

A New Bayesian Method to Identify the Environmental Factors That Influence Recent Migration

Pierre Faubet and Oscar E. Gaggiotti¹

*Génomique des Populations et Biodiversité Laboratoire d'Ecologie Alpine, CNRS UMR 5553,
Université Joseph Fourier, 38041 Grenoble, France*

Manuscript received September 29, 2007
Accepted for publication December 20, 2007

ABSTRACT

We present a new multilocus genotype method that makes inferences about recent immigration rates and identifies the environmental factors that are more likely to explain observed gene flow patterns. It also estimates population-specific inbreeding coefficients, allele frequencies, and local population F_{ST} 's and performs individual assignments. We generate synthetic data sets to determine the region of the parameter space where our method is and is not able to provide accurate estimates. Our simulation study indicates that reliable results can be obtained when the global level of genetic differentiation (F_{ST}) is $>1\%$, the number of loci is only 10, and sample sizes are of the order of 50 individuals per population. We illustrate our method by applying it to Pakistani human data, considering altitude and geographic distance as explanatory factors. Our results suggest that altitude explains better the genetic data than geographic distance. Additionally, they show that southern low-altitude populations have higher migration rates than northern high-altitude ones.

THE study of dispersal processes is an essential problem in ecology, population genetics, conservation, and management of wildlife. For this reason, the estimation of migration rates has been one of the most investigated problems in population biology. Migration parameters can be directly estimated using ecological approaches such as mark-release-recapture methods but they are not applicable to the study of large or extended metapopulations. In these cases, population genetics approaches provide a better alternative because the information contained in DNA can provide gene flow parameter estimates for different and complementary timescales. Methods based on coalescent theory provide long-term migration rates because they use the genealogical information contained in a sample of genes (*e.g.*, MIGRATE, BEERLI and FELSENSTEIN 2001). On the other hand, methods based on multilocus genotypes (*e.g.*, BAYESASS, WILSON and RANNALA 2003) provide estimates of recent immigration rates by extracting the gametic disequilibrium signal generated by immigrant individuals or their descendants.

Besides simply estimating migration rates, it is very important to identify the biotic and/or abiotic factors that influence them. This can be done by first obtaining gene flow estimates and then searching for correlations between them and various environmental variables (*e.g.*, GIORDANO *et al.* 2007). Such an approach requires the use of summary statistics that do not take advantage of all the information contained in genetic data. An alterna-

tive approach is to implement the joint analysis of genetic and nongenetic data. Several current methods that combine both genetic and geographic data can be used to detect recent migrants (*e.g.*, GENELAND, GUILLOT *et al.* 2005; TESS, FRANÇOIS *et al.* 2006) but they do not take into account other environmental factors.

In previous studies we presented methods that use genetic and environmental data to study colonization processes (GAGGIOTTI *et al.* 2002, 2004) and population genetic structure (FOLL and GAGGIOTTI 2006). These approaches based on hierarchical Bayesian methods (*e.g.*, GELMAN *et al.* 1995) estimate the probability that a given environmental factor influences the parameters of interest (*e.g.*, composition of colonizing groups or local population F_{ST} 's) because they explicitly model the relationship between them and the relevant ecological factors. In this article we present a new multilocus genotype method for inferring recent immigration rates and identifying the environmental factors that best explain observed gene flow patterns. We use a hierarchical Bayesian approach that introduces nongenetic data through the prior distribution of the migration rates. Following WILSON and RANNALA's (2003) approach we implement the estimation of inbreeding coefficients to allow for departures from Hardy–Weinberg equilibrium within local populations. Finally, the method infers the population ancestry of individuals by assigning their alleles to populations from which they originated. We carry out a simulation study to identify the region of parameter space where the method is and is not able to provide accurate posterior estimates. We also illustrate our method with a real data example.

¹Corresponding author: LECA, BP 53, 2233 Rue de la Piscine, 38041 Grenoble Cedex 9, France. E-mail: oscar.gaggiotti@ujf-grenoble.fr

DATA AND MODEL PARAMETERS

Inferring migration rates from genetic data: The method is based on a population genetics model that differs from that used by WILSON and RANNALA (2003). More specifically, instead of assuming that sampling takes place right after migration, we consider that this is done after reproduction and before migration. Let us consider a metapopulation of a diploid species with nonoverlapping generations that is subdivided into I demes that can exchange migrants. Let $\mathbf{X} = (\mathbf{X}_{hl})$ be the observed multilocus genotypes of n individuals scored at L marker loci, where \mathbf{X}_{hl} denotes the genotype of individual h at locus l . We assume that n_i individuals were sampled from population i and use the vector $\mathbf{S} = (S_h)$ to identify the population S_h where the individual h was sampled from.

Population allele frequencies are given by a matrix \mathbf{p} composed of vectors $\mathbf{p}_{il} = (p_{ila})$ that give the frequency of allele a at locus l for population i . Following FALUSH *et al.* (2003), we consider a model with correlated allele frequencies based on the approach introduced by BALDING and NICHOLS (1995). Thus, we assume that before the last generation, the population was at migration–drift equilibrium so that allele frequencies in each population are determined by the global allele frequencies in the metapopulation as a whole, $\hat{\mathbf{p}}_l = (\hat{p}_{la})$, and the degree of genetic differentiation between each local population and the overall metapopulation, $\boldsymbol{\theta} = (\theta_i)$, where $\theta_i = 1/F_{ST}^i - 1$. Finally, to allow departures from Hardy–Weinberg equilibrium, we introduce population-specific inbreeding coefficients $\mathbf{F} = (F_i)$, where F_i is the inbreeding coefficient for population i . Thus, we consider two levels of inbreeding, one at the population level corresponding to F_{ST} and another one at the individual level, corresponding to F_{IS} .

Instead of focusing directly on individual migration rates, we consider the probability that genes in a deme originated in another one over the last generation. Thus, migration is described by a matrix $\mathbf{m} = (m_{ij})$, where m_{ij} is the probability that alleles in population i came from population j during the previous generation. The ancestral state of the individuals is described by a matrix $\mathbf{M} = (\mathbf{M}_h)$, where $\mathbf{M}_h = (i, j)$ is a two-element vector identifying the source demes (i and j) for the two alleles of individual h . All possible ancestry states are considered: both alleles come from the deme where the individual was sampled, or both come from another deme, or they come from two different ones. Thus, migration rates for individuals are obtained as

$$\tilde{m}_{ijk} = \begin{cases} m_{ij}^2 & \text{if } j = k \\ 2m_{ij}m_{ik} & \text{if } j \neq k \end{cases}, \quad j \leq k, \quad (1)$$

where \tilde{m}_{ijk} is the probability that individuals sampled from population i belong to the ancestry class (j, k). Note that our approach estimates migration rates only over the last generation. Moreover, as opposed to WILSON

and RANNALA (2003) migration rates vary freely in the interval $(0, 1)$ and do not have to be small.

The model parameters described above ($\mathbf{p}, \hat{\mathbf{p}}, \boldsymbol{\theta}, \mathbf{F}, \mathbf{M}, \mathbf{m}$) are estimated from the genetic data using a Bayesian approach and Markov chain Monte Carlo (MCMC) techniques.

Likelihood: The likelihood is the probability of the observed genotypes given model parameters and is constructed by defining the probability of observing the genotype of individual h at locus l in terms of the ancestry classes. We note these genotypes $\mathbf{X}_{hl} = (X_{hl1}, X_{hl2})$, where X_{hlc} is the allele observed at locus l in chromosome $c = 1, 2$ of individual h . Thus, individual h genotype likelihood at locus l is given by

$$\Pr(\mathbf{X}_{hl} | \mathbf{M}_h, \mathbf{F}, \mathbf{p}) = \begin{cases} \phi(\mathbf{X}_{hl}, i) & \text{if } \mathbf{M}_h = (i, i) \\ p_{ilX_{hl1}}p_{jlX_{hl2}} + \gamma p_{ilX_{hl1}}p_{ilX_{hl2}} & \text{if } \mathbf{M}_h = (i, j), \end{cases} \quad (2)$$

where

$$\phi(\mathbf{X}_{hl}, i) = \begin{cases} (1 - F_i)p_{ilX_{hl1}}^2 + F_i p_{ilX_{hl2}} & \text{if } X_{hl1} = X_{hl2} \\ 2(1 - F_i)p_{ilX_{hl1}}p_{ilX_{hl2}} & \text{otherwise} \end{cases} \quad (3)$$

and

$$\gamma = \begin{cases} 0 & \text{if } X_{hl1} = X_{hl2} \\ 1 & \text{otherwise.} \end{cases} \quad (4)$$

The first case considered in Equation 2 corresponds to the scenario where both alleles originated in the same source population, in which case we need to take into account possible deviations from Hardy–Weinberg equilibrium (see Equation 3). The second case considers that the individual is the descendant of parents that come from two different source populations, in which case we need to take into account that there are two different ways of assigning the alleles to the parents.

If we assume that individuals were sampled at random and loci are unlinked, then the likelihood of the whole sample is obtained by multiplying across all loci and individuals,

$$\Pr(\mathbf{X} | \mathbf{M}, \mathbf{F}, \mathbf{p}) = \prod_{h=1}^n \prod_{l=1}^L \Pr(\mathbf{X}_{hl} | \mathbf{M}_h, \mathbf{F}, \mathbf{p}). \quad (5)$$

This likelihood can be used as the basis for inference using a Bayesian approach.

Combining genetic and environmental data: One can expect that migration patterns are influenced by environmental factors such as population densities, distances between local populations, etc. To identify which environmental factors have influenced gene flow we use GAGGIOTTI *et al.*'s (2004) approach. Let us suppose that our knowledge of the species under study leads us to think that R environmental factors $\mathbf{G} = (G^{(r)})$ may influence the migration process. We can then introduce their effect through the prior distribution of gene mi-

gration rates. More specifically, we focus on the ancestry of immigrant alleles by conditioning on not being a resident allele

$$m_{ij}^\star = \frac{m_{ij}}{1 - m_{ii}} \tag{6}$$

and assume that the vector $\mathbf{m}_i^\star = (m_{ij}^\star)_{j \neq i}$ follows a Dirichlet distribution; *i.e.*, $\mathbf{m}_i^\star \mid \psi_i \sim \text{Dir}(\psi_i)$, where $\psi_i = (\psi_{ij})_{j \neq i}$ are shape parameters for the Dirichlet distribution. Furthermore, we assume that each shape parameter ψ_{ij} follows a lognormal distribution; *i.e.*, for each pair of distinct populations $i \neq j$

$$\log \psi_{ij} \sim \mathcal{N}(\mu_{ij}, \sigma^2), \tag{7}$$

where the mean μ_{ij} is given by the generalized linear regression

$$\mu_{ij} = \alpha_0 + \sum_r \alpha_r G_{ij}^{(r)} + \sum_{r < s} \alpha_{rs} G_{ij}^{(r)} G_{ij}^{(s)}, \tag{8}$$

where α_r denotes the effect of environmental factor r and α_{rs} denotes the effect of first-order interactions between factors r and s ; these parameters are collected into a single vector $\alpha = (\alpha_r, \alpha_{rs})$. The sign and the magnitude of the α 's tell us about the direction and the strength of the environmental factors. Finally, σ^2 is the amount of variation that remains unexplained by the regression and $G_{ij}^{(r)}$ is the observed value for factor r , which is hypothesized to influence migration between populations i and j . To reduce posterior correlation and to simplify prior elicitation and posterior interpretation process, explanatory factors are normalized before analysis so that they have zero mean and variance one.

By excluding different regression terms we can define different alternative models. We note, however, that as opposed to previous applications of this approach (*cf.* GAGGIOTTI *et al.* 2004; FOLL and GAGGIOTTI 2006), the intercept α_0 is included in all models because it takes into account the effect of factors that act at a geographic scale larger than that of the metapopulation under study (see DISCUSSION for more details).

Other priors: We assume that there is no prior information on the shape of the other parameters and, therefore, adopt the vague priors that are given in the APPENDIX. Note that in the particular case of the probability to observe nonmigrant genes (*i.e.*, m_{qq}), we adopt a uniform prior between 0 and 1 because, although some environmental factors may influence whether or not an individual decides to emigrate, our method is aimed at estimating immigration rates and, therefore, cannot take into account this possibility.

Posterior distribution: The model is now expressed in terms of parameters $\Theta = (\mathbf{p}, \tilde{\mathbf{p}}, \theta, \mathbf{F}, \mathbf{M}, \mathbf{m}, \psi, \alpha, \sigma^2)$ and the corresponding posterior distribution is given by Bayes' rule:

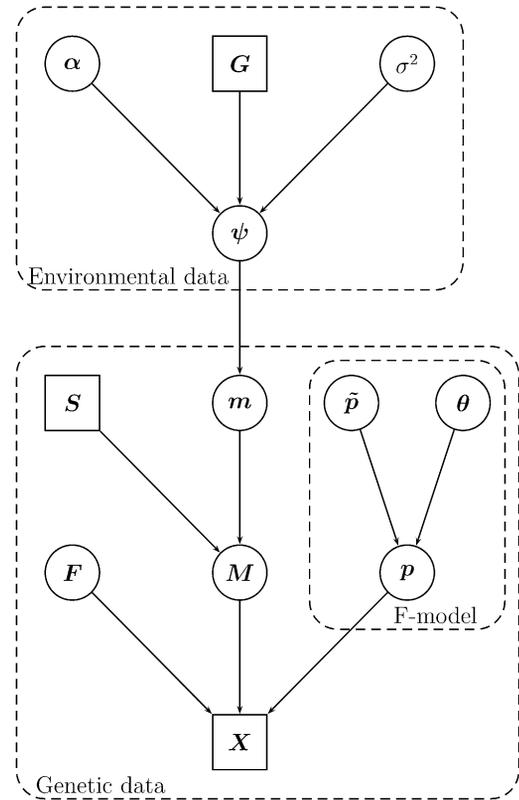


FIGURE 1.—The DAG for the model given in Equation 9. Square nodes denote known quantities (*i.e.*, data) and circles represent model parameters to be estimated. Arrows between nodes represent direct stochastic relationships within the model. The variables within each node correspond to model parameters discussed in the text.

$$f(\Theta \mid \mathbf{X}, \mathbf{S}, \mathbf{G})$$

$$\propto \underbrace{\Pr(\mathbf{X} \mid \mathbf{M}, \mathbf{F}, \mathbf{p}) \Pr(\mathbf{M} \mid \mathbf{S}; \mathbf{m})}_{\text{Genetic data}} \underbrace{f(\mathbf{F}) f(\mathbf{p} \mid \tilde{\mathbf{p}}, \theta) f(\tilde{\mathbf{p}}) f(\theta)}_{\text{F-model}} \times \underbrace{f(\mathbf{m} \mid \psi) f(\psi \mid \alpha, \sigma^2, \mathbf{G}) f(\alpha) f(\sigma^2)}_{\text{Environmental data}}. \tag{9}$$

The full model is represented by the directed acyclic graph (DAG) in Figure 1.

The posterior distributions of parameters given in Equation 9 are estimated using MCMC methods that are described in the supplemental information.

Posterior model probabilities: Besides estimating migration rates our method is aimed at identifying the environmental factors influencing gene flow. As we mentioned before, several alternative models can be obtained from the full regression (8) by canceling elements of the vector α . Note that models that include first-order interactions between factors r and s are allowed only if both factors are included. Thus for model \mathcal{M} , the corresponding posterior distribution is given by

$$\begin{aligned}
& f(\Theta_{\mathcal{M}} | \mathbf{X}, \mathbf{S}, \mathbf{G}) \\
& \propto \Pr(\mathbf{X} | \mathbf{M}, \mathbf{F}, \mathbf{p}) \Pr(\mathbf{M} | \mathbf{S}; \mathbf{m}) f(\mathbf{F}) f(\mathbf{p} | \hat{\mathbf{p}}, \boldsymbol{\theta}) f(\hat{\mathbf{p}}) f(\boldsymbol{\theta}) \\
& \quad \times f(\mathbf{m} | \boldsymbol{\psi}) f(\boldsymbol{\psi} | \alpha_{\mathcal{M}}, \sigma^2, \mathbf{G}) f(\alpha_{\mathcal{M}}) f(\sigma^2) \Pr(\mathcal{M}),
\end{aligned} \tag{10}$$

where $\Theta_{\mathcal{M}}$ is the parameter vector under model \mathcal{M} , $\alpha_{\mathcal{M}}$ is the corresponding regression vector, and $\Pr(\mathcal{M})$ denotes prior model probability. Posterior model probabilities are estimated using the reversible-jump (RJ) MCMC approach (GREEN 1995, detailed in supplemental information). Here we note only that one of the problems faced when estimating posterior model probabilities is that the prior for σ_{α}^2 can have a large effect on the estimates. When very vague priors are used, more posterior weight is placed on the model with the fewest parameters. This is the well-known Jeffreys–Lindley paradox (ROBERT 1994). This problem was avoided by first running an MCMC of the full model with vague priors and then using the posterior estimates of the α 's as informative priors for a new MCMC run.

SIMULATION STUDY

We evaluated the sensitivity of our method by generating synthetic data under a particular scenario in which gene flow is influenced by two factors. We considered various levels of genetic differentiation and migration rates. We are interested in the ability of our algorithm to find the correct scenario and to provide accurate posterior estimates for migration rates (with corresponding fairly narrow highest posterior density intervals, HPDI).

Generation of synthetic data: We simulated data following the inference model presented in Figure 1. We initially considered a scenario with $I = 4$ populations, each with local population sizes of $N_i = 5000$ individuals. The sample size per population was $n_i = 100$ and we assumed that each sampled individual was scored for $L = 10$ polymorphic loci, each with $K = 10$ alleles.

Generating migration rates from environmental factors: We consider two environmental factors that could be, for example, geographic distance (G_1) and population density (G_2). The pairwise geographic distances were generated using a standard normal distribution. The pairwise differences in population density were generated by first filling the top triangular matrix with values drawn from a standard normal and then filling the bottom triangular matrix with the opposite values. This procedure is equivalent to standardizing the observed pairwise differences in environmental factors before analysis.

To generate the migration matrix, we first chose the values for the diagonal elements (proportion of non-migrant genes) and then we calculated the values for the nondiagonal elements (immigration rates) using the following procedure. We set $\alpha_1 = -0.9$, $\alpha_2 = 1.1$, and $\alpha_{12} = 0$ (*i.e.*, no interaction effect) and calculated μ_{ij} 's using Equation 8. Assuming no deviation from the linear

regression (*i.e.*, $\sigma^2 = 0$), we set $\psi_{ij} = e^{\mu_{ij}}$. Finally we computed the means $E[m_{ij}^* | \boldsymbol{\psi}_i] = \psi_{ij} / \sum_{j \neq i} \psi_{ij}$ of the Dirichlet distribution used for the migration rates (Equation A2) and rescaled them so that they added up to $1 - m_{ii}$.

Genetic data: To generate multilocus genotypes with a given level of genetic differentiation, we need to generate parametric allele frequencies. This task was performed using BALDING and NICHOLS' (1997) sampling formula for F_{ST} . According to this formula, genes are sampled one by one during an iterative process. The probability that the next gene sampled is a after having sampled n genes of which n_a correspond to allele a is given by

$$p_a(n_a, n) = \frac{n_a F_{ST} + (1 - F_{ST}) p_a}{1 + (n - 1) F_{ST}}, \tag{11}$$

where p_a is the global frequency of allele a in the metapopulation.

To generate the parametric allele-frequency distribution of each locus in any given local population for a given F_{ST} value, we first sample an allele at random from the metapopulation allele-frequency distribution and then use Equation 11 to calculate the probability distribution for the type of the allele that will be sampled next. Using this distribution we obtain the next allele. This process is repeated iteratively until we obtain the $2N_i$ alleles present in local population i . We used uniform allele frequencies for the metapopulation.

Large departures from the target F_{ST} value were avoided by using the following iterative process. We generated the local population's allele-frequency distributions and calculated the global and pairwise F_{ST} 's. If one or more of these values were not within 10% of the target value we discarded the allele frequencies and generated new ones. If the new allele frequencies satisfied the requirement, we generated the genotypes; otherwise we continued this iterative procedure until the constraint was satisfied. This procedure was used to control for the effect of genetic differentiation on the performance of the method.

Using the gene migration rates calculated above, we obtained the proportion of migrant individuals in each population using Equation 1. Multilocus genotypes for each local population were generated assuming Hardy–Weinberg equilibrium. Genotypes of nonmigrant individuals were obtained by drawing two alleles from the local population allele-frequency distribution. For the migrant individuals with both parents coming from the same local population, the genotype was obtained by drawing two alleles at random from the parents' source population. For migrant individuals with parents coming from different source populations, we sampled one allele from the source population of each parent. Finally, samples were generated by drawing n_i individuals from each local population, keeping track of the ancestry of both alleles for each sampled individual.

TABLE 1
Posterior estimates for various levels of genetic differentiation and high gene flow

Factors included		F_{ST}			
		0.01	0.05	0.10	0.25
None		0.621	0.073	0.026	0.015
G_1		0.142	0.061	0.028	0.016
G_2		0.170	0.184	0.082	0.052
G_1 and G_2		0.045	0.472	0.637	0.701
With interaction		0.022	0.210	0.227	0.216
Parameter		Estimate/RMSE/95% HPDI			
		0.01	0.05	0.10	0.25
α_1	-0.900	^a	-0.921 0.057 (-1.754; -0.097)	-0.974 0.027 (-1.639; -0.324)	-0.945 0.003 (-1.503; -0.370)
α_2	1.100	^a	1.137 0.013 (0.285; 2.068)	1.178 0.011 (0.504; 1.869)	1.149 0.002 (0.554; 1.745)
σ^2	—	0.389 — (0.120; 1.159)	0.426 — (0.121; 2.858)	0.355 — (0.106; 2.010)	0.306 — (0.107; 1.578)
Assignments		0.01	0.05	0.10	0.25
Misassignments		0.754	0.280	0.110	0.002
Probabilities ^b		0.223	0.700	0.883	0.996

Posterior model probabilities, regression parameter mean estimates, and assignment accuracy for synthetic data generated with proportions of nonmigrant alleles set to $m_{ii} = 0.70$ are shown.

^aThe regression parameter is not included in the model with the highest posterior model probability.

^bMaximum posterior assignment probabilities averaged across all individuals.

Implementation details: Each MCMC was run for 1,030,000 iterations. The first 20,000 iterations consist of short pilot runs used to tune up the proposal distributions to obtain acceptance rates between 25 and 45%. The next 10,000 iterations were discarded as burn-in and the remaining observations were sampled every 100 iterations, giving a sample size of 10,000 for each analysis.

To take into account model uncertainty, parameters are estimated using Bayesian model-averaging methods. The only exception to this rule are the regression parameters, which are model specific and, therefore, were estimated using the subset of values corresponding to the model with the highest posterior probability. Finally, posterior model probabilities are obtained by observing the number of times the chain visits each alternative model.

Posterior estimates are based on the sample mean except for the deviation from the regression σ^2 , which usually has a highly asymmetric posterior distribution. In this latter case we used the posterior mode, which was estimated using kernel density estimation.

We investigated the effect of varying three parameters: the level of genetic differentiation, $F_{ST} = \{0.01, 0.05, 0.10, 0.25\}$; the proportion of nonmigrant alleles, $m_{ii} = \{0.7, 0.9\}$; and the number of populations, $I = \{4, 6\}$.

For each parameter set we generated 10 independent genetic data sets as described above. The results we present below are averages across these 10 replicates. As a measure of accuracy we also present the relative mean square errors (RMSE).

Results: We investigated the performance of our method to provide reliable estimates under different scenarios of migration and genetic differentiation and number of populations studied. We consider first the effects on model determination, and then we address the influence on migration rate estimates and finally on individual assignments.

When the immigration rate is high ($m_{ii} = 0.7$; see Table 1), estimates of posterior model probabilities are strongly influenced by the degree of genetic differentiation (F_{ST}). When differentiation is low ($F_{ST} = 0.01$), the method fails to identify the model used to generate the synthetic data. However, the correct model is identified when $F_{ST} > 0.01$, and, moreover, its posterior model probability increases steadily with increasing genetic differentiation. The estimation of regression parameters is also influenced by the magnitude of F_{ST} but to a lesser degree. The RMSE decreases with increasing genetic differentiation but the bias is largely unaffected.

TABLE 2
Posterior estimates for various levels of genetic differentiation and low gene flow

Factors included		F_{ST}			
		0.01	0.05	0.10	0.25
None		0.444	0.106	0.056	0.028
G_1		0.122	0.072	0.055	0.030
G_2		0.291	0.301	0.122	0.073
G_1 and G_2		0.075	0.323	0.521	0.638
With interaction		0.068	0.197	0.246	0.231
Parameter		Estimate/RMSE/95% HPDI			
		0.01	0.05	0.10	0.25
α_1	-0.900	^a	-0.858 0.123 (-1.984; 0.263)	-1.122 0.101 (-2.091; -0.191)	-1.010 0.015 (-1.723; -0.321)
α_2	1.100	^a	1.403 0.127 (0.197; 2.687)	1.314 0.071 (0.387; 2.279)	1.173 0.008 (0.473; 1.886)
σ^2	-	0.486 — (0.131; 3.067)	0.513 — (0.125; 4.117)	0.452 — (0.120; 3.090)	0.352 — (0.110; 1.956)
Assignments		0.01	0.05	0.10	0.25
Misassignments		0.808	0.134	0.046	0.002
Probabilities ^b		0.288	0.847	0.946	0.997

Posterior model probabilities, regression parameter mean estimates, and assignment accuracy for synthetic data generated with proportions of nonmigrant alleles set to $m_{ii} = 0.90$ are shown.

^aThe corresponding regression parameter is not included in the model with the highest posterior model probability.

^bMaximum posterior assignment probabilities averaged across all individuals.

Thus, it is the accuracy of the estimates (as illustrated by the HPDIs) that is influenced by F_{ST} . The proportion of the variance that remains unexplained by the model, σ^2 , decreases as genetic differentiation increases.

Decreasing the immigration rate ($m_{ii} = 0.90$) has a detrimental effect on estimates (Table 2). Although the true model is correctly identified for $F_{ST} > 0.01$, its posterior probability is lower than that observed when $m_{ii} = 0.70$. Estimates of regression parameters are more biased and less accurate (wider HPDIs), leading to higher RMSEs. Also, the proportion of the variance that remains unexplained, σ^2 , is larger. Note, however, that as was the case before, the quality of all estimates improves with increasing genetic differentiation.

Increasing the number of populations studied ($I = 6$) improves model determination (Table 3). More precisely, the posterior probability of the true model is strongly increased and the proportion of variance that remains unexplained decreases sharply (see Table 3 and last columns of Tables 1 and 2). However, the effect on the quality of the regression parameter estimates is somewhat decreased since the bias and the RMSE increase. Nevertheless, the width of the HPDIs decreases, indicating that the precision increases.

Estimates of gene migration rates improve with increasing genetic differentiation (Table 4). The bias decreases sharply between $F_{ST} = 0.01$ and 0.1 and then remains very low. Note that the only cases where the HPDI does not include the true value correspond to the case with the weakest genetic differentiation ($F_{ST} = 0.01$). When the number of nonmigrant genes decreases we observe the same pattern but in this case the number of estimates for which the HPDIs do not include the true value is smaller and corresponds only to the estimates of nonmigrant proportions (Table 5).

In terms of posterior individual assignments, increasing genetic differentiation improves the quality of the estimation (see the bottom three rows of Tables 1 and 2). That is, the proportion of individuals that are misassigned decreases while the average posterior assignment probability increases. Decreasing the proportion of migrant genes also improves the quality of assignments; the proportion of misassignments decreases and the average posterior probabilities with which individuals are assigned increase. The effect of varying the number of populations is very small, being somewhat more distinguishable when the proportion of nonmigrants is larger (see bottom three rows of Tables 1–3).

TABLE 3

Posterior estimates for the scenario with six populations

Factors included		$F_{ST} = 0.25$	
		$m_{ii} = 0.70$	$m_{ii} = 0.90$
None		0.000	0.000
G_1		0.000	0.000
G_2		0.000	0.001
G_1 and G_2		0.915	0.883
With interaction		0.085	0.116
Parameter		Estimate/RMSE/95%HPDI	
		True value	$m_{ii} = 0.70$
α_1	-0.900	-1.056 0.031 (-1.416; -0.707)	-1.022 0.022 (-1.521; -0.537)
α_2	1.100	1.244 0.018 (0.960; 1.531)	1.246 0.018 (0.856; 1.659)
σ^2	—	0.164 — (0.080; 0.381)	0.213 — (0.096; 0.566)
Assignments		$m_{ii} = 0.70$	$m_{ii} = 0.90$
Missassignments		0.010	0.002
Probabilities ^a		0.985	0.997

Model determination, regression parameters, mean estimates, and assignment accuracy for synthetic data when varying nonmigrant gene proportions m_{ii} .

^aMaximum posterior assignment probabilities averaged across all individuals.

We also investigated what is the effect of using explicative variables that are different from the ones used to generate the synthetic data. The results (Table 6) show that the highest posterior probability is assigned to the null model, which indicates that the method does not wrongly identify as important factors that are not responsible for the observed migration pattern.

It is also important to investigate the effect of the amount of data used for the estimation, which can be characterized by the sample sizes and number of loci scored. The effect of decreasing the sample size from 100 to 50 individuals per population does not have much of an effect on posterior model probabilities while estimates of regression parameters have a slightly larger bias and wider HPDIs leading to somewhat larger RMSEs (compare last column of Table 1 with Table 7). Migration rate estimates show no increase in bias but their HPDIs are larger (compare Tables 8 and 9). Finally, the quality of the assignments is barely influenced by a decrease in the sample sizes (compare Tables 1 and 7). The effect of increasing the number of loci scored from 10 to 20 does not have an effect on model determination, estimates of regression parameters, and migration rates when the

TABLE 4

Migration estimates for various levels of genetic differentiation and high migration rate

Migration rate	True value	F_{ST}			
		0.01	0.05	0.10	0.25
m_{11}	0.700	0.355 ^a	0.663	0.702	0.710
m_{12}	0.023	0.173	0.036	0.021	0.020
m_{13}	0.003	0.222 ^a	0.007	0.002	0.002
m_{14}	0.274	0.250	0.293	0.275	0.269
m_{21}	0.018	0.184	0.027	0.021	0.020
m_{22}	0.700	0.359 ^a	0.656	0.720	0.710
m_{23}	0.227	0.252	0.240	0.213	0.220
m_{24}	0.055	0.205	0.076	0.047	0.050
m_{31}	0.259	0.224	0.291	0.273	0.256
m_{32}	0.028	0.252 ^a	0.029	0.027	0.025
m_{33}	0.700	0.336 ^a	0.662	0.689	0.704
m_{34}	0.013	0.188 ^a	0.020	0.010	0.014
m_{41}	0.149	0.243	0.151	0.136	0.142
m_{42}	0.081	0.211	0.111	0.085	0.079
m_{43}	0.070	0.185	0.074	0.061	0.064
m_{44}	0.700	0.362 ^a	0.664	0.719	0.715

Posterior estimates averaged across analyses of 10 simulated data sets with proportions of nonmigrant alleles set to $m_{ii} = 0.70$ are shown.

^aThe 95% HPDI does not contain the true value.

level of genetic differentiation is moderate ($F_{ST} = 0.1$) (results not shown). The only result that changes is the proportion of individuals that are misassigned, which decreases from 0.002 to 0. We also carried out analysis of a

TABLE 5

Migration estimates for various levels of genetic differentiation and high migration rate

Migration rate	True value	F_{ST}			
		0.01	0.05	0.10	0.25
m_{11}	0.700	0.407 ^a	0.867	0.910	0.910
m_{12}	0.023	0.232	0.014	0.006	0.006
m_{13}	0.003	0.182	0.004	0.001	0.001
m_{14}	0.274	0.179	0.115	0.083	0.084
m_{21}	0.018	0.209	0.008	0.005	0.005
m_{22}	0.700	0.428 ^a	0.890	0.901	0.905
m_{23}	0.227	0.197	0.088	0.078	0.075
m_{24}	0.055	0.165	0.014	0.016	0.014
m_{31}	0.259	0.244	0.098	0.087	0.086
m_{32}	0.028	0.170	0.009	0.009	0.010
m_{33}	0.700	0.449 ^a	0.882	0.901	0.900
m_{34}	0.013	0.137	0.012	0.004	0.005
m_{41}	0.149	0.165	0.046	0.042	0.043
m_{42}	0.081	0.181	0.033	0.025	0.023
m_{43}	0.070	0.181	0.056	0.020	0.019
m_{44}	0.700	0.473 ^a	0.865	0.913	0.915

Posterior estimates averaged across analyses of 10 simulated data sets with proportions of nonmigrant alleles set to $m_{ii} = 0.90$ are shown.

^aThe corresponding 95% HPDI does not contain the true value.

TABLE 6

Posterior model estimates when testing for nonexplanatory factors

Factors included	F_{ST}			
	0.01	0.05	0.10	0.25
None	0.636	0.533	0.512	0.505
G_1	0.127	0.224	0.243	0.241
G_2	0.184	0.157	0.151	0.158
G_1 and G_2	0.039	0.065	0.071	0.076
With interaction	0.013	0.022	0.022	0.021

Assignments	F_{ST}			
	0.01	0.05	0.10	0.25
Misassignments	0.741	0.282	0.110	0.002
Probabilities	0.218	0.694	0.881	0.996

Posterior model probabilities and assignment accuracy when varying the level of genetic differentiation F_{ST} and testing for two nonexplanatory factors (*i.e.*, different from the ones we used for generating migration rates) are shown. Data were simulated with proportions of nonmigrant alleles set to $m_{ii} = 0.70$.

scenario with $F_{ST} = 0.05$ and in this case, increasing the number of loci from 10 to 20 decreased the width of the HPDIs for migration rate estimates and improved the accuracy of individual assignments.

TABLE 7

Model estimates when sampling 50 individuals per population

Factors included	Posterior probabilities	
None	0.021	
G_1	0.021	
G_2	0.074	
G_1 and G_2	0.651	
With interaction	0.233	

Parameter	True value	Estimate/RMSE/ 95%HPDI
α_1	-0.900	-0.960 0.008 (-1.608; -0.322)
α_2	1.100	1.208 0.014 (0.536; 1.909)
σ^2	-	0.349 (0.108; 1.922)

Assignments	
Misassignments	0.008
Probabilities	0.989

Posterior model probabilities, regression parameter mean estimates, and assignment accuracy for synthetic data generated with proportions of nonmigrant alleles set to $m_{ii} = 0.70$ and level of genetic differentiation set to $F_{ST} = 0.25$ are shown.

TABLE 8

Migration estimates when sampling 50 individuals per population

Parameter	True value	Posterior estimate	RMSE	95% HPDI
m_{11}	0.700	0.702	<0.001	(0.612; 0.790)
m_{12}	0.023	0.026	0.043	(0.003; 0.056)
m_{13}	0.003	0.002	0.070	(0.000; 0.010)
m_{14}	0.274	0.270	0.001	(0.185; 0.356)
m_{21}	0.018	0.011	0.149	(0.000; 0.029)
m_{22}	0.700	0.722	0.001	(0.632; 0.808)
m_{23}	0.227	0.220	0.001	(0.142; 0.301)
m_{24}	0.055	0.047	0.025	(0.013; 0.086)
m_{31}	0.259	0.238	0.007	(0.158; 0.322)
m_{32}	0.028	0.026	0.030	(0.003; 0.055)
m_{33}	0.700	0.725	0.001	(0.637; 0.811)
m_{34}	0.013	0.011	0.033	(0.000; 0.028)
m_{41}	0.149	0.136	0.008	(0.074; 0.201)
m_{42}	0.081	0.075	0.006	(0.031; 0.124)
m_{43}	0.070	0.067	0.001	(0.025; 0.115)
m_{44}	0.700	0.721	0.001	(0.634; 0.808)

Estimates based on the posterior mean, RMSE, and 95% HPDI are reported for synthetic data generated with proportions of nonmigrant alleles set to $m_{ii} = 0.70$ and the level of genetic differentiation set to $F_{ST} = 0.25$.

APPLICATION TO REAL DATA

We use the human genome diversity cell line panel—Centre d’Etude du Polymorphisme Humain (HGDP-CEPH) presented by CANN *et al.* (2002) to illustrate how our method can be used to make inferences about the

TABLE 9

Migration rate estimates when sampling 100 individuals per population

Parameter	True value	Posterior estimate	RMSE	95% HPDI
m_{11}	0.700	0.710	<0.001	(0.647; 0.772)
m_{12}	0.023	0.020	0.018	(0.004; 0.038)
m_{13}	0.003	0.002	0.218	(0.000; 0.006)
m_{14}	0.274	0.269	<0.001	(0.208; 0.323)
m_{21}	0.018	0.020	0.014	(0.005; 0.038)
m_{22}	0.700	0.710	<0.001	(0.647; 0.772)
m_{23}	0.227	0.220	0.001	(0.165; 0.277)
m_{24}	0.055	0.050	0.009	(0.023; 0.079)
m_{31}	0.259	0.256	<0.001	(0.198; 0.317)
m_{32}	0.028	0.025	0.009	(0.007; 0.046)
m_{33}	0.700	0.704	<0.001	(0.640; 0.766)
m_{34}	0.013	0.014	0.013	(0.002; 0.029)
m_{41}	0.149	0.142	0.002	(0.097; 0.190)
m_{42}	0.081	0.079	0.001	(0.045; 0.115)
m_{43}	0.070	0.064	0.007	(0.033; 0.098)
m_{44}	0.700	0.715	<0.001	(0.652; 0.777)

Estimates based on the posterior mean, RMSE, and 95% HPDI are reported for synthetic data generated with proportions of nonmigrant alleles set to $m_{ii} = 0.70$ and the level of genetic differentiation set to $F_{ST} = 0.25$.

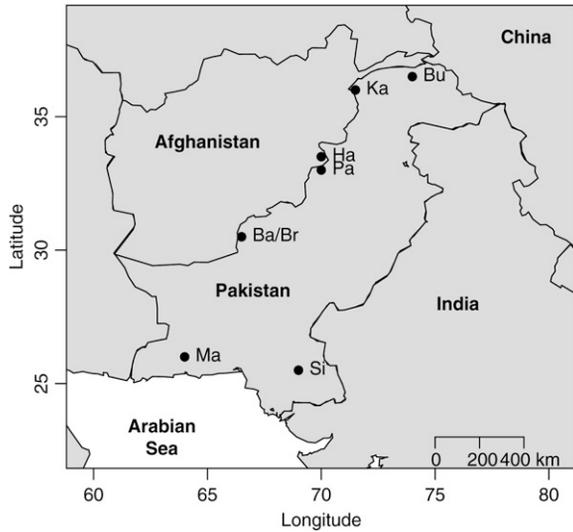


FIGURE 2.—Geographic locations of sampled populations. Solid circles represent centers of gravity of sampled areas of Pakistan. Abbreviations for population names are as follows: Ba, Balochi; Br, Brahui; Bu, Burusho; Ha, Hazara; Ka, Kalash; Ma, Makrani; Pa, Pathan; Si, Sindhi.

factors influencing migration patterns. In our example we selected a subset of eight populations, all from Pakistan (see Figure 2), corresponding to 200 individuals (25 per population). We grouped together the Balochi and Brahui samples because the STRUCTURE analyses carried out by ROSENBERG *et al.* (2002) place them in the same genetic cluster (see their Figure 2). Also, instead of using all 377 loci we did a first screening using an improved version of BEAUMONT and BALDING’s (2004) method to identify outlier loci that could be influenced by selection. On the basis of this screening we selected a total of 247 loci that were used in the analysis.

The effect of distance is supposed to be one of the main factors in determining gene flow in many species, but other factors such as altitude can influence geographic isolation and, therefore, migration patterns. We use our method to evaluate the relative importance of these two factors. We obtained pairwise geographic distances from latitude and longitude coordinates and also calculated the difference in altitude between each focal population and all other populations. CANN *et al.* (2002) give the geographic coordinates of each population as sample intervals; thus we used the gravity center of the area for the calculation of geographic distances between populations. With two parameters we can define five alternative models, which are presented in Table 10.

As was the case for the simulation study we used short pilot runs to tune up the proposal distributions to achieve reasonable acceptance ratios. To ensure convergence we increased the burn-in to 10^6 and the sample size to 20,000 and used a thinning interval of 50 iterations.

Some of the population-specific F_{ST} values are <0.01 (see Table 11), the level of genetic differentiation that our simulation study identified as problematic for the

TABLE 10

Posterior model probabilities for the Pakistani human data set

Factors included	Pr(\mathcal{M})	Estimate/95% HPDI		
		α_1	α_2	α_3
None	0.064			
Distance	0.093	-0.756 (-2.88; 0.417)		
Altitude	0.550		-1.74 (-3.35; 0.338)	
Distance and altitude	0.232	-0.469 (-2.95; 0.639)	-1.55 (-2.91; 0.747)	
With interaction	0.061	-0.524 (-2.66; 0.71)	-1.64 (3.4; 0.599)	0.302 (-1.2; 1.44)

Posterior model probabilities for the human data set when considering geographic distances and differences in altitudes as environmental factors are shown. Posterior estimates for regression parameters are based on the mode and 95% HPDI. The maximum *a posteriori* estimate of σ^2 is 0.657 with the 95% HPDI ranging from 0.089 to 11.1.

estimation of parameters. This example, therefore, provides us with an opportunity to illustrate the problems that may arise when our method (or any other MCMC-based method) is used in scenarios with weak genetic differentiation. In these situations, it is necessary to run many independent replicates and compare their results; in the present case we used 10 runs. In 6 of

TABLE 11

Local population F_{ST} ’s and inbreeding coefficient for Pakistani populations

Population	Mean/mode/95% HPDI	
	F_{ST}	F
Ba/Br	0.010	0.091
	0.010	0.099
	(0.006; 0.015)	(0.042; 0.132)
Bu	0.009	0.010
	0.009	0.009
	(0.005; 0.013)	(10^{-5} ; 0.033)
Ha	0.015	0.0155
	0.015	0.0156
	(0.010; 0.020)	(10^{-5} ; 0.040)
Ka	0.048	0.014
	0.049	0.0133
	(0.040; 0.058)	(10^{-6} ; 0.037)
Ma	0.007	0.116
	0.007	0.105
	(0.001; 0.016)	(0.065; 0.195)
Pa	0.009	0.0917
	0.008	0.0904
	(0.003; 0.015)	(0.046; 0.142)
Si	0.017	0.0579
	0.016	0.0586
	(0.009; 0.028)	(0.005; 0.125)

Estimates are based on the posterior mean and mode.

TABLE 12
Migration rates between Pakistani populations

From/into	Mean/mode/95% HPDI						
	Ba/Br	Bu	Ha	Ka	Ma	Pa	Si
Ba/Br	0.690 0.670 (0.427; 0.900)	0.000 0.000 (10 ⁻¹⁵ ; 10 ⁻⁸)	0.000 0.000 (10 ⁻⁹ ; 10 ⁻⁷)	0.000 0.000 (10 ⁻¹⁰ ; 10 ⁻⁷)	0.220 0.150 (0.008; 0.638)	0.060 0.020 (10 ⁻⁸ ; 0.339)	0.280 0.300 (0.021; 0.668)
Bu	0.000 0.000 (0.004; 0.086)	1.000 1.000 (1.000; 1.000)	0.000 0.000 (10 ⁻⁹ ; 10 ⁻⁷)	0.000 0.000 (10 ⁻¹⁰ ; 10 ⁻⁷)	0.010 0.000 (0.001; 0.190)	0.01 0.000 (10 ⁻⁴ ; 0.207)	0.010 0.000 (10 ⁻¹² ; 0.189)
Ha	0.010 0.000 (0.001; 0.135)	0.000 0.000 (10 ⁻¹⁶ ; 10 ⁻⁸)	1.000 1.000 (1.000; 1.000)	0.000 0.000 (10 ⁻¹⁴ ; 10 ⁻⁷)	0.010 0.000 (10 ⁻¹² ; 0.168)	0.040 0.020 (0.010; 0.246)	0.010 0.000 (10 ⁻¹² ; 0.151)
Ka	0.000 0.000 (10 ⁻⁴ ; 0.073)	0.000 0.000 (10 ⁻¹⁰ ; 10 ⁻⁸)	0.000 0.000 (10 ⁻¹⁵ ; 10 ⁻⁷)	1.000 1.000 (1.000; 1.000)	0.000 0.000 (10 ⁻⁴ ; 0.079)	0.010 0.000 (10 ⁻⁹ ; 0.128)	0.000 0.000 (10 ⁻⁴ ; 0.132)
Ma	0.220 0.230 (0.014; 0.488)	0.000 0.000 (10 ⁻¹⁵ ; 10 ⁻⁸)	0.000 0.000 (10 ⁻¹⁵ ; 10 ⁻⁷)	0.000 0.000 (10 ⁻¹⁰ ; 10 ⁻⁷)	0.680 0.740 (0.233; 0.935)	0.060 0.010 (10 ⁻⁹ ; 0.451)	0.150 0.030 (10 ⁻⁹ ; 0.560)
Pa	0.010 0.000 (0.003; 0.143)	0.000 0.000 (10 ⁻¹⁵ ; 10 ⁻⁸)	0.000 0.000 (10 ⁻⁹ ; 10 ⁻⁷)	0.000 0.000 (10 ⁻⁹ ; 10 ⁻⁷)	0.030 0.000 (0.001; 0.302)	0.740 0.760 (0.362; 0.936)	0.020 0.000 (0.003; 0.251)
Si	0.070 0.060 (0.009; 0.225)	0.000 0.000 (10 ⁻¹⁵ ; 10 ⁻⁸)	0.000 0.000 (10 ⁻¹⁴ ; 10 ⁻⁷)	0.000 0.000 (10 ⁻⁹ ; 10 ⁻⁷)	0.050 0.030 (0.001; 0.265)	0.070 0.050 (10 ⁻⁹ ; 0.339)	0.530 0.520 (0.145; 0.874)

Estimates are based on posterior mean and mode.

them, the most probable model included altitude only and in all cases there was a posterior probability of at least 50%. The second most probable model included both factors. However, in 4 other runs two other models, one including distance only and the other including both distance and altitude, gave similar high posterior probabilities while the model including altitude only was ranked third. Given these results, we followed FAUBET *et al.* (2007) and chose the run with the lowest deviance for estimation purposes. The Bayesian deviance has been proposed as a measure of model fit by a number of authors (FAUBET *et al.* 2007 and references therein) and in our specific case we considered the assignment component of the total deviance, $D_{\text{assign}} = -2 \log \Pr(M | S; m)$. Table 10 presents these results. The model with the highest posterior probability includes altitude only and the second most probable model includes both altitude and distance. In this latter case, the regression coefficients for the effects of altitude and distance are both negative, indicating that, as expected, both factors decrease migration rates between populations. Note, however, that the former seems to have a stronger effect (*i.e.*, larger absolute value).

Table 12 presents the mode and HPDI of migration rates between populations. Although the sum of maximum posterior estimates does not necessarily add up to one, we used them as estimators because of the inherent asymmetry of migration rate posterior distributions. There are three populations that do not receive migrants (Burusho, Bu; Hazara, Ha; and Kalash, Ka) and

they correspond to those located in high-altitude areas. Moreover, two of these populations (Bu and Ka) do not seem to send migrants either and a third one (Ha) seems to contribute very little to the gene pool of the Pathan. The population with the highest proportion of migrant genes is the Sindhi, which receives migrants mainly from Balochi/Brahui (Ba/Br). Three other populations have a somewhat lower proportion of migrant genes (Ba/Br; Makrani, Ma; and Pathan, Pa). In the case of Ba/Br, most of the genes come from Ma, and, conversely, most of the genes of Ma come from Ba/Br. Finally, the Pathan receive similar proportions of genes from Ba/Br, Ha, and Sindhi (Si). In general, there are frequent gene exchanges among southern populations while northern populations remain fairly isolated. The best explanation for this migration pattern is altitude differences with the most isolated populations being at high altitude and the least isolated ones at low altitude.

Finally, the mean and mode of inbreeding coefficient estimates are somewhat large when compared to F_{ST} estimates but this is not the case if we compare the lower bounds of the HPDIs (Table 11). Still, there are three local populations (Ba/Br, Ma, and Pa) for which the lower bound of F_{IS} HPDIs is >0.04 while that of F_{ST} 's is much lower. A potential explanation for this result could be that samples were taken from adult individuals and, therefore, the data set does not fit model assumptions concerning the moment at which sampling takes place. However, we do not have information concerning the age group involved in the sampling.

DISCUSSION

We present a new method for the estimation of recent migration rates that also allows for making inferences about the factors that influence gene flow in subdivided populations. It focuses on the F_1 descendants of migrant individuals and, therefore, estimates the probability that a given individual migrated during the previous generation. Our approach also estimates various other population-specific parameters such as local F_{ST} , inbreeding coefficients, and allele frequencies. The method requires data from codominant markers such as RFLPs, microsatellites, allozymes, and SNPs and environmental data specific to each local population. Note, however, that the modeling of dispersal barriers (mountains, roads, deforested areas, etc.) between pairs of populations can be introduced by considering landscape resistance measures usually used by landscape ecologists (see, *e.g.*, McRAE 2006).

We generated synthetic data following the inference model described above to investigate the effect of varying levels of genetic differentiation, proportions of non-migrant genes, and numbers of populations, loci, and individuals. The results of this simulation study indicate that the method can provide reliable estimates when global F_{ST} values are $>1\%$, the number of loci is only 10, and sample sizes are of the order of 50 individuals per population. Additionally, the identification of the environmental factors influencing migration is easier when migration rates are high and the number of local populations considered increases. We did not investigate the effect of varying the degree of polymorphism (*i.e.*, the number of allelic classes) or the effect of unsampled populations. We expect that increasing polymorphism will increase accuracy while the effect of unobserved populations is more likely to decrease it depending on true migration rates between unsampled and sampled populations. Our simulation study could be extended to take into account these considerations. Additionally, it would be desirable to consider demographic scenarios that differ from the one assumed by the inference model to test the robustness of our method.

We applied our method to a previously published microsatellite human data set for which local F_{ST} 's are within the range of values that our simulation study identified as problematic for parameter estimation. As expected, we observed convergence problems for this application and followed the approach of FAUBET *et al.* (2007) to minimize them (see previous section for a more detailed explanation). We found that altitude influences recent migration among Pakistani populations and that gene exchanges are more frequent in the south than in the north of Pakistan. Geographic distance seems to have little effect on migration, a result that can be explained by the limited geographic scale considered and the fact that even in poorly developed areas there are many means of transportation that facilitate movement of humans. On

the other hand, altitude can represent an important barrier particularly in winter when populations at high altitude can remain isolated for long periods of time.

The estimation of migration rates has proved to be a very difficult task. Several methods exist for this purpose; some of them estimate long-term migration rates and are based on coalescent theory (*e.g.*, MIGRATE, BEERLI and FELSENSTEIN 2001) while others provide recent migration rate estimates and are based on multilocus genotype approaches (*e.g.*, BAYESASS, WILSON and RANNALA 2003). All recent methods for estimating migration rates rely on MCMC approaches and require one to pay special attention to convergence issues (FAUBET *et al.* 2007). This is particularly important when genetic differentiation among populations is weak. This caveat also applies to our method, and the human example we present illustrates how to deal with these problems.

Being a multilocus genotype approach, our method resembles in many respects BAYESASS. It is important to note, however, that this resemblance is only superficial because we do not assume the same sampling scheme and we allow for high migration rates. Indeed, as opposed to WILSON and RANNALA (2003) we assume that sampling takes place after reproduction and before migration. This was done to avoid the low migration rate restriction underlying their method and to allow migration rates to vary between 0 and 1. More specifically, WILSON and RANNALA's (2003) formulation provides estimates of migration rates restricted to the interval $(0, \frac{1}{3})$ and assumes that m is very small because to account for individuals with mixed ancestry (*i.e.*, individuals whose alleles come from two different populations) they need to consider individuals that arrived one generation before sampling takes place. Thus, they are forced to assume that at most half of an individual's alleles comes from another population. In our case, we do not have this restriction because after reproduction the alleles of a given individual can come from any population. Doing this, however, precludes us from distinguishing between first-generation and second-generation migrants. Nevertheless, we can consider cases where parents are migrants from two different populations while BAYESASS considers only a single migrant ancestor.

The information used by our estimation method is the gametic disequilibrium generated by migration, which increases as genetic differentiation among local populations increases. Indeed, limited migration is very effective in increasing differentiation of gamete types among the subpopulations by random genetic drift (OHTA 1982). The strength of this gametic disequilibrium can be measured through the genotype of migrant individuals (or descendants from recent migrants) or through the gamete haplotype frequencies. Clearly, the former corresponds to short-term migration while the latter corresponds to the effect of long-term migration. All this implies that if the long-term migration is very high, the signature left by recent migration events will be weak. In

the case of our method, the simulation study indicates that reliable estimates can be obtained when the effective number of migrants is less than five (*i.e.*, $F_{ST} \geq 0.05$). The gametic disequilibrium due to long-term migration can also lead to a deviation from the hypothesis of independence among loci used to derive the likelihood function. This is a problem shared by all the methods that estimate migration rates from multilocus genotype data. The potential biases that could be introduced due to this problem require a very detailed simulation study, using an individually based model that produces synthetic data that allow for the estimation of gametic disequilibrium.

Another improvement introduced in our method is the use of the F-model first proposed by BALDING and NICHOLS (1995). This feature allows us to take into account the population admixture that may have taken place before the last generation of migration. Additionally, as pointed out by FALUSH *et al.* (2003), the implementation of this model permits identification of subtle population subdivisions and, therefore, improves the estimation of allele-frequency distributions when genetic differentiation is weak. This in turn improves the estimation of migration rates as shown by a pilot study comparing the performance of our method with and without the F-model (results not shown). All the improvements implemented by our method lead to good mixing properties of the MCMC and therefore minimize convergence problems. We stress, however, that users should always carefully check the convergence of the MCMC by running multiple analyses and comparing their results.

An important feature of our method is that besides simply estimating migration rates it also identifies the factors that influence them. We use the same approach as that first proposed by GAGGIOTTI *et al.* (2004), which consists of using a Dirichlet prior for the immigration rates and linking its shape parameters with the environmental data, using a generalized linear model. In the present case, however, we do not consider models without the constant factor (*i.e.*, the regression intercept). This was done because our experience with the application of this type of method (GAGGIOTTI *et al.* 2004; FOLL and GAGGIOTTI 2006) indicates that models excluding this parameter almost always had null posterior probabilities. These results can be explained by the fact that the regression intercept captures the effects of factors that act at a larger geographic scale than that considered for the metapopulation under study. It also takes into account behavioral characteristics of the species under study that remain the same regardless of the environment. In fact, the regression intercept influences only the variance of immigration rates, which increases as α_0 decreases. For example, we expect that the variance of the migration rate between two given populations will be larger for species that can disperse very long distances than for species with very poor dispersal abilities. In this case, then we expect to obtain estimates of the intercept that are smaller for the former.

In our approach we assumed that the probability of observing nonmigrant alleles in any given population is independent of environmental factors. The underlying rationale for this is that local environmental conditions will influence only emigration rates but do not have any effect on the immigration rates that are the focus of our estimation method. Ideally we would also like to estimate emigration rates. As WILSON and RANNALA (2003) point out, this could be done if we know local population sizes or, alternatively, if we could develop a method that can make use of temporal samples. However, such approaches are likely to involve much more complex likelihood functions that will necessarily lead to a worsening of convergence problems that are typical of complex methods that use MCMC approaches.

The software that implements the method incorporates features that facilitate the interpretation of results. For example, it provides estimates of both means and modes, which allows the user to choose the best parameter estimator depending on the shape of the posterior distribution (which is also provided by the software). Indeed, when posterior distributions are asymmetric, posterior estimates based on the mode and on the mean are rather different and the former provides a better way of describing the results. Thus, users should always have a look at the shape of posterior distributions to choose appropriate estimators.

Bayesian methods such as the one we present here are powerful tools for the study of natural populations. Users, however, should keep in mind that their application requires some expertise on the computational methods underlying their implementation, particularly on MCMC approaches. These issues are discussed more in detail in FAUBET *et al.* (2007) and also in the user manuals of several of the currently available methods. If these recommendations are followed, population biologists will be able to extract highly valuable information about the species under study.

Most of the computations presented in this article were performed on the cluster HealthPhy (CIMENT, Grenoble, France). We are grateful to Matthieu Foll for providing us the software to identify outlier loci in the human data set. We also thank Olivier Francois and two anonymous reviewers for their useful suggestions that helped to improve the manuscript. The software that implements the method is available for the three most popular operating systems at <http://www-leca.ujf-grenoble.fr/logiciels.htm>. This work was supported by the Fond National de la Science (grant ACI-Impbio-2004-42-ADGP). P.F. holds a Ph.D. studentship from the Ministère de la Recherche.

LITERATURE CITED

- BALDING, D. J., and R. A. NICHOLS, 1995 A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica* **96**: 3–12.
- BALDING, D. J., and R. A. NICHOLS, 1997 Significant genetic correlations among Caucasians at forensic DNA loci. *Heredity* **78**: 583–589.

- BEAUMONT, M. A., and D. J. BALDING, 2004 Identifying adaptive genetic divergence among populations from genome scans. *Mol. Ecol.* **13**: 969–980.
- 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.
- BROOKS, S. P., P. GIUDICI and G. O. ROBERTS, 2003 Efficient construction of reversible jump proposal distribution. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **65**: 3–55.
- CANN, H., C. TOMA, L. CAZES, M. LEGRAND, V. MOREL *et al.*, 2002 A human genome diversity cell line panel. *Science* **296**: 261–262.
- FALUSH, D., M. STEPHENS and J. K. PRITCHARD, 2003 Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**: 1567–1587.
- FAUBET, P., R. S. WAPLES and O. E. GAGGIOTTI, 2007 Evaluating the performance of a multilocus Bayesian method for the estimation of migration rates. *Mol. Ecol.* **16**: 1149–1166.
- FOLL, M., and O. E. GAGGIOTTI, 2006 Identifying the environmental factors that determine the genetic structure of populations. *Genetics* **174**: 875–891.
- FRANÇOIS, O., S. ANCELET and G. GUILLOT, 2006 Bayesian clustering using hidden Markov random fields in spatial population genetics. *Genetics* **174**: 805–816.
- GAGGIOTTI, O. E., F. JONES, W. M. LEE, W. AMOS, J. HARWOOD *et al.*, 2002 Patterns of colonization in a grey seal metapopulation. *Nature* **416**: 424–427.
- GAGGIOTTI, O. E., S. P. BROOKS, W. AMOS and J. HARWOOD, 2004 Combining demographic, environmental and genetic data to test hypotheses about colonization events in metapopulations. *Mol. Ecol.* **13**: 811–825.
- GELMAN, A., J. B. CARLIN, H. S. STERN and D. B. RUBIN, 1995 *Bayesian Data Analysis*. Chapman & Hall, London.
- GIORDANO, A. R., B. J. RIDENHOUR and A. STORFER, 2007 The influence of altitude and topography on genetic structure in the long-toed salamander (*Ambystoma macrodactylum*). *Mol. Ecol.* **16**: 1625–1637.
- GREEN, P. J., 1995 Reversible jump mcmc computation and bayesian model determination. *Biometrika* **82**: 711–732.
- GUILLOT, G., A. ESTOUP, F. MORTIER and J. F. COSSON, 2005 A spatial statistical model for landscape genetics. *Genetics* **170**: 1261–1280.
- MCRAE, B. H., 2006 Isolation by resistance. *Evolution* **60**: 1551–1561.
- OHTA, T., 1982 Linkage disequilibrium due to random genetic drift in finite subdivided populations. *Evolution* **79**: 1940–1944.
- ROBERT, C. P., 1994 *The Bayesian Choice: A Decision-Theoretic Motivation*. Springer, New York.
- ROSENBERG, N., J. PRITCHARD, J. WEBER, H. CANN, K. KIDD *et al.*, 2002 Genetic structure of human populations. *Science* **298**: 2981–2985.
- WILSON, G. A., and B. RANNALA, 2003 Bayesian inference of recent migration rates using multilocus genotypes. *Genetics* **163**: 1177–1191.

Communicating editor: L. EXCOFFIER

APPENDIX: PRIOR DISTRIBUTIONS FOR PARAMETERS

We take the following priors for each parameter discussed in the text.

Probability to observe nonmigrant genes: We assume that nonmigrant proportions are not influenced by environmental factors and therefore use a uniform distribution:

$$m_{ii} \sim \mathcal{U}(0, 1); \quad i.e., f(m_{ii}) = \begin{cases} 1 & \text{if } m_{ii} \in (0, 1) \\ 0 & \text{otherwise.} \end{cases} \quad (\text{A1})$$

Probability to observe migrant genes: We use a Dirichlet prior for the rate of migrant genes contributed by local populations other than the focal one, m_{ij}^* ,

$$\mathbf{m}_i^* | \boldsymbol{\psi}_i \sim \text{Dir}(\boldsymbol{\psi}_i); \quad i.e., f(\mathbf{m}_i^* | \boldsymbol{\psi}_i) = \Gamma\left(\sum_{j \neq i} \psi_{ij}\right) \prod_{j \neq i} \frac{m_{ij}^{\psi_{ij}-1}}{\Gamma(\psi_{ij})}, \quad (\text{A2})$$

where the m_{ij}^* 's are given by Equation 6.

Shape parameters for the Dirichlet prior: As ψ_{ij} 's must be positive we use a log-normal distribution,

$$\log \psi_{ij} | \boldsymbol{\alpha}, \sigma^2, \mathbf{G} \sim \mathcal{N}(\mu_{ij}, \sigma^2); \quad i.e., f(\psi_{ij} | \boldsymbol{\alpha}, \sigma^2, \mathbf{G}) = \frac{1}{\psi_{ij} \sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\log \psi_{ij} - \mu_{ij})^2}{2\sigma^2}\right), \quad (\text{A3})$$

where μ_{ij} is given by the regression (8).

Regression coefficients: We use a normal distribution,

$$\alpha \sim \mathcal{N}(0, \sigma_\alpha^2); \quad i.e., f(\alpha) = \frac{1}{\sqrt{2\pi\sigma_\alpha^2}} \exp\left(-\frac{\alpha^2}{2\sigma_\alpha^2}\right), \quad (\text{A4})$$

where $\sigma_\alpha^2 = 10$.

Deviation from the regression: We assume that σ^2 follows an inverse-gamma distribution,

$$\tau = \sigma^{-2} \sim \text{Gamma}(a_\tau, b_\tau); \quad i.e., f(\tau) = \frac{b_\tau^{a_\tau}}{\Gamma(a_\tau)} \tau^{a_\tau-1} \exp(-\tau b_\tau), \quad (\text{A5})$$

where $a_\tau, b_\tau = 1$

Local population F_{ST} 's: As θ_i 's must be positive we use a log-normal distribution,

$$\log \theta_i \sim \mathcal{N}(\omega, \xi); \quad i.e., f(\theta_i) = \frac{1}{\theta_i \sqrt{2\pi\xi}} \exp\left(-\frac{(\log \theta_i - \omega)^2}{2\xi}\right), \quad (\text{A6})$$

where $\omega = \xi = 1$.

Metapopulation allele frequencies: We use an uninformative Dirichlet prior,

$$\tilde{\mathbf{p}}_l \sim \text{Dir}(\lambda, \dots, \lambda); \quad i.e., f(\tilde{\mathbf{p}}_l) = \frac{\Gamma(k_l \lambda)}{\Gamma(\lambda)^{k_l}} \prod_{a=1}^{k_l} \tilde{p}_{la}^{\lambda-1}, \quad (\text{A7})$$

where k_l is the number of alleles observed at locus l in the metapopulation and $\lambda = 1$.

Population allele frequencies: We use a Dirichlet prior,

$$\mathbf{p}_{il} | \theta_i, \tilde{\mathbf{p}}_l \sim \text{Dir}(\theta_i \tilde{\mathbf{p}}_l); \quad i.e., f(\mathbf{p}_{il} | \theta_i, \tilde{\mathbf{p}}_l) = \Gamma(\theta_i) \prod_{a=1}^{k_l} \frac{\tilde{p}_{la}^{\theta_i \tilde{p}_{la} - 1}}{\Gamma(\theta_i \tilde{p}_{la})}. \quad (\text{A8})$$

Population-specific inbreeding coefficients: We use a uniform distribution,

$$F_i \sim \mathcal{U}(-1, 1); \quad i.e., f(F_i) = \begin{cases} \frac{1}{2} & \text{if } F_i \in (-1, 1) \\ 0 & \text{otherwise.} \end{cases} \quad (\text{A9})$$

Ancestry assignments: Following WILSON and RANNALA (2003), we use a multinomial prior,

$$\mathbf{M} | \mathbf{S}; \mathbf{m}_i \sim \text{Mult}(n_i, \mathbf{m}_i); \quad i.e., \Pr(\mathbf{M} | \mathbf{S}; \mathbf{m}_i) = n_i! \prod_{j \leq k} \frac{\tilde{m}_{ijk}^{n_{ijk}}}{n_{ijk}!}, \quad (\text{A10})$$

where n_{ijk} is the number of individuals sampled from population i that belongs to ancestry class (j, k) and \tilde{m}_{ijk} is given by Equation 1.