Inference of Population Structure Using Multilocus Genotype Data

Jonathan K. Pritchard, Matthew Stephens and Peter Donnelly

Department of Statistics, University of Oxford, Oxford OX1 3TG, United Kingdom

Manuscript received September 23, 1999
Accepted for publication February 18, 2000

ABSTRACT

We describe a model-based clustering method for using multilocus genotype data to infer population structure and assign individuals to populations. We assume a model in which there are K populations (where K may be unknown), each of which is characterized by a set of allele frequencies at each locus. Individuals in the sample are assigned (probabilistically) to populations, or jointly to two or more populations if their genotypes indicate that they are admixed. Our model does not assume a particular mutation process, and it can be applied to most of the commonly used genetic markers, provided that they are not closely linked. Applications of our method include demonstrating the presence of population structure, assigning individuals to populations, studying hybrid zones, and identifying migrants and admixed individuals. We show that the method can produce highly accurate assignments using modest numbers of loci—*e.g.*, seven microsatellite loci in an example using genotype data from an endangered bird species. The software used for this article is available from http://www.stats.ox.ac.uk/~pritch/home.html.

IN applications of population genetics, it is often useful to classify individuals in a sample into populations. In one scenario, the investigator begins with a sample of individuals and wants to say something about the properties of populations. For example, in studies of human evolution, the population is often considered to be the unit of interest, and a great deal of work has focused on learning about the evolutionary relationships of modern populations (*e.g.*, Cavalli *et al.* 1994). In a second scenario, the investigator begins with a set of predefined populations and wishes to classify individuals of unknown origin. This type of problem arises in many contexts (reviewed by Davies *et al.* 1999). A standard approach involves sampling DNA from members of a number of potential source populations and using these samples to estimate allele frequencies in each population at a series of unlinked loci. Using the estimated allele frequencies, it is then possible to compute the likelihood that a given genotype originated in each population. Individuals of unknown origin can be assigned to populations according to these likelihoods Paetkau *et al.* 1995; Rannala and Mountain 1997).

In both situations described above, a crucial first step is to define a set of populations. The definition of populations is typically subjective, based, for example, on linguistic, cultural, or physical characters, as well as the geographic location of sampled individuals. This subjective approach is usually a sensible way of incorporating diverse types of information. However, it may be difficult to know whether a given assignment of individuals to populations based on these subjective criteria represents a natural assignment in genetic terms, and it would be useful to be able to confirm that subjective classifications are consistent with genetic information and hence appropriate for studying the questions of interest. Further, there are situations where one is interested in "cryptic" population structure—*i.e.*, population structure that is difficult to detect using visible characters, but may be significant in genetic terms. For example, when association mapping is used to find disease genes, the presence of undetected population structure can lead to spurious associations and thus invalidate standard tests (Ewens and Spielman 1995). The problem of cryptic population structure also arises in the context of DNA fingerprinting for forensics, where it is important to assess the degree of population structure to estimate the probability of false matches (Balding and Nichols 1994, 1995; Foreman *et al.* 1997; Roeder *et al.* 1998).

Pritchard and Rosenberg (1999) considered how genetic information might be used to detect the presence of cryptic population structure in the association mapping context. More generally, one would like to be able to identify the actual subpopulations and assign individuals (probabilistically) to these populations. In this article we use a Bayesian clustering approach to tackle this problem. We assume a model in which there are K populations (where K may be unknown), each of which is characterized by a set of allele frequencies at each locus. Our method attempts to assign individuals to populations on the basis of their genotypes, while simultaneously estimating population allele frequencies. The method can be applied to various types of markers [*e.g.*, microsatellites, restriction fragment length polymorphisms (RFLPs), or single nucleotide polymorphisms (SNPs)], but it assumes that the marker

Corresponding author: Jonathan Pritchard, Department of Statistics, University of Oxford, 1 S. Parks Rd., Oxford OX1 3TG, United Kingdom. E-mail: pritch@stats.ox.ac.uk

loci are unlinked and at linkage equilibrium with one another within populations. It also assumes Hardy-Weinberg equilibrium within populations. (We discuss these assumptions further in background on clustering methods and the discussion.)

Our approach is reminiscent of that taken by Smouse *et al.* (1990), who used the EM algorithm to learn about the contribution of different breeding populations to a sample of salmon collected in the open ocean. It is also closely related to the methods of Foreman *et al.* (1997) and Roeder *et al.* (1998), who were concerned with estimating the degree of cryptic population structure to assess the probability of obtaining a false match at DNA fingerprint loci. Consequently they focused on estimating the amount of genetic differentiation among the unobserved populations. In contrast, our primary interest lies in the assignment of individuals to populations. Our approach also differs in that it allows for the presence of admixed individuals in the sample, whose genetic makeup is drawn from more than one of the K populations.

In the next section we provide a brief description of clustering methods in general and describe some advantages of the model-based approach we take. The details of the models and algorithms used are given in models and methods. We illustrate our method with several examples in applications to data: both on simulated data and on sets of genotype data from an endangered bird species and from humans. incorporating population information describes how our method can be extended to incorporate geographic information into the inference process. This may be useful for testing whether particular individuals are migrants or to assist in classifying individuals of unknown origin (as in Rannala and Mountain 1997, for example). Background on the computational methods used in this article is provided in the appendix.

BACKGROUND ON CLUSTERING METHODS

Consider a situation where we have genetic data from a sample of individuals, each of whom is assumed to have originated from a *single* unknown population (no admixture). Suppose we wish to cluster together individuals who are genetically similar, identify distinct clusters, and perhaps see how these clusters relate to geographical or phenotypic data on the individuals. There are broadly two types of clustering methods we might use:

1. *Distance-based methods.* These proceed by calculating a pairwise distance matrix, whose entries give the distance (suitably defined) between every pair of individuals. This matrix may then be represented using some convenient graphical representation (such as a tree or a multidimensional scaling plot) and clusters may be identified by eye.
2. *Model-based methods.* These proceed by assuming that observations from each cluster are random draws from some parametric model. Inference for the parameters corresponding to each cluster is then done jointly with inference for the cluster membership of each individual, using standard statistical methods (for example, maximum-likelihood or Bayesian methods).

Distance-based methods are usually easy to apply and are often visually appealing. In the genetics literature, it has been common to adapt distance-based phylogenetic algorithms, such as neighbor-joining, to clustering multilocus genotype data (*e.g.*, Bowcock *et al.* 1994). However, these methods suffer from many disadvantages: the clusters identified may be heavily dependent on both the distance measure and graphical representation chosen; it is difficult to assess how confident we should be that the clusters obtained in this way are meaningful; and it is difficult to incorporate additional information such as the geographic sampling locations of individuals. Distance-based methods are thus more suited to exploratory data analysis than to fine statistical inference, and we have chosen to take a model-based approach here.

The first challenge when applying model-based methods is to specify a suitable model for observations from each cluster. To make our discussion more concrete we introduce very briefly some of our model and notation here; a fuller treatment is given later. Assume that each cluster (population) is modeled by a characteristic set of allele frequencies. Let X denote the genotypes of the sampled individuals, Z denote the (unknown) populations of origin of the individuals, and P denote the (unknown) allele frequencies in all populations. (Note that X, Z, and P actually represent multidimensional vectors.) Our main modeling assumptions are Hardy-Weinberg equilibrium within populations and complete linkage equilibrium between loci within populations. Under these assumptions each allele at each locus in each genotype is an independent draw from the appropriate frequency distribution, and this completely specifies the probability distribution $\Pr(X|Z, P)$ (given later in Equation 2). Loosely speaking, the idea here is that the model accounts for the presence of Hardy-Weinberg or linkage disequilibrium by introducing population structure and attempts to find population groupings that (as far as possible) are not in disequilibrium. While inference may depend heavily on these modeling assumptions, we feel that it is easier to assess the validity of explicit modeling assumptions than to compare the relative merits of more abstract quantities such as distance measures and graphical representations. In situations where these assumptions are deemed unreasonable then alternative models should be built.

Having specified our model, we must decide how to perform inference for the quantities of interest (Z and P). Here, we have chosen to adopt a Bayesian approach,

by specifying models (priors) Pr(Z) and Pr(P), for both Z and P. The Bayesian approach provides a coherent framework for incorporating the inherent uncertainty of parameter estimates into the inference procedure and for evaluating the strength of evidence for the inferred clustering. It also eases the incorporation of various sorts of prior information that may be available, such as information about the geographic sampling location of individuals.

Having observed the genotypes, X, our knowledge about Z and P is then given by the posterior distribution

$$\Pr(Z, P|X) \propto \Pr(Z)\Pr(P)\Pr(X|Z, P). \quad (1)$$

While it is not usually possible to compute this distribution exactly, it *is* possible to obtain an approximate sample $(Z^{(1)}, P^{(1)}), (Z^{(2)}, P^{(2)}), \ldots, (Z^{(M)}, P^{(M)})$ from Pr(Z, $P|X$) using Markov chain Monte Carlo (MCMC) methods described below (see Gilks *et al.* 1996b, for more general background). Inference for Z and P may then be based on summary statistics obtained from this sample (see *Inference for Z, P, and Q* below). A brief introduction to MCMC methods and Gibbs sampling may be found in the appendix.

MODELS AND METHODS

We now provide a more detailed description of our modeling assumptions and the algorithms used to perform inference, beginning with the simpler case where each individual is assumed to have originated in a single population (no admixture).

The model without admixture: Suppose we genotype N diploid individuals at L loci. In the case without admixture, each individual is assumed to originate in one of K populations, each with its own characteristic set of allele frequencies. Let the vector X denote the observed genotypes, Z the (unknown) populations of origin of the individuals, and P the (unknown) allele frequencies in the populations. These vectors consist of the following elements,

$(x_l^{\{i,1\}}, x_l^{\{i,2\}})$ = genotype of the ith individual at the lth locus, where $i = 1, 2, \ldots, N$ and $l = 1, 2, \ldots, L$;
$z^{(i)}$ = population from which individual i originated;
p_{klj} = frequency of allele j at locus l in population k, where $k = 1, 2, \ldots, K$ and $j = 1, 2, \ldots, J_l$,

where J_l is the number of distinct alleles observed at locus l, and these alleles are labeled $1, 2, \ldots, J_l$.

Given the population of origin of each individual, the genotypes are assumed to be generated by drawing alleles independently from the appropriate population frequency distributions,

$$\Pr(x_l^{\{i,a\}} = j|Z, P) = p_{z(i)lj} \quad (2)$$

independently for each $x_l^{\{i,a\}}$. (Note that $p_{z(i)lj}$ is the frequency of allele j at locus l in the population of origin of individual i.)

Assume that before observing the genotypes we have no information about the population of origin of each individual and that the probability that individual i originated in population k is the same for all k,

$$\Pr(z^{(i)} = k) = 1/K, \quad (3)$$

independently for all individuals. (In cases where some populations may be more heavily represented in the sample than others, this assumption is inappropriate; it would be straightforward to extend our model to deal with such situations.)

We follow the suggestion of Balding and Nichols (1995) (see also Foreman *et al.* 1997 and Rannala and Mountain 1997) in using the Dirichlet distribution to model the allele frequencies at each locus within each population. The Dirichlet distribution $\mathcal{D}(\lambda_1, \lambda_2, \ldots, \lambda_J)$ is a distribution on allele frequencies $p = (p_1, p_2, \ldots, p_J)$ with the property that these frequencies sum to 1. We use this distribution to specify the probability of a particular set of allele frequencies $p_{kl\cdot}$ for population k at locus l,

$$p_{kl\cdot} \sim \mathcal{D}(\lambda_1, \lambda_2, \ldots, \lambda_{J_l}), \quad (4)$$

independently for each k,l. The expected frequency of allele j is proportional to λ_j, and the variance of this frequency decreases as the sum of the λ_j increases. We take $\lambda_1 = \lambda_2 = \cdots = \lambda_{J_l} = 1.0$, which gives a uniform distribution on the allele frequencies; alternatives are discussed in the discussion.

MCMC algorithm (without admixture): Equations 2, 3, and 4 define the quantities Pr($X|Z$, P), Pr(Z), and Pr(P), respectively. By setting $\theta = (\theta_1, \theta_2) = (Z, P)$ and letting $\pi(Z, P) = \Pr(Z, P|X)$ we can use the approach outlined in *Algorithm A1* to construct a Markov chain with stationary distribution Pr(Z, $P|X$) as follows:

Algorithm 1: *Starting with initial values $Z^{(0)}$ for Z (by drawing $Z^{(0)}$ at random using* (3) *for example), iterate the following steps for $m = 1, 2, \ldots$.*

Step 1. *Sample $P^{(m)}$ from* Pr($P|X$, $Z^{(m-1)}$).
Step 2. *Sample $Z^{(m)}$ from* Pr($Z|X$, $P^{(m)}$).

Informally, step 1 corresponds to estimating the allele frequencies for each population assuming that the population of origin of each individual is known; step 2 corresponds to estimating the population of origin of each individual, assuming that the population allele frequencies are known. For sufficiently large m and c, $(Z^{(m)}, P^{(m)})$, $(Z^{(m+c)}, P^{(m+c)})$, $(Z^{(m+2c)}, P^{(m+2c)})$, ... will be approximately independent random samples from Pr(Z, $P|X$). The distributions required to perform each step are given in the appendix.

The model with admixture: We now expand our model to allow for admixed individuals by introducing a vector Q to denote the admixture proportions for each individual. The elements of Q are

$q_k^{(i)}$ = proportion of individual i's genome that originated from population k.

It is also necessary to modify the vector Z to replace the assumption that each individual i originated in some unknown population $z^{(i)}$ with the assumption that each observed allele copy $x_l^{(i,a)}$ originated in some unknown population $z_l^{(i,a)}$:

$z_l^{(i,a)}$ = population of origin of allele copy $x_l^{(i,a)}$.

We use the term "allele copy" to refer to an allele carried at a particular locus by a particular individual.

Our primary interest now lies in estimating Q. We proceed in a manner similar to the case without admixture, beginning by specifying a probability model for (X, Z, P, Q). Analogues of (2) and (3) are

$$\Pr(x_l^{(i,a)} = j|Z, P, Q) = p_{z_l^{(i,a)}lj} \tag{5}$$

and

$$\Pr(z_l^{(i,a)} = k|P, Q) = q_k^{(i)}, \tag{6}$$

with (4) being used to model P as before. To complete our model we need to specify a distribution for Q, which in general will depend on the type and amount of admixture we expect to see. Here we model the admixture proportions $q^{(i)} = (q_1^{(i)}, \ldots, q_K^{(i)})$ of individual i using the Dirichlet distribution

$$q^{(i)} \sim \mathcal{D}(\alpha, \alpha, \ldots, \alpha) \tag{7}$$

independently for each individual. For large values of α ($\gg 1$), this models each individual as having allele copies originating from all K populations in equal proportions. For very small values of α ($\ll 1$), it models each individual as originating mostly from a single population, with each population being equally likely. As $\alpha \to 0$ this model becomes the same as our model without admixture (although the implementation of the MCMC algorithm is somewhat different). We allow α to range from 0.0 to 10.0 and attempt to learn about α from the data (specifically we put a uniform prior on $\alpha \in [0, 10]$ and use a Metropolis-Hastings update step to integrate out our uncertainty in α). This model may be considered suitable for situations where little is known about admixture; alternatives are discussed in the discussion.

MCMC algorithm (with admixture): The following algorithm may be used to sample from $\Pr(Z, P, Q|X)$.

Algorithm 2: *Starting with initial values $Z^{(0)}$ for Z (by drawing $Z^{(0)}$ at random using* (3) *for example), iterate the following steps for $m = 1, 2, \ldots$.*

Step 1. *Sample $P^{(m)}$, $Q^{(m)}$ from* $\Pr(P, Q|X, Z^{(m-1)})$.
Step 2. *Sample $Z^{(m)}$ from* $\Pr(Z|X, P^{(m)}, Q^{(m)})$.
Step 3. *Update α using a Metropolis-Hastings step.*

Informally, step 1 corresponds to estimating the allele frequencies for each population and the admixture proportions of each individual, assuming that the population of origin of each allele copy in each individual is known; step 2 corresponds to estimating the population of origin of each allele copy, assuming that the population allele frequencies and the admixture proportions are known. As before, for sufficiently large m and c, $(Z^{(m)}, P^{(m)}, Q^{(m)})$, $(Z^{(m+c)}, P^{(m+c)}, Q^{(m+c)})$, $(Z^{(m+2c)}, P^{(m+2c)}, Q^{(m+2c)})$, ... will be approximately independent random samples from $\Pr(Z, P, Q|X)$. The distributions required to perform each step are given in the appendix.

Inference: *Inference for Z, P, and Q:* We now discuss how the MCMC output can be used to perform inference on Z, P, and Q. For simplicity, we focus our attention on Q; inference for Z or P is similar.

Having obtained a sample $Q^{(1)}, \ldots, Q^{(M)}$ (using suitably large burn-in m and thinning interval c) from the posterior distribution of $Q = (q_1, \ldots, q_N)$ given X using the MCMC method, it is desirable to summarize the information contained, perhaps by a point estimate of Q. A seemingly obvious estimate is the posterior mean

$$E(q_i|X) \approx \frac{1}{M}\sum_{m=1}^{M} q_i^{(m)}. \tag{8}$$

However, the symmetry of our model implies that the posterior mean of q_i is $(1/K, 1/K, \ldots, 1/K)$ for all i, whatever the value of X. For example, suppose that there are just two populations and 10 individuals and that the genotypes of these individuals contain strong information that the first 5 are in one population and the second 5 are in the other population. Then either

$$q_1 \ldots q_5 \approx (1, 0) \quad \text{and} \quad q_6 \ldots q_{10} \approx (0, 1) \tag{9}$$

or

$$q_1 \ldots q_5 \approx (0, 1) \quad \text{and} \quad q_6 \ldots q_{10} \approx (1, 0), \tag{10}$$

with these two "symmetric modes" being equally likely, leading to the expectation of any given q_i being (0.5, 0.5). This is essentially a problem of nonidentifiability caused by the symmetry of the model [see Stephens (2000b) for more discussion].

In general, if there are K populations then there will be $K!$ sets of symmetric modes. Typically, MCMC schemes find it rather difficult to move between such modes, and the algorithms we describe will usually explore only one of the symmetric modes, even when run for a very large number of iterations. Fortunately this does not bother us greatly, since from the point of view of clustering all the symmetric modes are the same [compare the clusterings corresponding to (9) and (10)]. If our sampler explores only one symmetric mode then the sample means (8) will be very poor estimates of the posterior means for the q_i, but will be much better estimates of the *modes* of the q_i, which in this case turn out to be a much better summary of the information in the data. Ironically then, the poor mixing of the MCMC sampler between the symmetric modes gives the asymptotically useless estimator (8) some practical

value. Where the MCMC sampler succeeds in moving between symmetric modes, or where it is desired to combine results from samples obtained using different starting points (which may involve combining results corresponding to different modes), more sophisticated methods [such as those described by Stephens (2000b)] may be required.

Inference for the number of populations: The problem of inferring the number of clusters, K, present in a data set is notoriously difficult. In the Bayesian paradigm the way to proceed is theoretically straightforward: place a prior distribution on K and base inference for K on the posterior distribution

$$\Pr(K|X) \propto \Pr(X|K)\Pr(K). \tag{11}$$

However, this posterior distribution can be peculiarly dependent on the modeling assumptions made, even where the posterior distributions of other quantities (Q, Z, and P, say) are relatively robust to these assumptions. Moreover, there are typically severe computational challenges in estimating $\Pr(X|K)$. We therefore describe an alternative approach, which is motivated by approximating (11) in an *ad hoc* and computationally convenient way.

Arguments given in the appendix (*Inference on K, the number of populations*) suggest estimating $\Pr(X|K)$ using

$$\Pr(X|K) \approx \exp(-\hat{\mu}/2 - \hat{\sigma}^2/8), \tag{12}$$

where

$$\hat{\mu} = \frac{1}{M}\sum_{m=1}^{M} -2 \log \Pr(X|Z^{(m)}, P^{(m)}, Q^{(m)}) \tag{13}$$

and

$$\hat{\sigma}^2 = \frac{1}{M}\sum_{m=1}^{M} (-2 \log \Pr(X|Z^{(m)}, P^{(m)}, Q^{(m)}) - \hat{\mu})^2. \tag{14}$$

We use (12) to estimate $\Pr(X|K)$ for each K and substitute these estimates into (11) to approximate the posterior distribution $\Pr(K|X)$.

In fact, the assumptions underlying (12) are dubious at best, and we do not claim (or believe) that our procedure provides a quantitatively accurate estimate of the posterior distribution of K. We see it merely as an *ad hoc* guide to which models are most consistent with the data, with the main justification being that it seems to give sensible answers in practice (see next section for examples). Notwithstanding this, for convenience we continue to refer to "estimating" $\Pr(K|X)$ and $\Pr(X|K)$.

APPLICATIONS TO DATA

We now illustrate the performance of our method on both simulated data and real data (from an endangered bird species and from humans). The analyses make use of the methods described in *The model with admixture.*

Simulated data: To test the performance of the clustering method in cases where the "answers" are known, we simulated data from three population models, using standard coalescent techniques (Hudson 1990). We assumed that sampled individuals were genotyped at a series of unlinked microsatellite loci. Data were simulated under the following models.

Model 1: A single random-mating population of constant size.

Model 2: Two random-mating populations of constant effective population size 2N. These were assumed to have split from a single ancestral population, also of size $2N$ at a time N generations in the past, with no subsequent migration.

Model 3: Admixture of populations. Two discrete populations of equal size, related as in model 2, were fused to produce a single random-mating population. Samples were collected after two generations of random mating in the merged population. Thus, individuals have i grandparents from population 1, and $4 - i$ grandparents from population 2 with probability $\binom{4}{i}/16$, where $i \in \{0, 4\}$. All loci were simulated independently.

We present results from analyzing data sets simulated under each model. Data set 1 was simulated under model 1, with 5 microsatellite loci. Data sets 2A and 2B were simulated under model 2, with 5 and 15 microsatellite loci, respectively. Data set 3 was simulated under model 3, with 60 loci (preliminary analyses with fewer loci showed this to be a much harder problem than models 1 and 2). Microsatellite mutation was modeled by a simple stepwise mutation process, with the mutation parameter $4N\mu$ set at 16.0 per locus (*i.e.*, the expected variance in repeat scores within populations was 8.0). We did not make use of the assumed mutation model in analyzing the simulated data.

Our analysis consists of two phases. First, we consider the issue of model choice—*i.e.*, how many populations are most appropriate for interpreting the data. Then, we examine the clustering of individuals for the inferred number of populations.

Choice of K for simulated data: For each model, we ran a series of independent runs of the Gibbs sampler for each value of K (the number of populations) between 1 and 5. The results presented are based on runs of 10^6 iterations or more, following a burn-in period of at least 30,000 iterations. To choose the length of the burn-in period, we printed out $\log(\Pr(X|P^{(m)}, Q^{(m)}))$, and several other summary statistics during the course of a series of trial runs, to estimate how long it took to reach (approximate) stationarity. To check for possible problems with mixing, we compared the estimates of $P(X|K)$ and other summary statistics obtained over several independent runs of the Gibbs sampler, starting from different initial points. In general, substantial differences between runs can indicate that either the runs should

TABLE 1

Estimated posterior probabilities of K, for simulated data sets 1, 2A, 2B, and 3 (denoted X_1, X_{2A}, X_{2B}, and X_3, respectively)

K	log $P(K\|X_1)$	$P(K\|X_{2A})$	$P(K\|X_{2B})$	$P(K\|X_3)$
1	~1.0	~0.0	~0.0	~0.0
2	~0.0	0.21	0.999	~1.0
3	~0.0	0.58	0.0009	~0.0
4	~0.0	0.21	~0.0	~0.0
5	~0.0	~0.0	~0.0	~0.0

The numbers should be regarded as a rough guide to which models are consistent with the data, rather than accurate estimates of posterior probabilities.

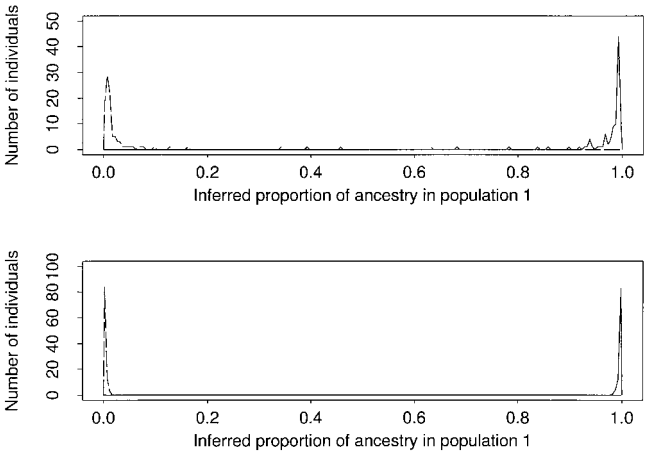

Figure 1.—Summary of the clustering results for simulated data sets 2A and 2B, respectively. For each individual, we computed the mean value of $q_1^{(i)}$ (the proportion of ancestry in population 1), over a single run of the Gibbs sampler. The dashed line is a histogram of mean values of $q_1^{(i)}$ for individuals from population 0; the solid line is for individuals from population 1.

be longer to obtain more accurate estimates or that independent runs are getting stuck in different modes in the parameter space. (Here, we consider the $K!$ modes that arise from the nonidentifiability of the K populations to be equivalent, since they arise from permuting the K population labels.)

We found that in most cases we obtained consistent estimates of $P(X|K)$ across independent runs. However, when analyzing data set 2A with $K = 3$, the Gibbs sampler found two different modes. This data set actually contains two populations, and when K is set to 3, one of the populations expands to fill two of the three clusters. It is somewhat arbitrary which of the two populations expands to fill the extra cluster: this leads to two modes of slightly different heights. The Gibbs sampler did not manage to move between the two modes in any of our runs.

In Table 1 we report estimates of the posterior probabilities of values of K, assuming a uniform prior on K between 1 and 5, obtained as described in *Inference for the number of populations*. We repeat the warning given there that these numbers should be regarded as rough guides to which models are consistent with the data, rather than accurate estimates of the posterior probabilities. In the case where we found two modes (data set 2A, $K = 3$), we present results based on the mode that gave the higher estimate of $\Pr(X|K)$.

With all four simulated data sets we were able to correctly infer whether or not there was population structure ($K = 1$ for data set 1 and $K > 1$ otherwise). In the case of data set 2A, which consisted of just 5 loci, there is not a clear estimate of K, as the posterior probability is consistent with both the correct value, $K = 2$, and also with $K = 3$ or 4. However, when the number of loci was increased to 15 (data set 2B), virtually all of the posterior probability was on the correct number of populations, $K = 2$.

Data set 3 was simulated under a more complicated model, where most individuals have mixed ancestry. In this case, the population was formed by admixture of two populations, so the "true" clustering is with $K = 2$, and Q estimating the number of grandparents from each of the two original populations, for each individual. Intuitively it seems that another plausible clustering would be with $K = 5$, individuals being assigned to clusters according to how many grandparents they have from each population. In biological terms, the solution with $K = 2$ is more natural and is indeed the inferred value of K for this data set using our *ad hoc* guide [the estimated value of $\Pr(X|K)$ was higher for $K = 5$ than for $K = 3$, 4, or 6, but much lower than for $K = 2$]. However, this raises an important point: the inferred value of K may not always have a clear biological interpretation (an issue that we return to in the discussion).

Clustering of simulated data: Having considered the problem of estimating the number of populations, we now examine the performance of the clustering algorithm in assigning particular individuals to the appropriate populations. In the case where the populations are discrete, the clustering performs very well (Figure 1), even with just 5 loci (data set 2A), and essentially perfectly with 15 loci (data set 2B).

The case with admixture (Figure 2) appears to be more difficult, even using many more loci. However, the clustering algorithm did manage to identify the population structure appropriately and estimated the ancestry of individuals with reasonable accuracy. Part of the reason that this problem is difficult is that it is hard to estimate the original allele frequencies (before admixture) when almost all the individuals (7/8) are admixed. A more fundamental problem is that it is difficult to get accurate estimates of $q^{(i)}$ for particular individuals because (as can be seen from the y-axis of Figure 2) for any given individual, the variance of how many of its alleles are *actually* derived from each population

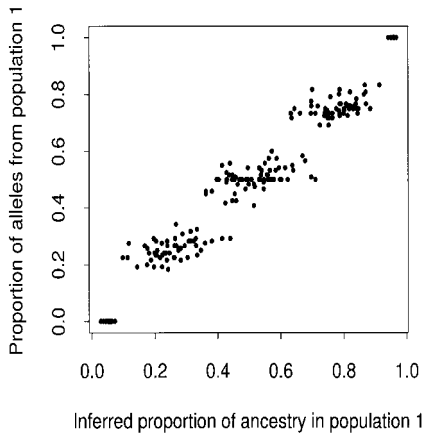

Figure 2.—Summary of the clustering results for simulated data set 3. Each point plots the estimated value of $q_1^{(i)}$ (the proportion of ancestry in population 1) for a particular individual against the fraction of their alleles that were actually derived from population 1 (across the 60 loci genotyped). The five clusters (from left to right) are for individuals with 0, 1, . . . , 4 grandparents in population 1, respectively.

can be substantial (for intermediate q). This property means that even if the allele frequencies were known, it would still be necessary to use a considerable number of loci to get accurate estimates of q for admixed individuals.

Data from the Taita thrush: We now present results from applying our method to genotype data from an endangered bird species, the Taita thrush, *Turdus helleri*. Individuals were sampled at four locations in southeast Kenya [Chawia (17 individuals), Ngangao (54), Mbololo (80), and Yale (4)]. Each individual was genotyped at seven microsatellite loci (Galbusera *et al.* 2000).

This data set is a useful test for our clustering method, because the geographic samples are likely to represent distinct populations. These locations represent fragments of indigenous cloud forest, separated from each other by human settlements and cultivated areas. Yale, which is a very small fragment, is quite close to Ngangao. Extensive data on ringed and radio-tagged birds over a 3-year period indicate low migration rates (Galbusera *et al.* 2000).

As discussed in background on clustering methods, it is currently common to use distance-based clustering methods to visualize genotype data of this kind. To permit a comparison between that type of approach and our own method, we begin by showing a neighbor-joining tree of the bird data (Figure 3). Inspection of the tree reveals that the Chawia and Mbololo individuals represent (somewhat) distinct clusters. Several individuals (marked by asterisks) appear to be classified with other groups. The four Yale individuals appear to fall within the Ngangao group [a view that is supported by summary statistics of divergence showing the Yale and Ngangao to be very closely related (Table 2)].

The tree illustrates several shortcomings of distance-

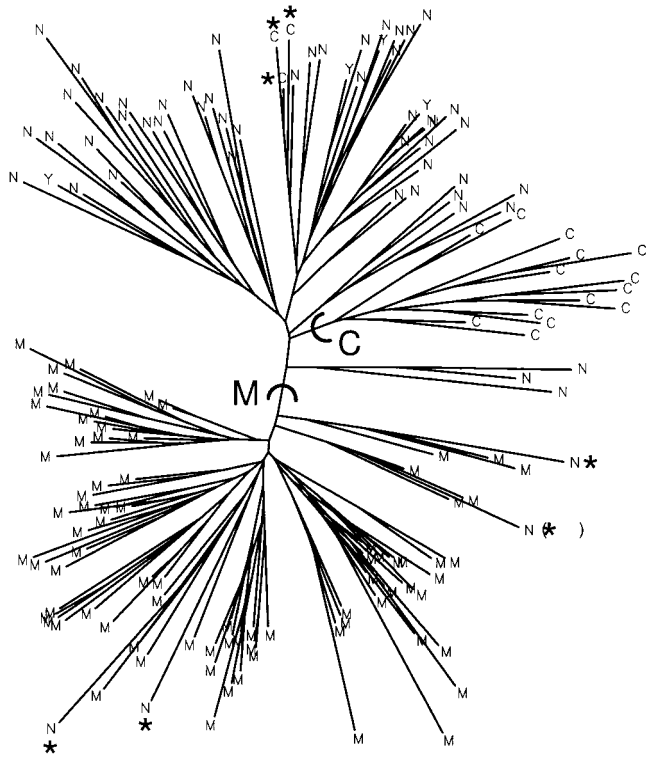

Figure 3.—Neighbor-joining tree of individuals in the *T. helleri* data set. Each tip represents a single individual. C, M, N, and Y indicate the populations of origin (Chawia, Mbololo, Ngangao, and Yale, respectively). Using the labels, it is possible to group the Chawia and Mbololo individuals into (somewhat) distinct clusters, as marked. However, it would not be possible to identify these clusters if the population labels were not available. Individuals who appear to be misclassified are marked *. One of these individuals [marked (*)] was also identified by our own algorithm as a possible migrant. The tree was constructed using the program *Neighbor* included in Phylip (Felsenstein 1993). The pairwise distance matrix was computed as follows (Mountain and Cavalli-Sforza 1997). For each pair of individuals, we added $1/L$ for each locus at which they had no alleles in common, $1/2L$ for each locus at which they had one allele in common (*e.g.*, AA:AB or AB:AC), and 0 for each locus at which they had two alleles in common (*e.g.*, AA:AA or AB:AB), where L is the number of loci compared.

TABLE 2

Summary statistics of variation within and between geographic groups

	Chawia	Mbololo	Ngangao	Yale
Chawia	5.1			
Mbololo	7.1	5.6		
Ngangao	3.1	1.6	5.5	
Yale	1.9	2.3	0.1	6.0

Diagonal, variance in repeat scores within groups; below diagonal, square of mean difference in repeat scores between populations [$(\delta\mu)^2$; Goldstein and Pollock 1997, Equation C3)].

based clustering methods. First, it would not be possible (in this case) to identify the appropriate clusters if the labels were missing. Second, since the tree does not use a formal probability model, it is difficult to ask statistical questions about features of the tree, for example: Are the individuals marked with asterisks actually migrants, or are they simply misclassified by chance? Is there evidence of population structure *within* the Ngangao group (which appears from the tree to be quite diverse)?

We now apply our clustering method to these data.

Choice of *K*, for Taita thrush data: To choose an appropriate value of K for modeling the data, we ran a series of independent runs of the Gibbs sampler at a range of values of K. After running numerous medium-length runs to investigate the behavior of the Gibbs sampler (using the diagnostics described in *Choice of K for simulated data*), we again chose to use a burn-in period of 30,000 iterations and to collect data for 10^6 iterations. We ran three to five independent simulations of this length for each K between 1 and 5 and found that the independent runs produced highly consistent results. At $K = 5$, a run of 10^6 steps takes ~70 min on our desktop machine.

Using the approach described in *Inference for the number of populations*, we estimated $\Pr(X|K)$ for $K = 1, 2, \ldots, 5$ and corresponding values of $\Pr(K|X)$ for a uniform prior on $K = 1, 2, \ldots, 5$. (In fact, this data set contains a lot of information about K, so that inference is relatively robust to choice of prior on K, and other priors, such as taking $\Pr(K)$ proportional to Poisson(1) for $K > 0$, would give virtually indistinguishable results.) From the estimates of $\Pr(K|X)$, shown in the last column of Table 3, it is clear that the models with $K = 1$ or 2 are completely insufficient to model the data and that the model with $K = 3$ is substantially better than models with larger K. Given these results, we now focus our subsequent analysis on the model with three populations.

Clustering results for Taita thrush data: Figure 4 shows a plot of the clustering results for the individuals in the sample, assuming that there are three populations (as inferred above). We did not use (and indeed, did not know) the sampling locations of individuals when we obtained these results. Our clustering algorithm seems to have performed very well, with just a few individuals (labeled 1–4) falling somewhat outside the obvious clusters. All of the points in the extreme corners (some of which may be difficult to resolve on the picture) are correctly assigned. The four Yale individuals were assigned to the Ngangao cluster, consistent with the neighbor-joining tree and the $(\delta\mu)^2$ distances. We return to this data set in incorporating population information to consider the question of whether the individuals that seem not to cluster tightly with others sampled from the same location are the product of migration.

TABLE 3

Inferring the value of *K*, the number of populations, for the *T. helleri* data

K	$\log P(X\|K)$	$P(K\|X)$
1	−3144	~0
2	−2769	~0
3	−2678	0.993
4	−2683	0.007
5	−2688	0.00005

The values in the last column assume a uniform prior for K ($K \in \{1, \ldots, 5\}$).

Application to human data: The next data set, taken from Jorde *et al.* (1995), includes data from 30 biallelic restriction site polymorphisms, genotyped in 72 Africans (Sotho, Tsonga, Nguni, Biaka and Mbuti Pygmies, and San) and 90 Europeans (British and French).

Application of our MCMC scheme with $K = 2$ indicates the presence of two very distinct clusters, corresponding to the Africans and Europeans in the sample (Figure 5). The model with $K = 2$ has vastly higher posterior probability than the model with $K = 1$.

Additional runs of the MCMC scheme with the models $K = 3$, 4, and 5 suggest that those models may be somewhat better than $K = 2$. This may reflect the presence of population structure within the continental groupings, although in this case the additional populations do not form discrete clusters and so are difficult to interpret.

Again it is interesting to contrast our clustering results with the neighbor-joining tree of these data (Figure 6). While our method finds it quite easy to separate the two continental groups into the correct clusters, it would not be possible to use the neighbor-joining tree to detect distinct clusters if the labels were not present. The data set of Jorde also contains a set of individuals of Asian origin (which are more closely related to Europeans than are Africans). Neither the neighbor-joining method nor our method differentiates between the Europeans and Asians with great accuracy using this data set.

INCORPORATING POPULATION INFORMATION

The results presented so far have focused on testing how well our method works. We now turn our attention to some further applications of this method.

Our clustering results (Figure 4) confirm that the three main geographic groupings in the thrush data set (Chawia, Mbololo, and Ngangao) represent three genetically distinct populations. The geographic labels correspond very closely to the genetic clustering in all but a handful of cases (1–4 in Figure 4). Individual 2 is also identified as a possible outlier on the neighbor-joining tree (Figure 3). Given this, it is natural to ask whether these apparent outliers are immigrants (or de-

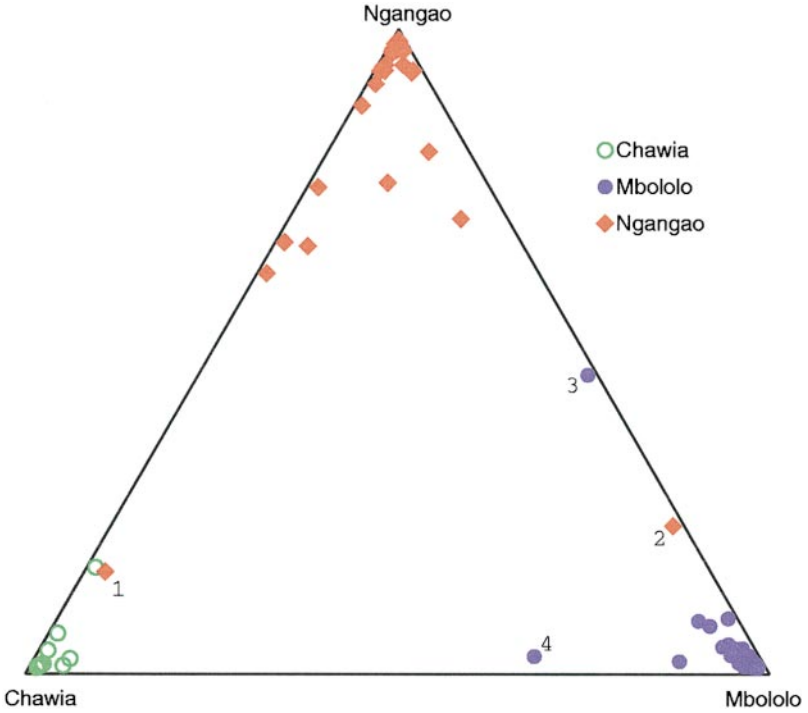

Figure 4.—Summary of the clustering results for the *T. helleri* data assuming three populations. Each point shows the mean estimated ancestry for an individual in the sample. For a given individual, the values of the three coefficients in the ancestry vector $q^{(i)}$ are given by the distances to each of the three sides of the equilateral triangle. After the clustering was performed, the points were labeled according to sampling location. Numbers 1–4 are individuals who appear to be possible outliers (see text). For clarity, the four Yale individuals (who fall into the Ngangao cluster) are not plotted. We were not told the sampling locations of individuals until after we obtained these results.

scendants of recent immigrants) from other populations. For example, given the genetic data, how probable is it that individual 1 is actually an immigrant from Chawia?

To answer this sort of question, we need to extend our algorithm to incorporate the geographic labels. By doing this, we break the symmetry of the labels, and we can ask specifically whether a particular individual is a migrant from Chawia (say). In essence our approach (described more formally in the next section) is to assume that each individual originated, with high probability, in the geographical region in which it was sampled, but to allow some small probability that it is an immigrant (or has immigrant ancestry). Note that this model is also suitable for situations in which individuals are classified according to some characteristic other than sampling location (physical appearance, for example). "Immigrants" in this situation would be individuals whose genetic makeup suggests they were misclassified. Thus, while we speak of "immigrants" and "immigrant ancestry," in some contexts these terms may relate to something other than changes in physical location.

Provided that geographic labels *usually* correspond to population membership, using the geographic infor-

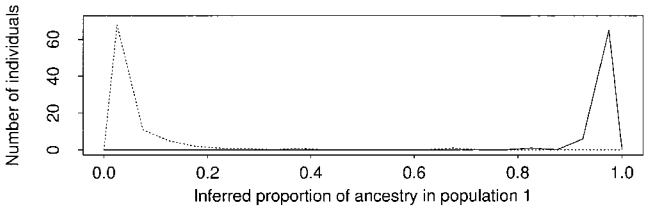

Figure 5.—Summary of the clustering results for the data set of Africans and Europeans taken from Jorde *et al.* (1995). For each individual, we computed the mean value of $q_1^{(i)}$ (the proportion of ancestry in population 1), over a single run of the Gibbs sampler. The dashed line is a histogram of mean values of $q_1^{(i)}$ for individuals of European origin; the solid line is for individuals of African origin.

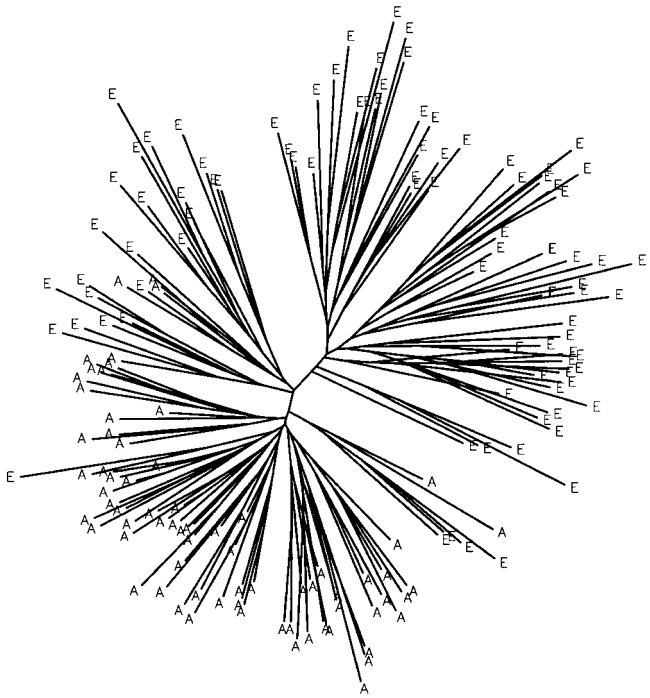

Figure 6.—Neighbor-joining tree of individuals in the data set of Jorde *et al.* (1995). Each tip represents a single individual. A and E indicate that individuals were African or European, respectively. The tree was constructed as in Figure 3.

mation will clearly improve our accuracy at assigning individuals to clusters; it will also improve our estimates of P, thus also giving us greater precision in assignment of individuals who do not have geographic information. However, in practice we suggest that before making use of such information, users of our method should first cluster the data without using the geographic labels, to check that the genetically defined clusters do in fact agree with geographic labels. We return to this issue in the discussion.

Rannala and Mountain (1997) also considered the problem of detecting immigrants and individuals with recent immigrant ancestors, taking a somewhat similar approach to that used here. However, rather than considering all individuals simultaneously, as we do here, they test each individual in the sample, one at a time, as a possible immigrant, assuming that all the other individuals are not immigrants. This approach will have reduced power to detect immigrants if the sample contains several immigrants from one population to another. In contrast, our approach can cope well with this kind of situation.

Model with prior population information: To incorporate geographic information, we use the following model. Our primary goal is to identify individuals who are immigrants, or who have recent immigrant ancestry, in the last G generations, say, where $G = 0$ is the present generation. [In practice there will only be substantial power to detect immigration for small G; *cf.* Rannala and Mountain (1997).]

First, we code each of the geographic locations by a (unique) integer between 1 and K, where K would usually be set equal to the number of locations. Using this coding, let $g^{(i)}$ represent the geographic sampling location of individual i. Now, let ν be the probability that an individual is an immigrant to population $g^{(i)}$ or has an immigrant ancestor in the last G generations. Otherwise, with probability $1 - \nu$, the individual is considered to be purely from population $g^{(i)}$. While in principle one could place a prior on ν and learn about it from the data as part of the MCMC scheme, in our current implementation the user must specify a fixed value for ν; we give some guidelines in the next section.

Assuming that migration is rare, we can use the approximation that each individual has at most one immigrant ancestor in the last G generations (where G is suitably small). Then, assuming a constant migration rate, the probability of an immigrant ancestor in generation t ($0 \le t \le G$) is proportional to 2^t, where $t = 0$ indicates that the individual migrated in the present generation. Thus, we set the prior on $q^{(i)}$ to be

$$q^{(i)}_{g^{(i)}} = 1, \quad q^{(i)}_k = 0 \quad (k \neq g^{(i)}) \tag{15}$$

with probability $1 - \nu$ and

$$q^{(i)}_{g^{(i)}} = 1 - 2^{-t},\ q^{(i)}_j = 2^{-t},\ q^{(i)}_k = 0 \quad (k \neq g^{(i)}, j) \tag{16}$$

for each $j \neq g^{(i)}$ with probability

$$\frac{2^t \nu}{(K-1)\Sigma_{T=0}^{G} 2^T}, \tag{17}$$

where $t \in \{0, \ldots, G\}$. As before, $q^{(i)}_l \ge 0$ for $l \in \{1, \ldots, K\}$, and $\Sigma q^{(i)}_l = 1$.

Again, we can sample from $\Pr(Q|X)$ using *Algorithm 2*. In this case, however, since there are a small number of possible values of $q^{(i)}$, we update $q^{(i)}$ by sampling directly from the posterior probability of $q^{(i)}|X,P$, rather than conditional on Z.

Note that in this framework, it is easy to include individuals for whom there is no geographic information by using the same prior and update steps as before (Equations 7 and A10).

Testing for migrants in the Taita thrush data: To apply our method, we must first specify a value for ν. In this case, based on mark-release-recapture data from these populations (Galbusera *et al.* 2000), migration seems relatively rare, and so ν is likely to be small. We performed analyses for $\nu = 0.05$ and $\nu = 0.1$; a summary of the results is shown in Table 4. Individuals 2 and 3 have moderate posterior probabilities of having migrant ancestry, but these probabilities are perhaps smaller than might be expected from examining Figure 4. This is due to a combination of the low prior probability for migration (from the mark-release-recapture data) and, perhaps more importantly, the fact that there is a limited amount of information in seven loci, so that the uncertainty associated with the position of the points marked 1, 2, 3, and 4 in Figure 4 may be quite large. A more definite conclusion could be obtained by typing more loci.

It is interesting to note that our conclusions here differ from those obtained on this data set using the package IMMANC (Rannala and Mountain 1997). IMMANC indicates that three individuals (1, 2, and 3 here) show significant evidence of immigrant ancestry at the 0.01 significance level (Galbusera *et al.* 2000). However, IMMANC does not make a multiple comparisons correction; such a correction would bring those results into line with ours.

We anticipate that our method might also be applied in situations where there is little data to help make an informed choice of ν. In such situations we suggest analyzing the data using several different values of ν, to see whether the conclusions are robust to choice of ν. The range of sensible values for ν will depend on the context, but typically we suggest values in the range 0.001–0.1 might be appropriate. Sensitivity to choice of ν indicates that the amount of information in the data is insufficient to draw strong conclusions.

DISCUSSION

We have described a method for using multilocus genotype data to learn about population structure and assign individuals (probabilistically) to populations.

TABLE 4

Testing whether particular individuals are immigrants or have recent immigrant ancestors

Individual	Geographic origin	Possible source	ν	No immigrant ancestry	Immigrant	Immigrant parent	Immigrant grandparent
1	Ngangao	Chawia	0.05	0.869	0.008	0.052	0.063
			0.10	0.739	0.019	0.106	0.123
2	Ngangao	Mbololo	0.05	0.673	0.029	0.126	0.168
			0.10	0.472	0.046	0.203	0.273
3	Mbololo	Ngangao	0.05	0.649	0.002	0.179	0.165
			0.10	0.464	0.003	0.271	0.253
4	Mbololo	Chawia	0.05	0.891	0.000	0.007	0.082
			0.10	0.791	0.000	0.014	0.157

The individuals are labeled as shown in Figure 4. "No immigrant ancestry" gives the probability that the ancestry of each individual is exclusively in the geographic origin population; the following columns show the probabilities that each individual has the given amount of ancestry in the possible source population. The rows do not add to 1 because there are small probabilities associated with individuals having ancestry in the third population.

Our method also provides a novel approach to testing for the presence of population structure ($K > 1$).

Our examples demonstrate that the method can accurately cluster individuals into their appropriate populations, even using only a modest number of loci. In practice, the accuracy of the assignments depends on a number of factors, including the number of individuals (which affects the accuracy of the estimate for P), the number of loci (which affects the accuracy of the estimate for Q), the amount of admixture, and the extent of allele-frequency differences among populations.

We anticipate that our method will be useful for identifying populations and assigning individuals in situations where there is little information about population structure. It should also be useful in problems where cryptic population structure is a concern, as a way of identifying subpopulations. Even in situations where there is nongenetic information that can be used to define populations, it may be useful to use the approach developed here to ensure that populations defined on an extrinsic basis reflect the underlying genetic structure.

As described in incorporating population information we have also developed a framework that makes it possible to combine genetic information with prior information about the geographic sampling location of individuals. Besides being used to detect migrants, this could also be used in situations where there is strong prior population information for some individuals, but not for others. For example, in hybrid zones it may be possible to identify some individuals who do not have mixed ancestry and then to estimate q for the rest (M. Beaumont, D. Gotelli, E. M. Barett, A. C. Kitchener, M. J. Daniels, J. K. Pritchard and M. W. Bruford, unpublished results). The advantage of using a clustering approach in such cases is that it makes the method more robust to the presence of misclassified individuals and should be more accurate than if only preclassified individuals are used to estimate allele frequencies (*cf.* Smouse *et al.* 1990).

Another type of application where the geographic information might be of value is in evolutionary studies of population relationships. Such analyses frequently make use of summary statistics based on population allele frequencies [*e.g.*, F_{ST} and $(\delta\mu)^2$]. In situations where the population allele frequencies might be affected by recent immigration or where population classifications are unclear, such summary statistics could be calculated directly from the population allele frequencies P estimated by the Gibbs sampler.

There are several ways in which the basic model that we have described here might be modified to produce better performance in particular cases. For example, in models and methods and applications to data we assumed relatively noninformative priors for q. However, in some situations, there might be quite a bit of information about likely values of q, and the estimation procedure could be improved by using that information. For example, in estimating admixture proportions for African Americans, it would be possible to improve the estimation procedure by making use of existing information about the extent of European admixture (*e.g.*, Parra *et al.* 1998).

A second way in which the basic model can be modified involves changing the way in which the allele frequencies P are estimated. Throughout this article, we have assumed that the allele frequencies in different populations are uncorrelated with one another. This is a convenient approximation for populations that are not extremely closely related and, as we have seen, can produce accurate clustering. However, loosely speaking, the model of uncorrelated allele frequencies says that we do not normally expect to see populations with very similar allele frequencies. This property has the result that the clustering algorithm may tend to merge subpopulations that share similar frequencies. An alternative,

which we have implemented in our software package, is to permit allele frequencies to be correlated across populations (appendix, *Model with correlated allele frequencies*). In a series of additional simulations, we have found that this allows us to perform accurate assignments of individuals in very closely related populations, though possibly at the cost of making us likely to overestimate *K*.

Our basic model might also be modified to allow for linkage among marker loci. Normally, we would not expect to see linkage disequilibrium within subpopulations, except between markers that are extremely close together. This means that in situations where there is little admixture, our assumption of independence among loci will be quite accurate. However, we might expect to see strong correlations among linked loci when there is recent admixture. This occurs because an individual who is admixed will inherit large chromosomal segments from one population or another. Thus, when the map order of marker loci is known, it should be possible to improve the accuracy of the estimation for such individuals by modeling the inheritance of these segments.

In this article we have devoted considerable attention to the problem of inferring *K*. This is an important practical problem from the standpoint of model choice. We need to have some way of deciding which clustering model is most appropriate for interpreting the data. However, we stress that care should be taken in the interpretation of the inferred value of *K*. To begin with, due to the very high dimensionality of the parameter space, we found it difficult to obtain reliable estimates of $\Pr(X \mid K)$ and have chosen to use a fairly *ad hoc* procedure that we have found gives sensible results in practice. Second, it has been observed that in Bayesian model-based clustering, the posterior distribution of *K* tends to be quite dependent on the priors and modeling assumptions, even though estimates of the other parameters (*e.g.*, *P* and *Q* here) may be reasonably robust (see Richardson and Green 1997; Stephens 2000a, for example).

There are also biological reasons to be careful interpreting *K*. The population model that we have adopted here is obviously an idealization. We anticipate that it will be flexible enough to permit appropriate clustering for a wide range of population structures. However, as we pointed out in our discussion of data set 3 (*Choice of K for simulated data*), clusters may not necessarily correspond to "real" populations. As another example, imagine a species that lives on a continuous plane, but has low dispersal rates, so that allele frequencies vary continuously across the plane. If we sample at *K* distinct locations, we might infer the presence of *K* clusters, but the inferred number *K* is not *biologically* interesting, as it was determined purely by the sampling scheme. All that can usefully be said in such a situation is that the migration rates between the sampling locations are not high enough to make the population act as a single unstructured population.

In summary, we find that the method described here can produce highly accurate clustering and sensible choices of *K*, both for simulated data and for real data from humans and from the Taita thrush. In the latter example, we find it particularly encouraging that using a relatively small number of loci (seven) we can detect a very strong signal of population structure and assign individuals appropriately.

The algorithms described in this article have been implemented in a computer software package *structure*, which is available at http://www.stats.ox.ac.uk/~pritch/home.html.

We thank Peter Galbusera and Lynn Jorde for allowing us to use their data, Augie Kong for a helpful discussion, Daniel Falush for suggesting comparison with neighbor-joining trees, Steve Brooks and Trevor Sweeting for helpful discussions on inferring *K*, and Eric Anderson for his extensive comments on an earlier version of the manuscript. This work was supported by National Institutes of Health grant GM19634 and by a Hitchings-Elion fellowship from Burroughs-Wellcome Fund to J.K.P., by a grant from the University of Oxford and a Wellcome Trust Fellowship (057416) to M.S., and by grants GR/M14197 and 43/MMI09788, from the Engineering and Physical Sciences Research Council and Biotechnology and Biological Sciences Research Council, respectively, to P.D. The work was initiated while the authors were resident at the Isaac Newton Institute for Mathematical Sciences, Cambridge, UK.

LITERATURE CITED

Balding, D. J., and R. A. Nichols, 1994 DNA profile match probability calculations: how to allow for population stratification, relatedness, database selection and single bands. Forensic Sci. Int. **64:** 125–140.

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.

Bowcock, A. M., A. Ruiz-Linares, J. Tomfohrde, E. Minch, J. Kidd *et al.*, 1994 High resolution of human evolutionary trees with polymorphic microsatellites. Nature **368:** 455–457.

Cavalli-Sforza, L. L., P. Menozzi and A. Piazza, 1994 *The History and Geography of Human Genes*. Princeton University Press, Princeton, NJ.

Chib, S., 1995 Marginal likelihood from the Gibbs output. J. Am. Stat. Assoc. **90:** 1313–1321.

Chib, S., and E. Greenberg, 1995 Understanding the Metropolis-Hastings algorithm. Am. Stat. **49:** 327–335.

Davies, N., F. X. Villablanca and G. K. Roderick, 1999 Determining the source of individuals: multilocus genotyping in nonequilibrium population genetics. TREE **14:** 17–21.

DiCiccio, T., R. Kass, A. Raftery and L. Wasserman, 1997 Computing Bayes factors by posterior simulation and asymptotic approximations. J. Am. Stat. Assoc. **92:** 903–915.

Ewens, W. J., and R. S. Spielman, 1995 The transmission/disequilibrium test: history, subdivision, and admixture. Am. J. Hum. Genet. **57:** 455–464.

Felsenstein, J., 1993 PHYLIP (phylogeny inference package) version 3.5c. Technical report, Department of Genetics, University of Washington, Seattle.

Foreman, L., A. Smith and I. Evett, 1997 Bayesian analysis of DNA profiling data in forensic identification applications. J. R. Stat. Soc. A **160:** 429–469.

Galbusera, P., L. Lens, E. Waiyaki, T. Schenck and E. Mattysen, 2000 Effective population size and gene flow in the globally,

critically endangered Taita thrush, *Turdus helleri*. Conserv. Genet. (in press).
Gilks, W. R., S. Richardson and D. J. Spiegelhalter, 1996a Introducing Markov chain Monte Carlo, pp. 1–19 in *Markov Chain Monte Carlo in Practice*, edited by W. R. Gilks, S. Richardson and D. J. Spiegelhalter. Chapman & Hall, London.
Gilks, W. R., S. Richardson and D. J. Spiegelhalter (Editors), 1996b *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London.
Goldstein, D. B., and D. Pollock, 1997 Launching microsatellites: a review of mutation processes and methods of phylogenetic inference. J. Hered. **88:** 335–342.
Green, P. J., 1995 Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika **82:** 711–732.
Hudson, R. R., 1990 Gene genealogies and the coalescent process, pp. 1–44 in *Oxford Surveys in Evolutionary Biology*, Vol. 7, edited by D. Futuyma and J. Antonovics. Oxford University Press, Oxford.
Jorde, L. B., M. J. Bamshad, W. S. Watkins, R. Zenger, A. E. Fraley *et al.*, 1995 Origins and affinities of modern humans: a comparison of mitochondrial and nuclear genetic data. Am. J. Hum. Genet. **57:** 523–538.
Mountain, J. L., and L. L. Cavalli-Sforza, 1997 Multilocus genotypes, a tree of individuals, and human evolutionary history. Am. J. Hum. Genet. **61:** 705–718.
Paetkau, D., W. Calvert, I. Stirling and C. Strobeck, 1995 Microsatellite analysis of population structure in Canadian polar bears. Mol. Ecol. **4:** 347–354.
Parra, E. J., A. Marcini, J. Akey, J. Martinson, M. A. Batzer *et al.*, 1998 Estimating African American admixture proportions by use of population-specific alleles. Am. J. Hum. Genet. **63:** 1839–1851.
Pritchard, J. K., and N. A. Rosenberg, 1999 Use of unlinked genetic markers to detect population stratification in association studies. Am. J. Hum. Genet. **65:** 220–228.
Raftery, A. E., 1996 Hypothesis testing and model selection, pp. 163–188 in *Markov Chain Monte Carlo in Practice*, edited by W. R. Gilks, S. Richardson and D. J. Spiegelhalter. Chapman & Hall, London.
Rannala, B., and J. L. Mountain, 1997 Detecting immigration by using multilocus genotypes. Proc. Natl. Acad. Sci. USA **94:** 9197–9201.
Richardson, S., and P. J. Green, 1997 On Bayesian analysis of mixtures with an unknown number of components. J. R. Stat. Soc. Ser. B **59:** 731–792.
Roeder, K., M. Escobar, J. B. Kadane and I. Balazs, 1998 Measuring heterogeneity in forensic databases using hierarchical Bayes models. Biometrika **85:** 269–287.
Smouse, P. E., R. S. Waples and J. A. Tworek, 1990 A genetic mixture analysis for use with incomplete source population-data. Can. J. Fish. Aquat. Sci. **47:** 620–634.
Spiegelhalter, D. J., N. G. Best and B. P. Carlin, 1999 Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models. Available from http://www.mrc-bsu.cam.ac.uk/publications/preslid.shtml.
Stephens, M., 2000a Bayesian analysis of mixtures with an unknown number of components—an alternative to reversible jump methods. Ann. Stat. (in press).
Stephens, M., 2000b Dealing with label-switching in mixture models. J. R. Stat. Soc. Ser. B (in press).

Communicating editor: M. K. Uyenoyama

APPENDIX

MCMC methods and Gibbs sampling:

MCMC methods are extremely useful for obtaining (approximate) samples from a probability distribution, $\pi(\theta)$, say, which cannot be simulated from directly [in our case $\theta = (Z, P, Q)$ and $\pi(\theta) = \Pr(Z, P, Q|X)$]. The idea is to construct a Markov chain $\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \ldots$ with stationary distribution $\pi(\theta)$. This is often surprisingly straightforward using standard methods devised for this purpose, such as the Metropolis-Hastings algorithm (*e.g.*, Chib and Greenberg 1995) and Gibbs sampling (*e.g.*, Gilks *et al.* 1996a), which we describe in more detail below. Intuitively, if the Markov chain $\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \ldots$ has stationary distribution $\pi(\theta)$, then $\theta^{(m)}$ will be approximately distributed as $\pi(\theta)$ provided m is sufficiently large. This can be formalized and shown to be true provided the Markov chain satisfies certain technical conditions (*ergodicity*) that hold for the Markov chains considered in this article. Furthermore, for sufficiently large c, $\theta^{(m)}, \theta^{(m+c)}, \theta^{(m+2c)}, \ldots$ will be reasonably independent samples from $\pi(\theta)$. The value of m used is often referred to as the *burn-in* period of the chain; c is often referred to as the *thinning* interval.

In general it is very difficult to know how large m and c should be. The values required to obtain reliable results depend heavily on the amount of correlation between successive states of the Markov chain. If successive states are relatively uncorrelated (that is, if the chain moves quickly between reasonably different values of θ), then the chain is said to *mix* well, and relatively small values of m and c will suffice. Conversely, if the chain mixes badly (sometimes known as being *sticky*, as the chain will tend to get stuck moving among very similar values of θ), then very large values of m and c will be required, possibly rendering the method impracticable. One strategy for investigating whether m and c are sufficiently large, and the strategy we adopt here, is to simulate several realizations of the Markov chain, each starting from a different value of $\theta^{(0)}$. If m and c are sufficiently large, then the results obtained should be independent of $\theta^{(0)}$ and should therefore be similar for the different runs. Substantial differences among the results obtained for the different runs indicate that m and c are too small. It is then necessary either to increase m and c or (if this makes the method computationally infeasible) to construct a Markov chain with better mixing properties. In the examples presented in this article we have chosen $c = 1$.

Gibbs sampling is a method of constructing a Markov chain with stationary distribution $\pi(\theta)$, which has proved particularly useful for clustering problems. Suppose that θ may be partitioned into $\theta = (\theta_1, \ldots, \theta_r)$, and that although it is not possible to simulate from $\pi(\theta)$ directly, it *is* possible to simulate a random value of θ_i directly from the full conditional distribution $\pi(\theta_i \mid \theta_1, \theta_2, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_r)$ for $i = 1, 2, \ldots, r$. Then the following algorithm may be used to simulate a Markov chain with stationary distribution $\pi(\theta)$:

Algorithm A1: *Starting with initial values* $\theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_r^{(0)})$, *iterate the following steps for* $m = 1, 2, \ldots$.

Step 1. *Sample* $\theta_1^{(m)}$ *from* $\pi(\theta_1|\theta_2^{(m-1)}, \theta_3^{(m-1)}, \ldots, \theta_r^{(m-1)})$.
Step 2. *Sample* $\theta_2^{(m)}$ *from* $\pi(\theta_2|\theta_1^{(m)}, \theta_3^{(m-1)}, \ldots, \theta_r^{(m-1)})$.

Step r. Sample $\theta_r^{(m)}$ *from* $\pi(\theta_r|\theta_1^{(m)}, \theta_2^{(m)}, \ldots, \theta_{r-1}^{(m)})$.

It is easy to show that if $\theta^{(m-1)} \sim \pi(\theta)$, then $\theta^{(m)} \sim \pi(\theta)$, and so $\pi(\theta)$ is the stationary distribution of this Markov chain.

Inference on *K*, the number of populations

We now provide further details regarding our approach to choosing K (see *Inference for the number of populations*).

The simplest way of estimating $\Pr(X|K)$ is the so-called harmonic mean estimator

$$\frac{1}{\Pr(X|K)} = \int \frac{\Pr(Z, P, Q|X, K)}{\Pr(X|Z, P, Q, K)} dZdPdQ \approx \frac{1}{M}\sum_{m=1}^{M} \frac{1}{\Pr(X|Z^{(m)}, P^{(m)}, Q^{(m)}, K)}. \quad \text{(A1)}$$

This estimator is notoriously unstable, often having infinite variance, and is thus of little use in practice. One theoretically attractive alternative involves estimating $\Pr(P, Q|X)$ for some P, Q (Chib 1995; Raftery 1996). However, our own implementation of versions of this approach has turned out to be computationally infeasible, due to the very high-dimensional parameter space of our problem. While alternative approaches to estimating $\Pr(X|K)$, such as variable-dimension MCMC methods (Green 1995; Stephens 2000a) or importance sampling (DiCiccio *et al.* 1997), may lead to computationally feasible algorithms, the high-dimensional parameter space makes designing efficient versions of these schemes rather challenging. For this reason we take a more *ad hoc* approach, which begins by considering the *Bayesian deviance*

$$D(Z, P, Q) = -2 \log \Pr(X|Z, P, Q). \quad \text{(A2)}$$

The conditional mean and variance of D given X are easily estimated using

$$E(D(Z, P, Q)|X) \approx \frac{1}{M}\sum_{m=1}^{M} -2 \log \Pr(X|Z^{(m)}, P^{(m)}, Q^{(m)}) = \hat{\mu}, \text{ say}, \quad \text{(A3)}$$

and

$$\mathrm{Var}(D(Z, P, Q)|X) \approx \frac{1}{M}\sum_{m=1}^{M} (-2 \log \Pr(X|Z^{(m)}, P^{(m)}, Q^{(m)}) - \hat{\mu})^2 = \hat{\sigma}^2, \text{ say}. \quad \text{(A4)}$$

If we make the (admittedly dubious) assumption that the conditional distribution of D given X is normal, then it follows from (A1) that

$$-2 \log(\Pr(X|K)) \approx \hat{\mu} + \hat{\sigma}^2/4. \quad \text{(A5)}$$

(Replacing the assumption of normality with the assumption of being gamma-distributed may be more asymptotically justifiable and gives similar results.) We then use this to estimate the posterior distribution of K from (11). An alternative interpretation of this method is that model selection is based on penalizing the mean of the Bayesian deviance by a quarter of its variance (*cf.* Spiegelhalter *et al.* 1999, who suggested investigating model fit using a different penalization of the mean of the Bayesian deviance).

Details of the MCMC algorithms

Algorithm A2: Step 1 may be performed by simulating $p_{kl\cdot}$ independently for each (k, l), from

$$p_{kl\cdot}|X, Z \sim \mathcal{D}(\lambda_1 + n_{kl1}, \ldots, \lambda_{l_l} + n_{kll_l}), \quad \text{(A6)}$$

where

$$n_{klj} = \#\{(i,a) : x_l^{(i,a)} = j \text{ and } z^{(i)} = k\} \quad \text{(A7)}$$

is the number of copies of allele j at locus l observed in individuals assigned (by Z) to population k.

Step 2 may be performed by simulating $z^{(i)}$, independently for each i, from

$$\Pr(z^{(i)} = k|X, P) = \frac{\Pr(x^{(i)}|P, z^{(i)} = k)}{\sum_{k'=1}^{K}\Pr(x^{(i)}|P, z^{(i)} = k')}, \quad \text{(A8)}$$

where $\Pr(x^{(i)}|P, z^{(i)} = k) = \prod_{l=1}^{L} p_{klx(i,1)}p_{klx(i,2)}$.

Note that Equation A8 makes an implicit assumption that an equal fraction of the sample is drawn from each population. Alternatively, it might be natural to introduce an additional parameter for the fraction of the sample drawn from each population.

Algorithm A3: Step 1 may be performed by updating P and Q independently. Updating P is achieved as before, using Equation A6 but where the definition (A7) of n_{klj} is modified in the obvious way to

$$n_{klj} = \#\{(i, a) : x_l^{(i,a)} = j \text{ and } z_l^{(i,a)} = k\}. \quad \text{(A9)}$$

Updating Q involves simulating from

$$q^{(i)}|X, Z \sim \mathcal{D}(\alpha + m_1^{(i)}, \ldots, \alpha + m_K^{(i)}), \quad \text{(A10)}$$

where $m_k^{(i)}$ is the number of allele copies in individual i that originated (according to Z) in population k:

$$m_k^{(i)} = \#\{(l, a) : z_l^{(i,a)} = k\}. \quad \text{(A11)}$$

Step 2 may be performed by simulating $z_l^{(i,a)}$, independently for each i, a, l, from

$$\Pr(z_l^{(i,a)} = k|X, P) = \frac{q_k^{(i)}\Pr(x_l^{(i,a)}|P, z_l^{(i,a)} = k)}{\sum_{k'=1}^{K} q_{k'}^{(i)}\Pr(x_l^{(i,a)}|P, z_l^{(i,a)} = k')}, \quad \text{(A12)}$$

where $\Pr(x_l^{(i,a)}|P, z_l^{(i,a)} = k) = p_{klx_l^{(i,a)}}$.

Step 3 may be performed by simulating a proposal α', from a normal distribution with mean α, and some variance σ_α^2. The proposal is automatically rejected if $\alpha' \leq 0$, and otherwise it is accepted with the appropriate Metropolis-Hastings probability.

Model with correlated allele frequencies

For very closely related populations it is natural to assume that allele frequencies are correlated across populations. For completeness, we describe a model that is implemented in the program *structure*, allowing allele-frequency correlations.

Recall that we model allele frequencies by $p_{kl\cdot} \sim \mathcal{D}(\lambda_1, \lambda_2, \ldots, \lambda_{J_l})$. For all the results presented in this article, we took $\lambda_1 = \lambda_2 = \cdots = \lambda_{Jl} = 1.0$, which gives a uniform distribution on allele frequencies, where J_l is the number of alleles at lows l. To model closely related populations, we consider an alternative model, where

$$p_{kl\cdot} \sim \mathcal{D}(f^{(l)}J_l\mu_1^{(l)}, f^{(l)}J_l\mu_2^{(l)}, \ldots, f^{(l)}J_l\mu_{J_l}^{(l)}). \quad \text{(A13)}$$

Here, $\mu_i^{(l)}$ is the mean sample frequency of allele i at locus l, and $f^{(l)} > 0$ determines the strength of the correlations across populations at locus l. When $f^{(l)}$ is large, the allele frequencies in all populations tend to be similar to the mean allele frequencies in the sample. In our implementation of this model, we placed a gamma prior on each $f^{(l)}$ and used a Metropolis-Hastings update step. The proposal $f^{(l)'}$ was chosen from a normal with mean $f^{(l)}$ and some variance σ_f^2. It was automatically rejected if $f^{(l)'} \leq 0$.

There are several possible alternative models to considering a factor f for each locus. One would be to consider a factor for each population, and another would be to give each type of locus (*e.g.*, SNPs and dinucleotide and trinucleotide repeats) a shared value of f.