tests.

The results of our limited simulation study indicate that very accurate estimates of migration rates and individual migrant ancestries can be obtained when levels of genetic differentiation among populations are large, migration rates are low, and 20 or more loci are examined. If 5 or fewer loci are examined little information may be available, even if a large number of individuals are sampled. To explore the robustness of migration rate estimates and assignments for particular data sets, it may be advisable to carry out preliminary simulations to determine the expected accuracy of the method, given the observed level of genetic differentiation among populations. In our simulation study, we considered only diallelic loci and it is likely that accuracy may increase with increasing numbers of alleles (*e.g.*, with microsatellite loci *vs.* SNPs).

There are a number of ways in which the approach presented here could be extended in the future. First, we have ignored preexisting patterns of genetic differentiation among populations; our population-specific inbreeding coefficients consider only identity by descent (IBD) of alleles (making up genotypes) within populations. One could take direct account of population structure by introducing additional *F*-statistics that describe the probabilities of IBD of alleles sampled from different populations (in the case of individuals with mixed migrant ancestry). This could improve performance because the allele frequencies in populations with low levels of differentiation are not independent and genotype sample information can therefore be effectively "combined" across populations (through the use of *F*-statistics) to provide improved estimates of allele frequencies (in the extreme case, imagine two populations with no differentiation; a sample from the first population can be used to estimate allele frequencies in the second).

Another extension of our approach could be to allow immigration rates to vary over time. Posterior probabilities under models with constant or variable immigration rates could then be compared, using predictive posterior probabilities (see BERNARDO and SMITH 2000) to test the hypothesis of constant immigration rates during the last few generations. This might potentially allow one to directly address the relationship between immigration and gene flow. Strictly speaking, gene flow involves both immigration and local reproduction. If the rates of migration in the current and previous generations are similar this suggests that there is no difference in breeding success between residents and migrants (gene flow equals immigration rate), etc.

A disadvantage of our method, as currently formulated, is that it allows only the proportions of immigrants in a population to be estimated; it does not allow one to estimate directly the total proportion of individuals that emigrate from a population or the proportion that emigrate from one particular population to another. For example, a small population may have a large proportion of the total individuals in the population migrating to a particular large population but the fraction of migrants detected in the large population will be low

(because of the relative difference in the population sizes) and will provide no indication of the large proportion of actual emigration from the source population. One way to deal with this would be to estimate emigration rates that are corrected for the relative population sizes (if known). Alternatively, if temporal samples were collected, with replacement, information from unique individual genotypes (or mark-recapture tagging) could be combined with a method such as ours to jointly estimate population sizes and relative migration rates. Another assumption of the method is that all populations exchanging migrants have been sampled. The effect of this assumption may be important and a goal of future research should be to devise models that allow some degree of migration from "unobserved" populations for which no reference allele frequencies are available. This could likely be done using the flexible MCMC framework presented here.

To derive the prior probability distribution for individual immigrant ancestries, we have considered the distribution of migrant ancestries in a population in the limit of low migration rates and random mating between migrants and residents. Simulation studies are needed to determine the robustness of this approximation in the face of high migration rates and local inbreeding among migrant founders. Despite some outstanding issues of interpretation and reliability, methods for estimating recent migration rates using multilocus genotypes, such as we have presented here, should provide a useful (and complementary) alternative to existing methods, on the basis of diffusion approximations or coalescent theory, aimed at estimating historical migration rates under particular demographic scenarios. On balance, we are optimistic that new methods for inferring contemporary migration rates and gene flow will ultimately require fewer assumptions and will yield information that is highly relevant to conservation biologists, ecologists, human geneticists, and others dealing with practical problems involving recent (or ongoing) migration and admixture among study populations.

The program BayesAss, written in C, is available from our website at http://rannala.org.

We thank Helene Freville and Isabelle Olivieri for providing us with their plant microsatellite data and Lindsey Carmichael and Curtis Strobeck for providing us with their wolf microsatellite data. We are grateful to the two anonymous reviewers. This research was supported by the National Institutes of Health grant HG01988 and Canadian Institutes of Health research grant MOP 44064 to B.R.

LITERATURE CITED

Anderson, E. C., and E. A. Thompson, 2002 A model-based approach for identifying species hybrids using multilocus genetic data. Genetics **160:** 1217–1229.

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

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

Bernardo, J. M., and A. F. M. Smith, 2000 *Bayesian Theory*. Wiley, New York.

Carmichael, L. E., J. A. Nagy, N. C. Larter and C. Strobeck, 2001 Prey specialization may influence patterns of gene flow in wolves of the Canadian Northwest. Mol. Ecol. **10:** 2787–2798.

Casella, G., and R. L. Berger, 1990 *Statistical Inference*. Duxbury Press, Belmont, MA.

Cornuet, J. M., S. Piry, G. Luikart, A. Estoup and M. Solignac, 1999 New methods employing multilocus genotypes to select or exclude populations as origins of individuals. Genetics **153:** 1989–2000.

Dawson, K. J., and K. Belkhir, 2001 A Bayesian approach to the identification of panmictic populations and the assignment of individuals. Genet. Res. **78:** 59–77.

Freville, H., F. Justy and I. Olivieri, 2001 Comparative allozyme and microsatellite population structure in a narrow endemic plant species, *Centaurea corymbosa* Pourret (Asteraceae). Mol. Ecol. **10:** 879–889.

Gaggiotti, O. E., F. Jones, W. M. Lee, W. Amos, J. Harwood *et al.*, 2002 Patterns of colonization in a metapopulation of grey seals. Nature **416:** 424–427.

Gamerman, D., 1997 *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman & Hall, New York.

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

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

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

Paetkau, D., W. Calvert, I. Stirling and C. Strobeck, 1995 Microsatellite analysis of population structure in Canadian polar bears. Mol. Ecol. **4:** 347–354.

Pritchard, J. K., M. Stephens and P. Donnelly, 2000 Inference of population structure using multilocus genotype data. Genetics **155:** 945–959.

Rannala, B., and J. L. Mountain, 1997 Detecting immigration by using multilocus genotypes. Proc. Natl. Acad. Sci. USA **94:** 9197–9201.

Slatkin, M., and N. H. Barton, 1989 A comparison of three indirect methods for estimating average levels of gene flow. Evolution **43:** 1349–1368.

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

Vitalis, R., and D. Couvet, 2001 Estimation of effective population size and migration rate from one- and two-locus identity measures. Genetics **157:** 911–925.

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

Wright, S., 1969 *Evolution and Genetics of Populations: The Theory of Gene Frequencies*, Vol. 2. University of Chicago Press, Chicago.

Communicating editor: J. Hey

APPENDIX A

Expected migrant proportions: Define the probability (per generation) that an individual is a migrant as m and let ϕ be the expected fraction of alleles at a locus that is derived from migrants. By enumerating all possible patterns of ancestry (individual is a migrant, individual is a nonmigrant but both parents are migrants, etc.) that result in a given value of ϕ, we obtain

$$\begin{aligned}\Pr(\phi = 1) &= m + (1 - m)m^2 + \ldots \\ &= m + O(m^2).\end{aligned}$$

$$\begin{aligned}\Pr\left(\phi = \frac{1}{2}\right) &= 2m(1 - m)^2 + \binom{4}{2}m^2(1 - m)^5 + \ldots \\ &= 2m + O(m^2).\end{aligned}$$

$$\Pr\left(\phi = \frac{1}{4}\right) = 4m(1 - m)^6 + \binom{8}{2}m^2(1 - m)^{13} + \ldots$$
$$= 4m + O(m^2),$$

where the notation $O(m^2)$ denotes terms of order m^2 and higher. The first term in each series is the probability of a single migrant ancestor: In the case of $\phi = 1$ the individual is a migrant (at generation 1); in the case of $\phi = ½$ the individual has a migrant parent (at generation 2); and in the case of $\phi = ¼$ the individual has a migrant grandparent (at generation 3). Several possible ancestries leading to $\phi = 1$ and $\phi = ½$ are shown in Figure A1. The other terms allow possibilities such as two immigrant parents at generation 2 (in the case of $\phi = 1$), etc.

The first term in each of the three possibilities listed above (migrant, migrant parent, and migrant grandparent) is a linear function of m, and the remaining terms are of order m^2 and higher. If m is small the higher-order terms can be neglected and we need consider only possibilities involving a single migrant ancestor at some generation (this approximation is implicit in the method of Rannala and Mountain 1997). In the limit of small m, we expect a fraction m of individuals in the population to be first-generation migrants, a fraction $2m$ to have one migrant parent, a fraction $4m$ to have one migrant grandparent, and so on. Individuals with migrant ancestry beyond parents will have only one-quarter of their genome derived from migrant ancestors, on average, and for smaller numbers of loci such individuals will be statistically indistinguishable from nonmigrants; we have therefore chosen to use only the previous two generations of migrant ancestry to estimate m, although more distant generations could also be included with sufficient numbers of loci.

Constant allele frequencies: Assume that a Fisher-Wright population of constant size N_e receives migrants at rate m. The deterministic change in allele frequency in the population due to migration in each generation is

$$\Delta p = -m\Delta p_0,$$

where $\Delta p = p_1 - p_0$ is the change in the population allele frequency in a single generation and $\Delta p_0 = p_0 - p_m$ is the difference in allele frequency between the population that is the migrant source and the population from which individuals are sampled. For a single diallelic locus, the measure of population differentiation, F_{ST}, is defined as

$$F_{ST} = \frac{\sigma_p^2}{\bar{p}(1 - \bar{p})}$$

(see Wright 1969), where $\bar{p}$ is the average allele frequency across populations and σ_p^2 is the variance of allele frequencies across populations. We can write σ_p^2 for our pair of populations as

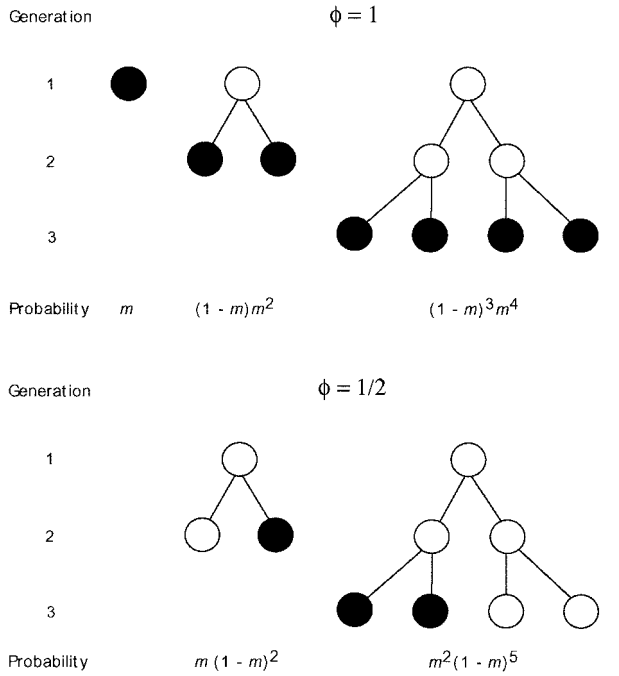

Figure A1.—Several possible patterns of immigrant ancestry that would each result in either all of an individual's genes arising from an immigrant source (top of figure, $\phi = 1$) or one-half of an individual's genes arising from an immigrant source (bottom of figure, $\phi = ½$). Immigrants are denoted by solid circles and nonimmigrants by open circles. The probability of each pattern, given a migration rate m and assuming random mating, is given below each part.

$$\sigma_p^2 = \frac{1}{2}\left[(\bar{p} - p_0)^2 + (\bar{p} - p_m)^2\right],$$
$$= \frac{1}{2}\left[\left(\frac{p_0 + p_m}{2} - p_0\right)^2 + \left(\frac{p_0 + p_m}{2} - p_m\right)^2\right],$$
$$= \frac{1}{4}\Delta p_0^2.$$

For a given value of F_{ST}, the difference Δp takes on its most extreme values when $\bar{p} = ½$ and the value of F_{ST} is then Δp_0^2. In that case, we can rewrite Δp as

$$|\Delta p| = m\sqrt{F_{ST}}.$$

We now have an expression for the magnitude of the change of allele frequency (per generation) under migration pressure in a population that receives migrants from another population with a specified level of differentiation between the populations. If $m < 0.05$ and $F_{ST} < 0.05$ then $\Delta p < 0.01$ and the change of allele frequency over a few generations will be negligibly small. Similarly, twice the standard deviation of the allele frequency change due to drift will be

$$2\sigma_p = 2\sqrt{\frac{p_0(1 - p_0)}{2N_e}} = \sqrt{\frac{2p_0(1 - p_0)}{N_e}}.$$

The change in allele frequency under drift will be great-

est when $p_0 = \frac{1}{2}$ and in that case $2\sigma_p = \sqrt{1/2N_e}$. If $N_e > 5000$ then $2\,\sigma_p < 0.01$ and the change of allele frequency over a few generations will be negligibly small. These values define boundaries beyond which the approximations underlying the proposed method will be well satisfied. In such cases, the resulting estimates should be accurate. The method may provide reasonable estimates for larger values of m and F_{ST} (or smaller N_e) as well but the specific range of applicability remains to be shown. Simulation studies are needed to evaluate the performance of the method under a range of conditions.

APPENDIX B

MCMC algorithm: The Metropolis-Hastings (MH) algorithm (Metropolis *et al.* 1953; Hastings 1970) was used to numerically calculate the posterior probability density of the parameters in our analyses. The basic idea is to construct a Markov chain with a stationary distribution that is the joint posterior distribution of the parameters to be estimated. This chain is simulated and samples from the chain are used to make inferences about joint or marginal posterior probabilities of parameters. The implementation of the MH algorithm used in our program has four steps at each iteration of the chain. At each step (outlined below) a particular set of parameters are potentially modified.

Modifying population migration rates: The matrix of population migration rates at iteration a, denoted as $\mathbf{m}[a]$, is modified to be $\mathbf{m}[a+1] = \mathbf{m}^*$ with probability

$$\alpha_{\mathbf{m}}(\mathbf{m}^*|\mathbf{m}[a]) = \min\left\{1, \frac{\Pr(\mathbf{M}, \mathbf{t}|\mathbf{m}^*)}{\Pr(\mathbf{M}, \mathbf{t}|\mathbf{m}[a])}\right\}.$$

The nominating function $g(\mathbf{m}^*|\mathbf{m}[a])$ is as follows: Choose one of the I^2 elements of the migration matrix to be modified with uniform probability $1/I^2$. The migration rates are constrained by our model such that

$$m_{ll} = 1 - \sum_{q\neq l}(m_{lq} + 2m_{lq}) = 1 - 3\sum_{q\neq l} m_{lq} \quad \text{and} \quad m_{ll} \geq 0.$$

It follows that

$$1 - 3\sum_{q\neq l} m_{lq} \geq 0, \quad \sum_{q\neq l} m_{lq} \leq \frac{1}{3}, \quad m_{ll} \in \left(\frac{2}{3}, 1\right).$$

To maintain these constraints, we used the following proposal scheme. If element l, q is chosen ($l \neq q$), the proposed value is $m_{lq}^* = m_{lq}[a] + z$, where z is chosen on a uniform interval $(-\delta_{\mathbf{m}}, +\delta_{\mathbf{m}})$ with reflecting boundaries, where $\delta_{\mathbf{m}} = \max\{0.10, 1 - m_{ll}\}$. If $m_{lq}^* > \frac{1}{3}$ or $m_{lq}^* < 0$ then m_{lq}^* is reflected back onto the interval $(0, \frac{1}{3})$ by an amount $m_{lq}[a] + z - \frac{1}{3}$ or $-z - m_{lq}[a]$. The remaining elements $j \neq q$ of row l are adjusted so that they sum to 1 by using the transformation

$$m_{lj}^* = \frac{m_{lj}(1 - m_{ll} - m_{lq}^*)}{\sum_{j\neq q, j\neq l} m_{lj}}, \quad \text{for all } j \neq q,\ j \neq l.$$

If element m_{ll} is chosen, the proposed value is $m_{ll}^* = m_{ll}[a] + z$, where z is chosen on a uniform interval $(-\delta_{\mathbf{m}}, +\delta_{\mathbf{m}})$ with reflecting boundaries, where $\delta_{\mathbf{m}} \leq \frac{1}{3}$. If $m_{ll}^* > 1$ or $m_{ll}^* < \frac{2}{3}$ then m_{lq}^* is reflected back onto the interval $(\frac{2}{3}, 1)$ by an amount $m_{lq}[a] + z - 1$ or $\frac{2}{3} - z - m_{lq}[a]$. The remaining elements are adjusted to sum to 1 using the transformation

$$m_{lj}^* = \frac{m_{lj}(1 - m_{ll}^*)}{\sum_{j\neq q} m_{lj}}.$$

We assumed a uniform Dirichlet prior for $\mathbf{m}$ and a uniform prior (on the integers 0, 1, 2) for t_i so that the terms in the MH ratio involving the priors for $\mathbf{m}$ and $\mathbf{t}$ cancel. The nominating function $g(m^*|m[a])$ described above is symmetrical so that these terms also cancel from the MH ratio.

Modifying individual migrant ancestries: The matrix of individual migrant ancestries at iteration a, denoted by the composite parameters $\mathbf{M}[a]$ and $\mathbf{t}[a]$, are modified to be $\mathbf{M}[a+1] = \mathbf{M}^*$ and $\mathbf{t}[a+1] = \mathbf{t}^*$ with probability

$$\alpha_{\mathbf{M},\mathbf{t}}(\mathbf{M}^*, \mathbf{t}^*|\mathbf{M}[a], \mathbf{t}[a]) = \min\left\{1, \frac{\Pr(\mathbf{M}^*, \mathbf{t}^*|m)\Pr(\mathbf{X}|\mathbf{S}; \mathbf{M}^*, \mathbf{t}^*, \mathbf{F}, \mathbf{p})g(\mathbf{M}[a], \mathbf{t}[a]|\mathbf{M}^*, \mathbf{t}^*)}{\Pr(\mathbf{M}[a], \mathbf{t}[a]|\mathbf{m})\Pr(\mathbf{X}|\mathbf{S}; \mathbf{M}[a], \mathbf{t}[a], \mathbf{F}, \mathbf{p})g(\mathbf{M}^*, \mathbf{t}^*|\mathbf{M}[a], \mathbf{t}[a])}\right\},$$

where

$$\frac{g(\mathbf{M}[a], \mathbf{t}[a]|\mathbf{M}^*, \mathbf{t}^*)}{g(\mathbf{M}^*, \mathbf{t}^*|\mathbf{M}[a], \mathbf{t}[a])} = \frac{n_{l^*q^*t^*} + 1}{n_{lqt}}.$$

The nominating function $g(\mathbf{M}^*, \mathbf{t}^*|\mathbf{M}[a], \mathbf{t}[a])$ is as follows: Choose one of the n sampled individuals to have its migrant ancestry modified with uniform probability $1/n$. There are $2I - 1$ possible states for the migrant ancestry of the chosen individual; it can be a nonmigrant or a first- or second-generation migrant from one of the remaining $I - 1$ populations. The proposed change for an individual must be to one of the $2I - 2$ states other than its present state and each possibility is assigned a uniform probability $1/(2I - 2)$.

Modifying population allele frequencies: The matrix of population allele frequencies at iteration a, denoted as $\mathbf{p}[a]$, is modified to be $\mathbf{p}[a+1] = \mathbf{p}^*$ with probability

$$\alpha_{\mathbf{p}}(\mathbf{p}^*|\mathbf{p}[a]) = \min\left\{1, \frac{\Pr(\mathbf{X}|\mathbf{S}; \mathbf{M}, \mathbf{t}, \mathbf{F}, \mathbf{p}^*)}{\Pr(\mathbf{X}|\mathbf{S}; \mathbf{M}, \mathbf{t}, \mathbf{F}, \mathbf{p}[a])}\right\}.$$

The nominating function $g(\mathbf{p}^*|\mathbf{p}[a])$ is as follows: Choose one of the I populations with uniform probability $1/I$, choose one of the J loci with uniform probability $1/J$, and choose one of the k_{lj} alleles at locus j in population l with uniform probability $1/k_{lj}$. If allele i at locus j in population l is chosen the proposed value is $p_{ijl}^* = p_{ijl}[a] + z$, where z is chosen on a uniform interval $(-\delta_{\mathbf{p}}, +\delta_{\mathbf{p}})$ with reflecting boundaries and the remaining allele frequencies are adjusted so that the proposed allele frequencies sum to 1.

Modifying population inbreeding coefficients: The vector of population inbreeding coefficients at iteration a, denoted as $\mathbf{F}[a]$, is modified to be $\mathbf{F}[a + 1] = \mathbf{F}^*$ with probability

$$\alpha_{\mathbf{F}}(\mathbf{F}^*|\mathbf{F}[a]) = \min\left\{1, \frac{\Pr(\mathbf{X}|\mathbf{S}; \mathbf{M}, \mathbf{t}, \mathbf{F}^*, \mathbf{p})}{\Pr(\mathbf{X}|\mathbf{S}; \mathbf{M}, \mathbf{t}, \mathbf{F}[a], \mathbf{p})}\right\}.$$

The nominating function $g(\mathbf{F}^*|\mathbf{F}[a])$ is as follows: Choose one of the I populations with uniform probability $1/I$. The proposed value is $F_l^* = F_l[a] + z$, where z is chosen on a uniform interval $(-\delta_{\mathbf{F}}, +\delta_{\mathbf{F}})$ with reflecting boundaries such that F_l^* remains on the interval $(-1, +1)$.

Modifying genotypes with missing data: If $\mathbf{X}_-$ is a submatrix of $\mathbf{X} = \{\mathbf{X}_-, \mathbf{X}_+\}$ containing the missing genotypes for each individual, the proposed genotypes at these loci at iteration a, denoted as $\mathbf{X}_-[a]$, were modified to be $\mathbf{X}_-[a + 1] = \mathbf{X}_-^*$ with probability

$$\alpha_{\mathbf{X}_-}(\mathbf{X}_-^*|\mathbf{X}_-[a]) = \min\left\{1, \frac{\Pr(\mathbf{X}_+, \mathbf{X}_-^*|\mathbf{S}; \mathbf{M}, \mathbf{t}, \mathbf{F}, \mathbf{p})}{\Pr(\mathbf{X}_+, \mathbf{X}_-[a]|\mathbf{S}; \mathbf{M}, \mathbf{t}, \mathbf{F}, \mathbf{p})}\right\}.$$

The nominating function $g(\mathbf{X}_-^*|\mathbf{X}_-[a])$ is as follows: Choose any one of the $L_T = \Sigma L_i$ loci with missing data with uniform probability $1/L_T$, where L_i is the number of loci with missing data for individual i. Modify the locus to become genotype u, v with uniform probabilities $2/[k_l(k_l - 1)]$ if $u \neq v$ and $1/k_l^2$ if $u = v$ where k_l is the number of alleles (in all sampled populations) at locus l.