Inference of Population Structure Using Genetic Markers and a Bayesian Model Averaging Approach for Clustering

GUZMÁN SANTAFÉ, JOSE A. LOZANO, and PEDRO LARRAÑAGA

The analysis of the structure of populations on the basis of genetic data is essential in population genetics. It is used, for instance, to study the evolution of species or to correct for population stratification in association studies. These genetic data, normally based on DNA polymorphisms, may contain irrelevant information that biases the inference of population structure. In this paper we adapt a recently proposed algorithm, named multi-start EMA, to be used in the inference of population structure. This algorithm is able to deal with irrelevant information when obtaining the (probabilistic) population partition. Additionally, we present a maker selection test able to obtain the most relevant markers to retrieve that population partition. The proposed algorithm is compared with the widely used STRUCTURE software on the basis of the F_{ST} metric and the log-likelihood score. It is shown that the proposed algorithm improves the obtention of the population structure. Moreover, information about relevant markers obtained by the multi-start EMA can be used to improve the results obtained by other methods, correct for population stratification or even also reduce the economical cost of sequencing new samples. The software presented in this paper is available online at *http://www.sc.ehu.es/ccwbayes/members/guzman*.

1. INTRODUCTION

THE DNA POLYMORPHISMS ARE VARIATIONS IN DNA sequence along individuals within species.

Polymorphism between individuals can arise through several different mechanisms which include single nucleotide changes (SNPs), deletions and insertions of nucleotides and, above all, through variable numbers of short nucleotide sequence repeats (microsatellites). All these polymorphisms occurred during the history of species and are inherited among generations.

Worldwide human population is usually defined in terms of subjective aspects such as language, culture, physical appearance or geographic location. However, human populations also tend to be genetically distant. Genetic differences are caused by a fairly independent evolution under population genetic forces, such as

mutation, recombination, random drift and selection. This variation within and between populations can be observed at genetic marker locations. Recent studies using a variety of genetic markers, have shown that individuals sampled worldwide fall into clusters that roughly correspond to continental lines as well as to self-identifying racial groups (Bamshad et al., 2003; Corander and Marttinen, 2006; Rosenberg et al., 2002, 2005).

The information about population structure, namely population stratification and admixture, is useful not only in evolutionary studies or subspecies classification (Pritchard et al., 2000b; Rosenberg et al., 2002) but also in association studies of disease genes (Patterson et al., 2004; Price et al., 2006; Riddle et al., 2006; Sillanpää et al., 2001). Association studies often use a case-control design to identify genetic variants related to a specific disease by comparing allele frequencies between unrelated individuals that are affected and those unaffected. However, the presence of population stratification can lead to spurious allelic association between candidate marker and a phenotype (Cardon and Palmer, 2003; Pritchard et al., 2000a).

From a machine learning point of view, the inference of population structure can be seen as a clustering process where individuals are assigned to their population of origin according to their DNA polymorphisms. In the literature two main clustering approaches to the inference of population structures can be found; distance-based methods and model-based methods. Distance-based methods use pairwise distances between individuals to obtain a clustering partition of the population (Bowcock et al., 1994). These methods are highly dependent on the selected distance measure and therefore it is very difficult to know if the obtained clustering partition is meaningful. On the other hand, model-based clustering assumes that there is a generative probabilistic model underlying the genetic information of the individuals. Another key modeling assumption is linkage and Hardy-Weinberg disequilibrium. Since these models are based on probability theory, a large amount of methods from statistical learning, sampling theory and Bayesian statistics can be used. Bayesian statistical methods based on Markov chain Monte Carlo (MCMC) are commonly used for the inference of population structure. Particularly, STRUCTURE (Falush et al., 2003; Pritchard et al., 2000b) is one of the most widely used algorithms based on MCMC. However, there are other proposals such as PARTITION (Dawson and Belkhir, 2001), BAPS 2 (Corander and Marttinen, 2006; Corander et al., 2004), a spatial statistical model for landscape genetics proposed by Guillot et al. (2004), or the learning of mixtures of trees (Kollin and Koivisto, 2006). Additionally, there are other algorithms, for instance methods based on the EM algorithm (Dempster et al., 1977) such as PSMIX (Wu et al., 2006), or based on information theory (O'Rourke et al., 2005).

Even when high-throughput technologies offer the possibility of measuring a large number of polymorphisms simultaneously, it has sometimes been observed, for certain ancestry inference procedures, that accuracy of inference does not necessarily increase as markers are accumulated. Indeed, in an increasing number of species, the number of markers from which allele frequencies are available exceeds those required for accurate assignments. However, it is possible to find robust clustering patterns by using a panel with only a part of the markers that are available (Rosenberg, 2005; Rosenberg et al., 2003; Turakulov and Easteal, 2003). Thus, not only may the accuracy and efficiency of the population inference method be improved but also the genotyping cost reduced. Nevertheless, there is not a clear criteria to select the set of markers needed to obtain a robust clustering partition. Furthermore, the fact that not all the makers available are needed, suggests that there is redundant information and/or makers that are not relevant to cluster the individuals into their population of origin. This redundant and irrelevant information may damage the ability of the clustering methods to infer the population structure.

In this work, we adapt the multi-start Expectation Model Averaging (multi-start EMA) algorithm (Santafé et al., 2006) to infer the structure of a population. The multi-start EMA is a recently proposed algorithm which approximates Bayesian model averaging for clustering based on the naive Bayes model, which is a simple Bayesian network successfully used in many other biological problems (Barash and Friedman, 2002).

Naive Bayes model assumes that allele frequencies of any two markers are independent once the population of origin is known. Additionally, the Bayesian model averaging method used to learn the naive Bayes model underlying the population structure takes model uncertainty into account. That is, it takes into account the uncertainty about the usefulness of including and excluding each marker from the clustering model. This process incorporates a kind of implicit feature selection in the model and it can be used to filter out those genetic markers which are considered irrelevant for obtaining the population

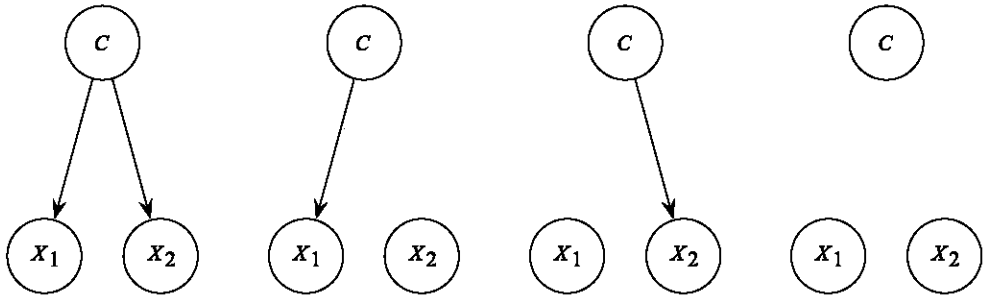

FIG. 1. Selective naive Bayes structures with two marker loci: each marker can be dependent on or independent of the clustering assignment, C. That is, each marker can be considered relevant or irrelevant for clustering purposes.

structure. The information about relevant markers may be useful in many other genetic studies. Hence, we also purpose a two-step test based on mutual information that can be used to retrieve this information from the clustering model.

2. METHODS

The marker loci are denoted by $\boldsymbol{X} = \{X_1, \ldots, X_n\}$ and the cluster variable, C, represents the population grouping of the N polyploid individuals $D = \{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(N)}\}$. Since we would like to use the proposed method with human genetic data, from now on we assume diploid data. Therefore, the lth individual of the population is characterized by n genetic markers, $\boldsymbol{x}^{(l)} = (x_1^{(l)}, \ldots, x_n^{(l)})$ and each marker X_i, with $i = 1, \ldots, n$, contains information for two alleles $x_i^{(l)} = \{x_i^{(l),h_1}, x_i^{(l),h_2}\}$. Capital letters, X_i, represent a marker locus and small letters, x_i, denote a specific allele value for the marker locus.

The model underlying the population structure is assumed to be a naive Bayes. However, the learning process performed by the multi-start EMA includes a kind of implicit selection of relevant marker loci in the model. We say that it is an implicit selection because all the markers are included in the naive Bayes model learned by the multi-start EMA, but the learning process gives rise to the fact that not all the markers have the same significance when obtaining the population structure.

2.1. *EMA algorithm*

The aim of the algorithm is to obtain a clustering model that provides a posterior distribution over the dataset and thus, allows to infer the population structure of the individuals under study. For this purpose, we adapt the EMA algorithm originally proposed in Santafé et al. (2006) to deal with polyploid[1] and missing data. Thus, we provide a useful and realistic tool for the inference of population structure.

The EMA is a greedy iterative algorithm that learns a naive Bayes model as a result of a Bayesian model averaging over all the possible selective naive Bayes models (see Figure 1 below for a graphical description of selective naive Bayes models). The method itself is an adaptation of the well-known EM algorithm (Dempster et al., 1977) that allows us to extend efficient model averaging techniques for supervised classification (Dash and Cooper, 2004) to clustering problems. Each iteration of the EMA algorithm comprises two steps: Expectation (E) and Model Averaging (MA) steps. At iteration t, the parameters of the model $\boldsymbol{\Theta}^{(t)} = (\theta_{ijk}^{(t)}, \theta_{C-j}^{(t)})$ denote the frequency of every allele k at marker X_i in a population j, $\theta_{ijk}^{(t)}$, and the prior probability of each population j, $\theta_{C-j}^{(t)}$. In order to calculate these parameters, the algorithm takes into account the possibility that any marker X_i can be relevant or irrelevant for the inference of the population structure. In the calculations, the frequency at marker X_i of the allele k in population j when marker X_i is considered relevant is represented as θ_{ijk}. However, if it is considered irrelevant, we take into consideration the overall frequency of allele k, represented as θ_{i-k}. The EMA algorithm is based on the

decomposability of model averaging calculations that can be performed by making two main assumptions. On the one hand, the allele frequencies are assumed to follow a Dirichlet distribution with parameters α_{ijk} if the marker X_i is considered relevant for the inference of population structure or α_{i-k} if it is considered irrelevant. Similarly, the prior probability over populations also follows a Dirichlet distribution with α_{C-j} parameters. On the other hand, we assume that θ_{ijk}, for any population j and allele k, is independent of $\theta_{i'jk}$ with $i \neq i'$, and similarly for θ_{i-k} and $\theta_{i'-k}$. This is known as parameter independence assumption.

The EMA algorithm successively performs the E and MA steps until the difference between the parameter set of the models calculated in two consecutive iterations is less than a given parameter, ϵ. The first parameter set, $\boldsymbol{\Theta}^{(0)}$, is usually taken at random. For a better understanding of the EMA algorithm, we describe in more detail the E and MA steps at each iteration t.

2.1.1. E step. This step includes the main modification with respect to the original algorithm. This modification allows the use of the EMA algorithm with polyploid and missing data. Intuitively, we can see this step as a probabilistic assignment of each individual to each population on the basis of the current model parameters $\boldsymbol{\Theta}^{(t)}$. Actually, this step computes, given the current model parameters $\boldsymbol{\Theta}^{(t)}$, the expected number of individuals from a population j that present allele k at the ith marker when the marker is considered relevant for clustering purposes, $E(N_{ijk}|\boldsymbol{\Theta}^{(t)})$, or irrelevant $E(N_{i-k}|\boldsymbol{\Theta}^{(t)})$, and the expected number of individuals classified into population j, $E(N_{C-j}|\boldsymbol{\Theta}^{(t)})$.

$$
\begin{aligned}
E(N_{ijk}|\boldsymbol{\Theta}^{(t)}) &= \sum_{l=1}^{N} p(c^j, x_i^{k,h_1}|\boldsymbol{x}^{(l)}, \boldsymbol{\Theta}^{(t)}) + p(c^j, x_i^{k,h_2}|\boldsymbol{x}^{(l)}, \boldsymbol{\Theta}^{(t)}) \\
E(N_{i-k}|\boldsymbol{\Theta}^{(t)}) &= \sum_{l=1}^{N} p(x_i^{k,h_1}|\boldsymbol{x}^{(l)}, \boldsymbol{\Theta}^{(t)}) + p(x_i^{k,h_2}|\boldsymbol{x}^{(l)}, \boldsymbol{\Theta}^{(t)}) \\
E(N_{C-j}|\boldsymbol{\Theta}^{(t)}) &= \sum_{l=1}^{N} p(c^j|\boldsymbol{x}^{(l)}, \boldsymbol{\Theta}^{(t)})
\end{aligned}
\tag{1}
$$

We abuse the notation by using c^j and x_i^{k,h_g}, with $g = 1, 2$, to denote the fact that the cluster variable C takes the jth value and marker X_i takes the kth value for the genetic copy h_g respectively. Although we avoid the use of superscript l in order to clarify the notation, the information contained in c^j and x_i^{k,h_g} is assumed to belong to the lth individual. Moreover, for a simpler notation, when we write $\boldsymbol{x}$, we assume that each marker X_i takes its kth allele value. Note that, both copies of the genetic information for each marker, $x_i = \{x_i^{k,h_1}, x_i^{k,h_2}\}$, are taken into account for the calculation of $E(N_{ijk}|\boldsymbol{\Theta}^{(t)})$ and $E(N_{i-k}|\boldsymbol{\Theta}^{(t)})$. In the original EMA algorithm, as missing values were not allowed, the values of $E(N_{i-k}|\boldsymbol{\Theta}^{(t)})$ were constant throughout the iterations of the algorithm. However, the current modification proposed in this paper allows the presence of missing data and therefore not only the value of $E(N_{ijk}|\boldsymbol{\Theta}^{(t)})$ and $E(N_{C-j}|\boldsymbol{\Theta}^{(t)})$ but also $E(N_{i-k}|\boldsymbol{\Theta}^{(t)})$ may change at each iteration. From now on, $D^{(t)}$ denotes the dataset after the E step at the tth iteration of the algorithm.

2.1.2. MA step. In this step, the EMA algorithm calculates a new set of parameters, $\boldsymbol{\Theta}^{(t+1)}$, for the model by averaging over all the selective naive Bayes models. These calculations are given by the following equation:

$$
\begin{aligned}
p(c_j, \boldsymbol{x}|D^{(t)}) &= \sum_S \underbrace{p(c_j, \boldsymbol{x}|D^{(t)}, S)}_{(1)} p(S|D^{(t)}) \\
&= \sum_S \underbrace{\int p(c_j, \boldsymbol{x}|S, \boldsymbol{\Theta}) \;\; p(\boldsymbol{\Theta}|S, D^{(t)}) d\boldsymbol{\Theta}}_{(1)} \underbrace{p(D^{(t)}|S)}_{(2)} P(S)
\end{aligned}
\tag{2}
$$

where S denotes a specific selective naive Bayes model that sets which markers are considered relevant for clustering purpose and which are not.

The general idea of an efficient model averaging over selective naive Bayes is that Equation (2) can be approximated in terms that only depend on each marker (X_i) or on each marker and the population membership (X_i and C).

On the one hand, part 1 in Equation (2) can be approximated by the *maximum a posteriori* (MAP) parameter configuration:

$$p(c_j, \boldsymbol{x}|D^{(t)}, S) \approx \frac{\alpha_{C-j} + E(N_{C-j}|\boldsymbol{\Theta}^{(t)})}{\alpha_C + N} \prod_{i=1}^{n} \frac{\alpha_{ijk} + E(N_{ijk}|\boldsymbol{\Theta}^{(t)})}{\alpha_{ij} + E(N_{ij}|\boldsymbol{\Theta}^{(t)})}$$

$$= \tilde{\theta}^S_{C-j} \prod_{i=1}^{n} \tilde{\theta}^S_{ijk} \tag{3}$$

where $\tilde{\theta}^S_{ijk}$ and $\tilde{\theta}^S_{C-j}$ denote the MAP parameter configuration for a selective naive Bayes structure S. Additionally, $\alpha_{ij} = \sum_k \alpha_{ijk}$ and $E(N_{ij}|\boldsymbol{\Theta}^{(t)}) = \sum_k E(N_{ijk}|\boldsymbol{\Theta}^{(t)})$, and similarly for values related to C where $\alpha_C = \sum_j \alpha_{C-j}$. Note that S determines if a marker X_i is dependent on or independent of C. Therefore, if S determines that X_i is independent of C, we should use $\tilde{\theta}^S_{i-k}$ and $E(N_{i-k}|\boldsymbol{\Theta}^{(t)})$ instead of $\tilde{\theta}^S_{ijk}$ and $E(N_{ijk}|\boldsymbol{\Theta}^{(t)})$ respectively in Equation (3), and substitute $E(N_{i-}|\boldsymbol{\Theta}^{(t)})$ for $E(N_{ij}|\boldsymbol{\Theta}^{(t)})$ with $E(N_{i-}|\boldsymbol{\Theta}^{(t)}) = \sum_k E(N_{i-k}|\boldsymbol{\Theta}^{(t)})$.

On the other hand, the marginal likelihood (part 2 in Equation (2)) can also be written in terms that only depend on X_i or on X_i and C. This is given by the well-known close formula for $p(D|S)$ (Cooper and Herskovits, 1992) adapted to our specific problem. The reader may pay attention to the fact that, while $p(D|S)$ is resolvable in closed form when there are no missing values and the clustering assignation is known (the dataset is complete), in our case, $D^{(t)}$ is not a complete dataset, therefore we are not able to calculate the sufficient statistics N_{ijk} and N_{C-j} but only approximations given the current model $\boldsymbol{\Theta}^{(t)}$. Hence, the adaptation of Cooper and Herskovits' formula (Cooper and Herskovits, 1992) gives an approximation to $p(D^{(t)}|S)$.

$$p(D^{(t)}|S) \approx \frac{\Gamma(\alpha_C)}{\Gamma(\alpha_C + N)} \prod_j \frac{\Gamma(\alpha_{C-j} + E(N_{C-j}|\boldsymbol{\Theta}^{(t)}))}{\Gamma(\alpha_{C-j})}$$

$$\cdot \prod_{i=1}^{n} \prod_j \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + E(N_{ij}|\boldsymbol{\Theta}^{(t)}))} \prod_k \frac{\Gamma(\alpha_{ijk} + E(N_{ijk}|\boldsymbol{\Theta}^{(t)}))}{\Gamma(\alpha_{ijk})} \tag{4}$$

where $\Gamma(\cdot)$ represents the gamma function. Of course, it is S again which establishes if α_{i-k}, $E(N_{i-k}|\boldsymbol{\Theta}^{(t)})$ and $E(N_{i-}|\boldsymbol{\Theta}^{(t)})$ should be used instead of the original ones in Equation (4). The approximation given by this equation in the model averaging process has been compared to a Monte Carlo approximation, which is a more accurate and computationally expensive technique to approximate $p(D|S)$, obtaining similar results (Santafé et al., 2006).

At this point, as a consequence of parameter independence assumption, we can state that if two different selective naive Bayes structures set the same relationship between marker X_i and C (in both structures X_i is relevant or irrelevant for inferring population structure) the calculations related to marker X_i in Equations (3) and (4) are the same for both structures. This is essential for an efficient model averaging calculation since it allows to eliminate the dependence on S in the calculations performed in Equation (2). That is, using the approximations given by Equations (3) and (4) in Equation (2), and grouping these calculations in terms that depend on each marker X_i, each one of these groups will contain only two kinds of terms: the ones that consider X_i relevant for clustering, ρ_{ijk}, and the ones that consider X_i irrelevant for clustering, ρ_{i-k}. Therefore, the model averaging calculations from Equation (2) can be approximated as follows:

$$p(c_j, \boldsymbol{x}|D^{(t)}) \approx \rho_{C-j} \prod_{i=1}^{n} (\rho_{i-k} + \rho_{ijk}) \tag{5}$$

where ρ_{C-j} is the term which groups the calculations related only to the cluster variable. Thus, the resulting model of the model averaging process is a naive Bayes and its parameters at $t+1$ step are given by:

$$\theta_{ijk}^{(t+1)} \propto \rho_{i-k} + \rho_{ijk} \qquad \theta_{C-j}^{(t+1)} \propto \rho_{C-j}$$

See Santafé et al. (2006) for complete details of the MA step.

2.2. *Multi-start EMA*

The EMA is a greedy algorithm that is likely to be trapped in a local optima. The results obtained by the algorithm depend on the random initialization of the parameters. Therefore, we propose the use of a multi-start algorithm where m different runs of the algorithm with different random initializations are performed. In Santafé et al. (2006), different criteria to obtain the final model from the multi-start process are proposed.

In our case, we use the best choice multi-start EMA, where the best model, in terms of likelihood, among the m models calculated by the multi-start process, is selected to be the final model.

2.3. *Selecting the most relevant markers for population inference*

The model averaging process performed by the EMA algorithm can also be seen as an implicit unsupervised feature selection that is incorporated in the final model. In fact, although the EMA algorithm, and consequently the multi-start EMA, obtains a naive Bayes model where all the maker loci are independent given population assignment, the parameters of the resultant model are calculated by a model averaging over selective naive Bayes. Thus, these parameters should reflect the significance of each marker for the inference of population structure.

In this section we propose a two-step test that can be used to obtain information about relevant markers that is implicitly contained in the final naive Bayes model calculated by the multi-start EMA. This test is based on mutual information. It is known (Cover and Thomas, 1991) that the statistic $2NI(X_i, C)$, where $I(X_i, C)$ is the mutual information between X_i and C, asymptotically follows a Chi-square probability distribution with $(r_i - 1)(r_C - 1)$ degrees of freedom. In our case, r_i is the number of different alleles that a marker X_i can present and r_C the number of clusters.

The mutual information between a marker X_i and the cluster variable, or the mutual information between two markers X_i and $X_{i'}$ with $i \neq i'$ can be calculated using the naive Bayes model obtained by the multi-start EMA. Thus, a Chi-square test can be performed to decide which marker loci are relevant for the clustering process. In the first step, a test threshold p_{rel} is set and a Chi-square test is used to filter out those markers which are considered not relevant for clustering purposes. This first step selects the relevant markers but the set of selected markers may contain redundant information. As a consequence, we develop a second step to filter out redundant information by again using a Chi-square test with a test threshold p_{red}. In this second step the pairwise mutual information of the markers selected in the first step is calculated and is used to decide whether or not two markers are redundant. As described below in Figure 2, the two-step algorithm is used to obtain the set of markers, $\boldsymbol{X}_{rel}$, which are relevant to obtain the underlying population structure.

The thresholds p_{rel} and p_{red} can be used to control the number of selected markers. On the one hand, the higher the p_{rel} value is, the more markers are selected as relevant for clustering. On the other hand, as p_{red} decreases, the number of markers considered redundant increases and therefore, the final number of selected markers is smaller.

2.4. *Number of clusters and genetic distance*

The multi-start EMA algorithm requires the specificaion of the number of clusters underlying the population. Since the real number of groups is usually unknown, we propose to investigate the number of subpopulations by evaluating runs of the algorithms with a different number of clusters. The different configurations for the number of clusters can be compared by using the genetic distance F_{ST}. This genetic distance is a measure of the dissimilarity of genetic material between different species or individuals of

$\boldsymbol{X}_{rel} = (X_1, \ldots, X_n)$

– STEP 1 –

for $i = 1$ **to** n
  if $2NI(X_i, C) < \chi^2_{(r_i-1)(r_C-1);1-p_{rel}}$
    remove X_i from $\boldsymbol{X}_{rel}$
  end if
end for

– STEP 2 –

for all $X_i, X_{i'}$ **with** $X_i, X_{i'} \in \boldsymbol{X}_{rel}$ **and** $i \neq i'$
  if $2NI(X_i, X_{i'}) < \chi^2_{(r_i-1)(r_{i'}-1);1-p_{red}}$
    if $I(X_i, C) < I(X_{i'}, C)$
      remove X_i from $\boldsymbol{X}_{rel}$
    else
      remove $X_{i'}$ from $\boldsymbol{X}_{rel}$
    end if
  end if
end for

FIG. 2. Pseudo-code for marker selection algorithm.

the same species (Reynolds et al., 1983; Weir, 1996), and it can be interpreted as the proportion of the total genetic variance contained in subpopulations relative to the total genetic variance.

F_{ST} metric is computed as $F_{ST} = -\ln(1 - \frac{\sum_i a_i}{\sum_i (a_i + b_i)})$. Note that, $\frac{\sum_i a_i}{\sum_i (a_i + b_i)}$ is the estimator for coancestry coefficient usually named as $\hat{\theta}$. However, we decide not to use the classical notation for the coancestry coefficient since it is against the definition of the parameters for the model used to infer the population structure. The calculations of a_i and b_i are taken from Reynolds et al. (1983) and adapted to our specific notation.

$$a_i = \frac{2\left[\sum_j N\theta_{C-j} \sum_k \left(\theta_{ijk} - \sum_j \theta_{ijk}\right)^2 - b_i(r_C - 1)\right]}{2N(r_C - 1)\left(1 - \sum_j \theta^2_{C-j}\right)}$$

$$b_i = \frac{2\sum_j N\theta_{C-j} \quad 1 - \sum_k \theta^2_{ijk}\Big)\Big)}{2N - r_C}$$

The F_{ST} distance gives some intuition about how far the analyzed subpopulations are. Certainly, the quantity can range from 0 to 1 and it increases as the sample allele frequencies of individuals from different subpopulations diverges. Since we aim to obtain clusters which represents well differentiated populations, the F_{ST} may be a good metric to evaluate the quality of the clustering partition from a biological point of view. Therefore, in order to decide the number of clusters, we propose to compare the mean F_{ST} metric over a set of ten independent runs among experiments with several numbers of clusters. Additionally, it is possible to perform a Mann-Whitney test to decide if the differences in the F_{ST} metric are statistically significant. Thus, we may be able to decide the proper number of clusters in the dataset.

3. DATASET

Data were taken from the HGDP-CEPH human genome diversity cell line panel. The diversity panel is a large and widely used collection of DNA samples from individuals distributed around the world. The properties of the sample of individuals were first reported by Cann et al. (2002). However, new genotypes have been reported since then. Specifically, we employ a dataset included in the HGDP-CEPH which has been used in recent studies of genetic structure of human populations (Rosenberg et al., 2005). The dataset contains 993 markers, including 783 microsatellites and 210 insertion/deletion polymorphisms corresponding to 1048 individuals from 53 different human population distributed around the world. Although the individuals are classified into 53 human populations or ethnic groups, they correspond to seven major regions: Africa, Europe, Middle East, Central/South Asia, East Asia, Oceania, and America.

4. RESULTS AND DISCUSSION

The aim of the study is to group individuals into genetic clusters in such a way that each individual is given an estimated membership coefficient for each cluster. This probability can be seen as an admixture coefficient since each individual may have genetic information belonging to different population sources.

The proposed algorithm is compared with the well-known STRUCTURE software (Falush et al., 2003; Pritchard et al., 2000b), version 2.1. Both the multi-start EMA and STRUCTURE require a prespecified number of clusters, therefore, we run experiments with two, three and four numbers of clusters. Since the obtained clustering results depend on the random initialization of the models, we run ten executions of each algorithm with each number of clusters. For the multi-start EMA we use a $\epsilon = 0.01$ and 500 iterations in the multi-start process, $m = 500$. For STRUCTURE, we use the configuration reported in Rosenberg et al. (2005), where an allele frequency correlated model is used. Additionally, the parameters are calculated using 1000 iterations after a burn-in period of 5000.

Figure 3 shows the estimation of population structure obtained by the multi-start EMA with two, three and four clusters ($r_C = 2$, $r_C = 3$, and $r_C = 4$). Similarly, Figure 4 shows the results obtained by STRUCTURE. The plots were generated with DISTRUCT (Rosenberg, 2004), where each individual is represented by a segment partitioned into r_C colored parts that represents the estimated membership of the individual to each one of the r_C clusters. For each number of clusters, only the best run of ten on the basis of the F_{ST} measure is shown. In these experiments, the multi-start EMA tends to assign each individual to a cluster with a very high probability being the membership probability for the rest of the clusters very small. Thus, the admixture proportions detected by multi-start EMA are also very small. By contrast, STRUCTURE detects a higher level of admixture among populations. The results obtained by STRUCTURE may be biologically correct since admixture is usual in population genetics. However not always detected admixture proportions represents genuine contributions from corresponding ancestral sources since the uncertainty about the allele frequencies in two particular source populations may cause the overestimation of the admixture proportions (Corander and Marttinen, 2006). Moreover, although the admixed ancestry detected by multi-start EMA is very small, the genetic distance given by the F_{ST} metric[2] is much higher (Tables 1 and 2). This suggests that the populations obtained by the multi-start EMA are genetically more distant between themselves than the populations obtained by STRUCTURE.

On the other hand, the population structure obtained by the multi-start EMA mainly correspond to major geographical regions. Nevertheless, it is surprising that most of the individuals from Uygur and Hazara ethnic group are classified in the same cluster as ethnic groups from East Asia even when they are located in Central/South Asia. Similarly, a few Mozabite individuals are clustered into a group dominated

[2]Since the dataset includes two different types of polymorphisms, F_{ST} genetic distance is calculated taking into account only genetic information regarding microsatellite markers and therefore, insertion/deletion polymorphisms are ignored.

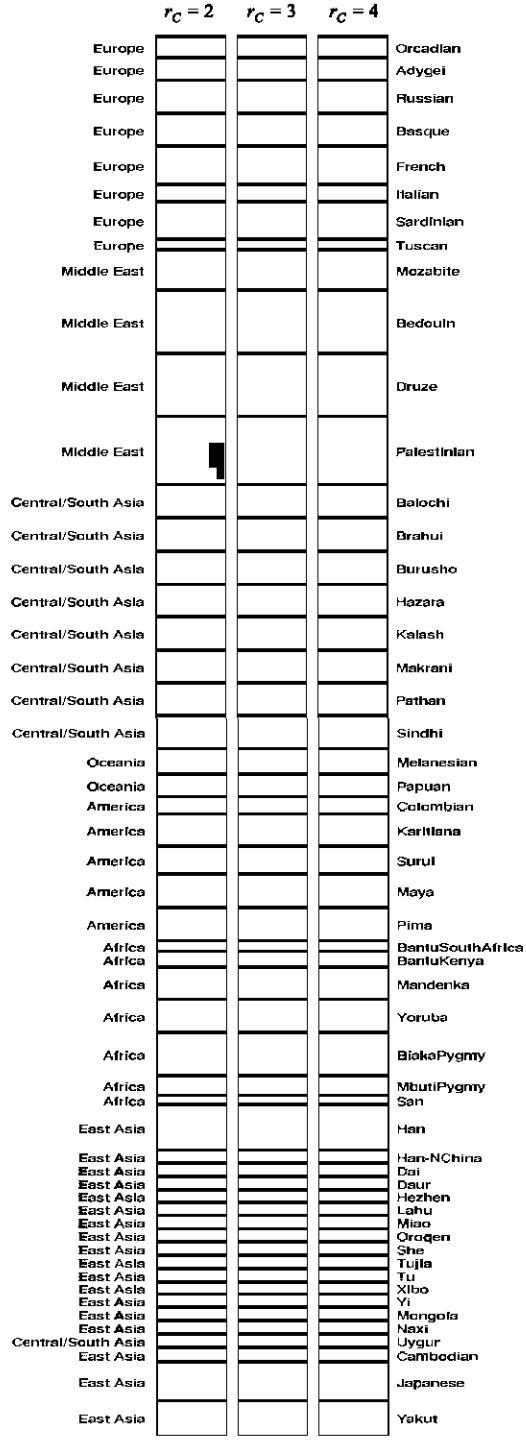

FIG. 3. Inferred population structure using the multi-start EMA.

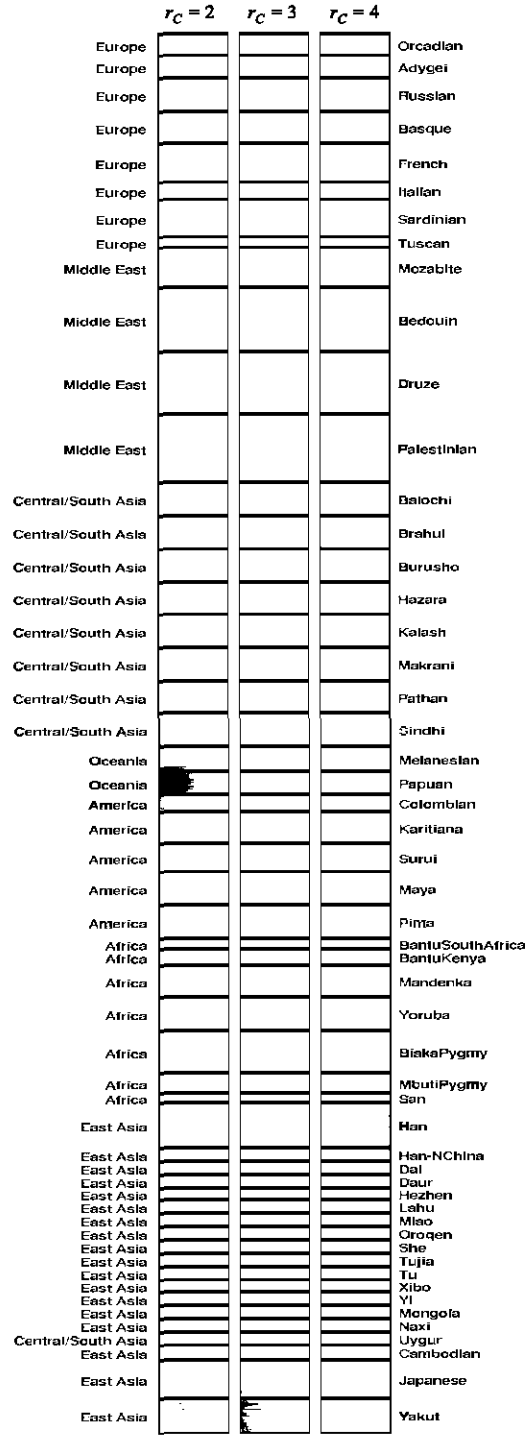

FIG. 4. Inferred population structure using STRUCTURE.

TABLE 1. F_{ST} VALUES OBTAINED ON TEN RUNS OF THE MULTI-START EMA WITH TWO, THREE, AND FOUR CLUSTERS ($r_C = 2$, $r_C = 3$, AND $r_C = 4$)

	$r_C = 2$	$r_C = 3$	$r_C = 4$
Run 1	0.02714400	0.03193830	0.03100300
Run 2	0.02714400	0.03193830	0.02937500
Run 3	0.02714400	0.03193830	0.03463900
Run 4	0.02714400	0.03193830	0.02989200
Run 5	0.02715100	0.03156800	0.03100300
Run 6	0.02714400	0.03193830	0.03100300
Run 7	0.02714400	0.03193830	0.03100300
Run 8	0.02714400	0.03541020	0.03100300
Run 9	0.02714400	0.03193830	0.03100300
Run 10	0.02715200	0.03193830	0.03905100
Mean	0.02714550	0.03224446	0.03189750
STD	0.00000317	0.00111700	0.00286500

The mean and standard deviation values over the ten runs are also reported.

by Africans. This situation may be apparently strange. However, these results obtained by the multi-start EMA agree with those obtained by STRUCTURE, where the membership coefficients of Hazara and Yaruba individuals are higher for the cluster dominated by East Asia individuals than for the cluster dominated by Central/South Asia individuals.

4.1. Number of clusters in the dataset

As it was stated before, the multi-start EMA and STRUCTURE assume that the number of clusters, r_C, is given. However, this is not usually true in real problems. In order to select the best number of clusters, we evaluate the results with two, three, and four clusters. We have also extended the experiments to more clusters but the multi-start EMA algorithm, in these experiments, converges to different solutions in separate runs when the number of clusters is higher than four. Actually, a few runs with five clusters

TABLE 2. F_{ST} VALUES OBTAINED ON TEN RUNS OF STRUCTURE WITH TWO, THREE, AND FOUR CLUSTERS ($r_C = 2$, $r_C = 3$, AND $r_C = 4$)

	$r_C = 2$	$r_C = 3$	$r_C = 4$
Run 1	0.00069648	0.00315150	0.00579690
Run 2	0.00069517	0.00315350	0.00579590
Run 3	0.00069517	0.00392830	0.00579690
Run 4	0.00069517	0.00311270	0.00579690
Run 5	0.00069517	0.00315350	0.00579690
Run 6	0.00069517	0.00316680	0.00579690
Run 7	0.00069517	0.00313840	0.00579690
Run 8	0.00067972	0.00315350	0.00579690
Run 9	0.00069517	0.00315350	0.00582990
Run 10	0.00069517	0.00315350	0.00579690
Mean	0.00069375	0.00322652	0.00580010
STD	0.00000000	0.00024699	0.00001047

The mean and standard deviation values over the ten runs are also reported.

yield similar results than with four clusters but a new group is created with individuals of American origin (data not shown). These results are also similar to those obtained by Rosenberg et al. (2002, 2005) with five clusters. Nevertheless, most of the runs with five and six clusters results in a partition with four clusters (similar to the one shown in Figure 1 with $r_C = 4$) and one or two empty clusters, respectively. Therefore, the multi-start EMA is detecting that there is no more than four clear clusters.

In order to decide the number of clusters, we use the F_{ST} distance. Table 1 shows the F_{ST} values over ten independent runs of the multi-start EMA for each number of clusters (two, three, and four). The best mean F_{ST} value corresponds to the experiments with three clusters. Clustering the data into four groups produces, except for run 10, slightly lower F_{ST} values than with three clusters. However, the difference of the F_{ST} values with three and four clusters is not statistically very significant (the p-value of the Mann-Whitney test is 0.72). By the contrary, the difference between F_{ST} values with two and three clusters is statistically significant (p-value = 0.00). Therefore, we think that according to the multi-start EMA and the proposed metric, there are three or four clear clusters underlying the dataset. These results are compatible with those presented in Rosenberg et al. (2005), where the quality of the clustering partition is given by the clusteredness.[3] Although Rosenberg et al. (2005) also consider the presence of more than four clusters, and the best partition, on the basis of the clusteredness, is obtained with only two clusters, the populations obtained with three and four clusters are considered as good partitions too.

4.2. *Selection of relevant markers*

The use of the multi-start EMA algorithm allows to select the most relevant markers needed to obtain the clustering partition. The number of selected markers is controlled by two parameters, p_{rel} and p_{red}, which represent the thresholds for the statistical tests, being p_{rel} the threshold to control the selection of relevant markers and p_{red} the threshold to control the redundancy between the selected markers. STRUCTURE software is not able to filter out irrelevant or redundant markers and the presence of this irrelevant and/or redundant information may damage its ability to obtain the clustering partition. In order to show how STRUCTURE software behaves in these situations and how the information about irrelevant and redundant markers provided by the multi-start EMA helps to obtain a better clustering partition, we proceed as follows: first, we select, for each number of clusters, the best model on the basis of the F_{ST} metric (these are the models used to obtain the population partitions represented in Fig. 3). Then, we perform a selection of the most relevant markers by using the maker selection algorithm where the parameter of the test p_{rel} varies in $\{0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$ and p_{red} in $\{0, 0.01, 0.05, 0.1\}$. Each parameter configuration gives rise to a different set of selected markers (Fig. 5). Finally, for each set of selected markers we run ten independent executions of STRUCTURE and measure the mean value over the ten runs of the F_{ST} metric.

Figure 6 shows the mean F_{ST} value for each one of the selected marker sets. A clear trend can be observed in the plots: marker selection using small values of p_{rel} and p_{red} improves the F_{ST} values obtained by STRUCTURE. Additionally, as the value of p_{rel} increases, the number of selected markers also increases including more redundant information. Consequently, the mean value of the F_{ST} metric decreases approaching the F_{ST} value obtained by STRUCTURE with the whole set of markers. It should be noted that the sets of selected markers are used only to obtain the clustering partition with STRUCTURE, but the F_{ST} metric is always calculated taking into account all the microsatellite markers.

According to the experimental results, we can say that the information about relevant markers implicitly included in the model calculated by the multi-start EMA helps to obtain the clustering partition. STRUCTURE, as well as other algorithms for the inference of population structure, does not take into account the existence of irrelevant and redundant information in the dataset. Therefore, the presence of irrelevant and/or redundant information damages their ability to retrieve the population structure underlying the dataset.

[3] Clustering quality metric that measures the extent to which individuals were estimated to belong to a single cluster rather than to a combination of clusters.

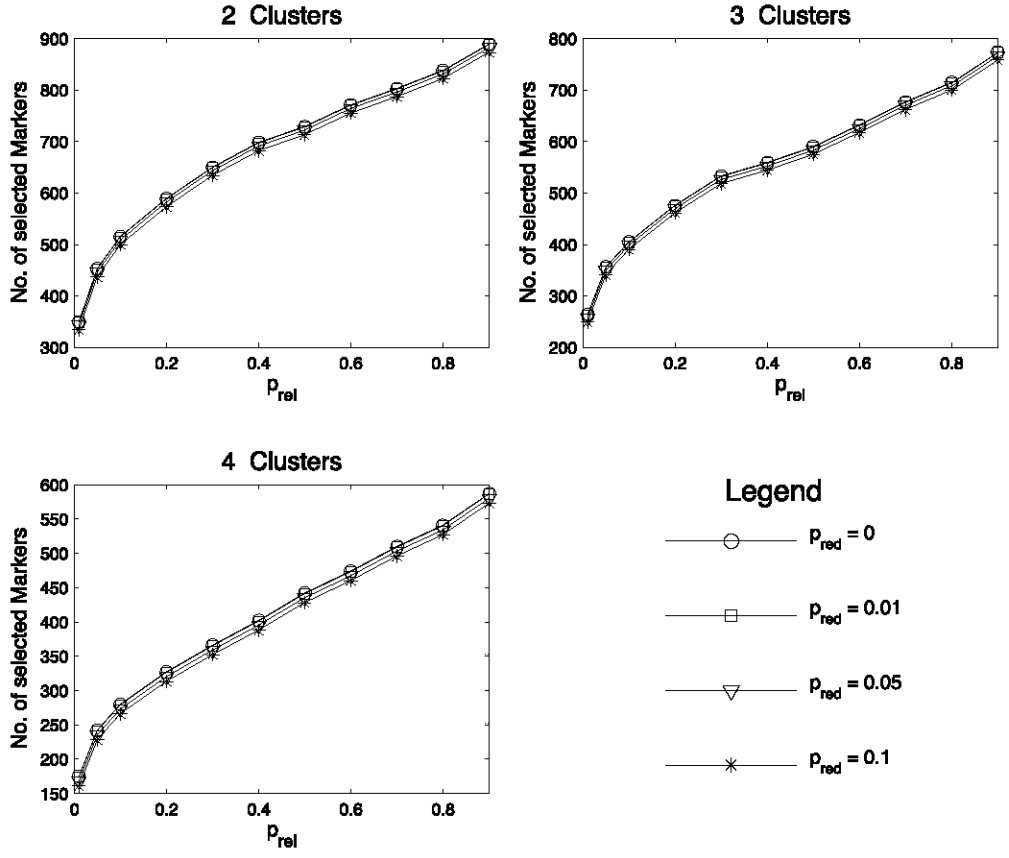

FIG. 5. Number of selected markers.

5. CONCLUSIONS

The inference of population structure is a very important and highly studied problem in population genetics. In the current paper we describe a new approach to learn about population structure and assign individuals (probabilistically) to populations. This new approach uses multilocus genotype data and a Bayesian model averaging technique to obtain the population (probabilistic) partition and the allele frequencies. The method, named multi-start EMA, was originally proposed in Santafé et al. (2006), but in this work, it has been tailored in order to work with polyploid and missing data. Thus, we provide a useful and realistic tool for the inference of population structure.

One of the main drawbacks of the multi-start EMA, as well as other popular algorithms for the inference of population structure, is the fact that the number of clusters has to be fixed in advance. However, we provide a method to investigate the number of groups underlying the data and thus overcome this multi-start EMA restriction. Alternatively, the main advantage of the proposed algorithm is its skill at dealing with irrelevant data for clustering purposes. This irrelevant information may damage the ability of other methods to obtain the underlying population structure. Moreover, we propose a marker selection algorithm based on mutual information which is able to obtain the most relevant markers needed to retrieve the population structure.

The performance of the multi-start EMA is evaluated in a real problem and compared with STRUCTURE, which is the most widely used software for the inference of population structure. The results from the experiments show that the populations obtained by the multi-start EMA have higher values for the F_{ST} metric than populations obtained by STRUCTURE. This suggests that populations obtained by the multi-start EMA are genetically more distant than populations obtained by STRUCTURE. By contrast, in the experiments, the multi-start EMA is not able to obtain population partitions with more than four clusters. It may suggest that the multi-start EMA does not perform as well as STRUCTURE when subpopulations of individuals in the dataset are genetically very close.

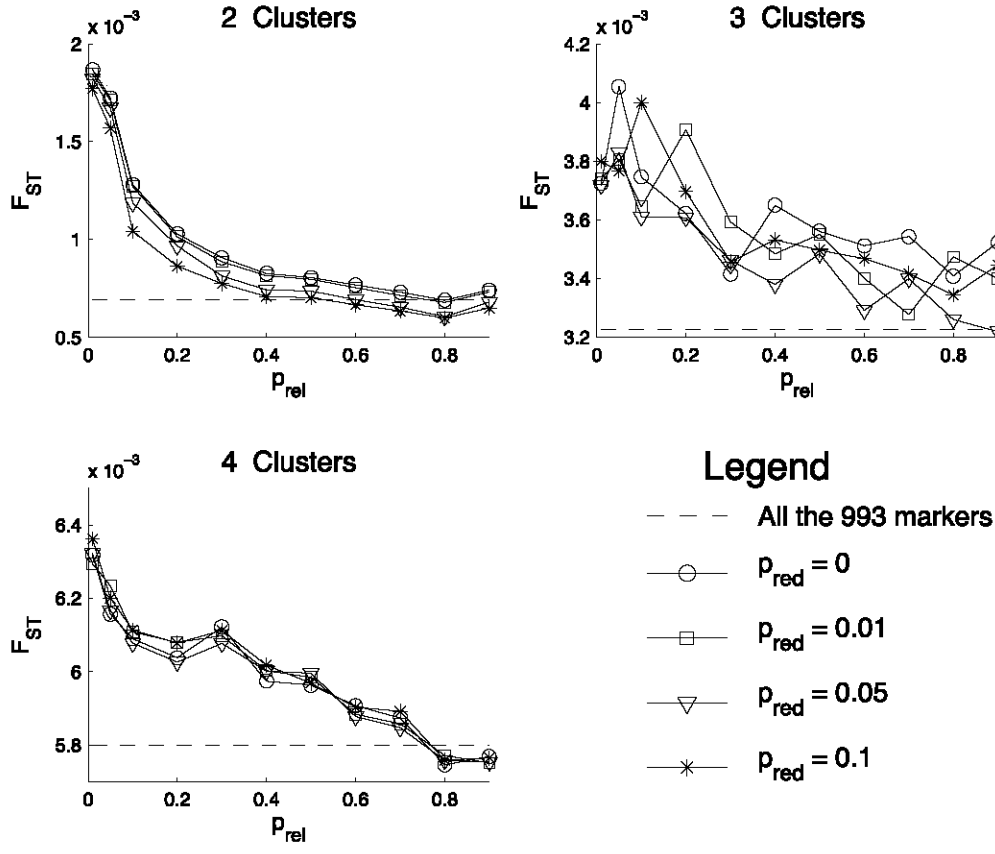

FIG. 6. Evolution of F_{ST} metric in the marker selection process.

ACKNOWLEDGMENTS

The authors wish to thank Professor Jose Miguel-Alonso for providing computing resources, funded by the Spanish Ministerio de Educación y Ciencia (TIN2004-07440-C02-02). This work was supported by the SAIOTEK-Autoinmune (II) 2006 and Etortek research projects from the Basque Government. It has been also supported by the Spanish Ministerio de Educación y Ciencia (grant TIN 2005-03824) and by the Government of Navarra (under a Ph.D. grant awarded to the first author).

Bamshad, M.J., Wooding, S., Watkins, W.S., et al. 2003. Human population genetic structure and inference of group membership. *Am. J. Hum. Genet.* 72, 578–589.

Barash, Y., and Friedman, N. 2002. Context-specific Bayesian clustering for gene expression data. *J. Comput. Biol.* 9, 169–191.

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

Cann, H.M., de Toma, C., Cazes, L., et al. 2002. A human genome diversity cell line panel. *Science* 296, 261–262.

Cardon, L.R., and Palmer, L.J. 2003. Population stratification and spurious allelic association. *Lancet* 361, 598–604.

Cooper, G.F., and Herskovits, E. 1992. A Bayesian method for the induction of probabilistic networks from data. *Mach. Learn.* 9, 309–347.

Corander, J., and Marttinen, P. 2006. Bayesian identification of admixture events using multilocus molecular markers. *Mol. Ecol.* 15, 2833–2843.

Corander, J., Waldmann, P., Marttinen, P., et al. 2004. BAPS 2: enhanced possibilities for the analysis of genetic population structure. *Bioinformatics* 20, 2363–2369.

Cover, T.M., and Thomas, J.A. 1991. *Elements of Information Theory.* John Wiley & Sons, New York.

Dash, D., and Cooper, G.F. 2004. Model averaging for prediction with discrete Bayesian networks. *J. Mach. Learn. Res.* 5, 1177–1203.

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

Dempster, A.P., Laird, N.M., and Rubin, D.B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. B* 39, 1–38.

Falush, D., Stephens, M., and Pritchard, J.K. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 164, 1567–1587.

Guillot, G., Estoup, A., Mortier, F., et al. 2004. A spatial statistical model for landscape genetics. *Genetics* 170, 1261–1280.

Kollin, J., and Koivisto, M. 2006. Bayesian learning with mixtures of trees. *Proc. 17th Eur. Conf. Mach. Learn.* 294–305.

O'Rourke, S., Chechik, G., and Eskin, E. 2005. Separation of overlapping subpopulation by mutual information. *Proc. NIPS Workshop Comput. Biol. Anal. Heterogeneous Data.*

Patterson, N., Hattangadi, N., Lane, B., et al. 2004. Methods for high-density admixture mapping of disease genes. *Am. J. Hum. Genet.* 74, 1001–1013.

Price, A.L., Patterson, N.J., Plenge, R.M., et al. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38, 904–909.

Pritchard, J., Stephens, M., Rosenberg, N., et al. 2000a. Association mapping in structured populations. *Am. J. Hum. Genet.* 67, 170–181.

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

Reynolds, J., Weir, B.S., and Cockerham, C.C. 1983. Estimation of the coancestry coefficient: basis for a short-term genetic distance. *Genetics* 105, 767–779.

Riddle, E.L., Murthy, K.K., Eskin, E., et al. 2006. Single nucleotide polymorphisms and ethnic-specific haplotypes of the regulator of G-protein signaling-2 (rgs2) gene in hypertensive and normotensive subjects. *Hypertension* 47, 415–420.

Rosenberg, N.A. 2004. Distruct: a program for the graphical display of population structure. *Mol. Ecol. Notes* 4, 137–138.

Rosenberg, N.A. 2005. Algorithms for selecting informative marker panels for population assignment. *J. Comput. Biol.* 12, 1183–1201.

Rosenberg, N.A., Li, L.M., Ward, R., et al. 2003. Informativeness of genetic markers for inference of ancestry. *Am. J. Hum. Genet.* 73, 1402–1422.

Rosenberg, N.A., Mahajan, S., Ramachandran, S., et al. 2005. Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS Genet.* 1, 661–671.

Rosenberg, N.A., Pritchard, J.K., Weber, J.L., et al. 2002. Genetic structure of human population. *Science* 298, 2381–2385.

Santafé, G., Lozano, J.A., and Larrañaga, P. 2006. Bayesian model averaging of naive Bayes for clustering. *IEEE Trans. Syst. Man Cybernet. B* 36, 1149–1161.

Sillanpää, M.J., Kilpikari, R., Ripati, S., et al. 2001. Bayesian association mapping for quantitative traits in a mixture of two populations. *Genet. Epidemiol.* 21, S692–S699.

Turakulov, R., and Easteal, S. 2003. Number of SNPs loci needed to detect population structure. *Hum. Hered.* 55, 37–45.

Weir, B. 1996. *Genetic Data Analysis II: Methods for Discrete Population Genetic Data*, 2nd ed. Sinauer Associates, Sunderland, MA.

Wu, B., Liu, N., and Zhao, H. 2006. Psmix: An r package for population structure inference via maximum likelihood method. *BMC Bioinform.* 7, 317.