April 2005

Bayesian models for the analysis of genetic structure when populations are correlated

Rongwei Fu
University of Connecticut

Dipak Dey
University of Connecticut, DIPAK.DEY@uconn.edu

Kent E. Holsinger
University of Connecticut, kent.holsinger@uconn.edu

Bayesian models for the analysis of genetic structure when populations are correlated

Rongwei Fu* and Dipak K. Dey
Department of Statistics, U-4120
University of Connecticut
Storrs, CT 06269-4120

Kent E. Holsinger
Department of Ecology & Evolutionary Biology, U-3043
University of Connecticut
Storrs, CT 06269-3043

September 28, 2004

*Current address: Department of Public Health and Preventive Medicine, #CB669, Oregon Health & Science University, Portland, OR, 97239

ABSTRACT

Motivation: Populations are often correlated due to shared history or gene exchange. However, consideration of such correlation has not been adequate and satisfactory in probability models for allele frequency and inference about population structure. Recent study shows that correlation among populations could be very high, which in turn affect the estimate of measures of genetic variation. In this study we propose a mixture beta model to characterize allele frequency and incorporate the correlation among populations into account as well as extending the model to data with different clusters.

Results: Using simulated data, we show that in general, the mixture model provides a good approximation of the allele frequency and a good estimate of correlation among populations. Results from fitting the mixture model to a data set of phenotypes at 377 autosomal microsatellite loci from human populations indicates high correlation among populations, which may not be appropriate to neglect. Also traditional measures of population structure tend to overestimate the amount of genetic differentiation when correlation is neglected. Inference is performed in a Bayesian framework.

Contact: fur@ohsu.edu

INTRODUCTION

Organisms usually form different populations in various habitats. Inevitably some genetic differentiation occurs and there are differences of allele frequencies among these populations. It has been one of the main interests throughout the history of theoretical population genetics to describe and understand the nature of genetic differentiation observed in natural populations. Furthermore, advances in molecular technology in the last decade have resulted in large amounts of data for accessing genetic variation. On the other hand, populations are often correlated due to shared history or gene flow. Fu *et al.* (2003) derived equations to solve for correlation in allele frequencies for a set of populations subject to drift, mutation and migration and developed exact expressions of correlation for several special cases under finite island model. Indeed, the correlation in allele frequencies among populations can be very large for realistic rates of mutation and migration unless an enormous number of populations are exchanging genes.

However, consideration of such correlation has not been adequate and satisfactory in probability models for allele frequency and often ignored when genetic differentiation is accessed by using Wright's F_{ST} (Wright, 1951) or similar measures. Balding and Nichols (1995) developed a beta distribution to describe allele frequency at biallelic loci. This beta distribution and its multiallelic version have been widely used to make inference of genetic differentiation and population structure in both likelihood-based and Bayesian approaches (Balding and Nichols, 1995; Balding and Nichols, 1997; Roeder *et al.*, 1998; Holsinger, 1999; Holsinger *et al.*, 2002; Falush *et al.*, 2003). Nicholson *et al.* (2002) proposed a truncated normal model to describe the allele frequency for single nucleotide polymorphisms. For both models, the populations have been interpretated as dependent in the sense

that the mean of allele frequencies from a set of populations is assumed to be the allele frequency of a hypothetical "ancestral" population and the populations have each diverged from the "ancestral" population. Hence, the populations are related to one another by having the common "ancestral" population. However, these models does not incorporate the correlation across populations induced by gene flow. Actually it does not build a correlation across populations *statistically* by assuming a common "ancestral" population, as shown later. As a result, estimates of genetic differentation based on these models do not incorporate the correlation among populations. Beerli and Felsenstein (1999) estimated migration rates and effective population numbers by using a coalesent approach. In their model, the correlation across populations was implicitly accounted for, but they did not estimate the magnitude of the correlation.

Unfortunately, the correlation across populations affects the estimates of genetic differentation. Fu et al. (2003) discussed some implications of this correlation for measures of genetic differentation based on Wright's F_{ST} (Wright, 1969), and showed that when populations are small, the estimates of population structure measures are remarkably different depending on whether the correlation is incorporated or not. Nicholson *et al.* (2002) also expressed their concern, more than once, that the correlation across populations due to shared history or gene flow is not typically accounted for. Therefore, in this study we propose a new mixture beta model to approximate the allele frequency in which the correlation among populations induced by shared history or gene flow is incorporated and explicitly estimated. The allele frequency at each locus in any population is described by the sum of two beta variables in which one of them is common across populations. In general, such a mixture of beta distributions forms a very rich class of distributions (Diaconis and Ylvisaker, 1985). We also extend our approach to genetic data with clusters. The performance of the model

is evaluated by using three sets of simulated data from finite island model and illustrated by a real data set with phenotypes at 377 autosomal microsatellite loci from 52 human populations. All the analysis are done in a Bayesian framework.

MIXTURE BETA MODEL FOR ALLELE FREQUENCY

Assume that we have allele frequencies in K populations at I loci, each locus having two alleles A_1 and A_2. Let $\mathbf{p}_{I\times K}$ denote the allele frequencies of A_1, i.e., the ikth element of $\mathbf{p}$, p_{ik}, is the allele frequency of A_1 at locus i in population k, $i = 1, \cdots, I$; $k = 1, \cdots, K$. It is sufficient to work only with $\mathbf{p}$ since the allele frequencies of A_1 and A_2 sum to 1. To incorporate the correlation among populations into the analysis, we describe allele frequency p_{ik} by using the following mixture model

$$p_{ik} = wx_{ik} + (1 - w)y_i, \tag{1}$$

where x_{ik} is a set of independent beta variates and y_i is another set of independent beta variates. Here w is the mixture coefficient of the two beta variates, and a number between 0 and 1. Further x_{ik} and y_i are assumed to be independent from each other. That is, for any locus, the allele frequency at each population could be expressed as the weighted sum of two independent components, an individual component for that population (x_{ik}) and a common (y_i) component across all populations. Thus we build correlation among populations through the common component y_i. Heuristically, the common component could be considered as the contribution of shared history or gene flow to allele frequency, as both shared history or gene flow make populations more similar to each other. Smaller w is expected to be associated with high correlation. Note that we assume a

common w across all loci. It is possible to specify a different w for each locus, namely w_i, to have a more flexible model, but we do not pursue this possibility. Precisely estimating w_i for each locus needs information from a large number of populations, which may not be available in most studies.

For each locus, we assume that they have the same probabilistic structure and focus on loci that have been subjected to similar evolutionary process. In particular, we assume

$$\begin{aligned} x_{ik} &\sim \text{Beta}(\frac{1-\theta^x}{\theta^x}\pi_k, \frac{1-\theta^x}{\theta^x}(1-\pi_k)), \\ y_i &\sim \text{Beta}(\frac{1-\theta^y}{\theta^y}\pi, \frac{1-\theta^y}{\theta^y}(1-\pi)), \end{aligned} \tag{2}$$

where θ^x, θ^y, π_k, $k = 1, \cdots, K$, and π are all between 0 and 1. It follows that

$$\begin{aligned} \text{E}(p_{ik}) &= w\pi_k + (1-w)\pi, \\ \text{Var}(p_{ik}) &= w^2\theta^x\pi_k(1-\pi_k) + (1-w)^2\theta^y\pi(1-\pi), \\ \text{Cov}(p_{ik}, p_{i\prime k}) &= 0, \\ \text{Cov}(p_{ik}, p_{ik'}) &= (1-w)^2\theta^y\pi(1-\pi), \\ \text{Cov}(p_{ik}, p_{i\prime k\prime}) &= 0. \end{aligned} \tag{3}$$

In other words, (3) shows that the same loci from different populations are correlated but different loci from the same or different populations are not. This agrees with the results from Fu *et al.* (2003) for a set of populations subject to migration, mutation and random drift for independent loci. We assume a common covariance among any pair of populations at each loci but correlation among any two populations could be different due to possible differences of π_k for different populations. Again,

it is possible to have a different covariance among any pair of populations by assuming w_i. Results from (3) also shows that our formulation only allows positive correlation among populations, which, again, agrees with results from Fu *et al.* (2003).

When $w = 1$, $p_{ik} = x_{ik}$ and (1) gives the usual beta model. The major difference between our formulation and previous ones (Balding and Nichols, 1995; Roeder *et al.*, 1998; Holsinger, 1999; Holsinger *et al.*, 2002; Falush *et al.*, 2003) is that in our formulation, $\mathrm{E}(x_{ik}) = \pi_k$, e.g., the mean allele frequency is calculated across loci for each population and in previous ones, the beta model is given by

$$p_{ik} \sim \mathrm{Beta}(\frac{1-\theta}{\theta}\pi_i, \frac{1-\theta}{\theta}(1-\pi_i)) \tag{4}$$

and $\mathrm{E}(p_{ik}) = \pi_i$ is the mean allele frequency across populations for each loci. The model of Nicholson *et al.* (2002) has similar specification and the two models agree to first and second moments. In these models, π_i is interpreted as the allele frequency in a "ancestral" populaiton from which the sampled populations have each independently diverged. Thus these populations are related by sharing the "ancestral" populaiton. Conditional on π_i, the allele frequencies are independent. However, related to a common "ancestral" populaiton does not *statistically* build correlation in allele frequencies since, marginally,

$$\begin{aligned} \mathrm{Cov}(p_{ik}, p_{ik\prime}) &= \mathrm{E}(\mathrm{Cov}(p_{ik}, p_{ik\prime}|\pi_i)) + \mathrm{Cov}(\mathrm{E}(p_{ik}|\pi_i), \mathrm{E}(p_{ik\prime}|\pi_i)) \\ &= \mathrm{Var}(\pi_i) \\ &= 0 \end{aligned} \tag{5}$$

as π_i is considered as *the* allele frequency of the ancestor population and a parameter in the beta

model. The same argument applies to the truncated normal distribution by Nicholson *et al.* (2002), and they recognized that their model does not account for correlation across populations induced by shared history or by gene glow, which could be the most likely deviation from real data.

When $\mathrm{E}(p_{ik}) = \pi_i$, estimate of θ in (4) is analogous to Weir and Cockerham's (1984) θ and is interpreted as a measure of population structure (e.g., Roeder *et al.*, 1998; Holsinger, 1999). In our formulation, $\mathrm{E}(p_{ik}) = \pi_k$. As a result, we are unable to construct a measure (as a function of θ^x, θ^y and w), which has the same interpretation as θ in (4). However, the traditional method to estimate F_{ST} typically ignores correlation across populations. When correlation is present, the goodness of the traditional estimation method is subject to more investigation.

Recall that Wright's definition (1951) of F_{ST} for one locus with two alleles is given by

$$F_{ST} = \frac{\sigma_p^2}{\mu_p(1-\mu_p)} \quad . \tag{6}$$

The parameter θ in (4) agrees with this definition, which is equivalent to an intraclass correlation coefficient. For a finite set of K populations, a natural analog of (6) is

$$F_{ST} = \frac{\sum_{k=1}^{K}(p_k - \bar{p})^2/K}{\bar{p}(1-\bar{p})} \quad , \tag{7}$$

where $\bar{p} = (1/K)\sum p_i$. Equation (7) has also been used to estimate F_{ST} using data from a sample of populations. It explicitly shows that F_{ST} measures relative reduction in heterozygosity resulting from population structure (Wright, 1943) since $\sum_{k=1}^{K}(p_k - \bar{p})^2/K$ calculates reduction in heterozygosity.

On the other hand, we must also note that the equivalency of (6) and (7) for a finite set

of populations only holds with the implicit assumption that these populations are independent. When correlation across populations (ρ) is present, $\mathrm{E}\left(\sum_{k=1}^{K}(p_k-\bar{p})^2/K\right) = \frac{K-1}{K}\sigma_p^2(1-\rho)$ for a sample of populations and $\sum_{k=1}^{K}(p_k-\bar{p})^2/K$ is not equivalent to σ_p^2 for a finite set of populations. Actually, σ_p^2 overestimates reduction in heterozygosity across populations. Similarly, Weir and Cockerham's (1984) θ estimates relative reduction in heterozygosity only when the correlation among populations is zero. In Fu *et al.*(2003), three definitions of (7) were considered. Denote the numerator on the right-hand side of (7) as Num and the denominator $Denom$, we showed that $\mathrm{E}(Num/Denom)$ is the definition that fully reflects the correlation among populations. When correlation among populations is not fully accounted for, estimates of F_{ST} overestimates the amount of genetic differentiation among populations and substantially so when populations are small. It is hard to construct a measure of population structure corresponding to $\mathrm{E}(Num/Denom)$ through parameters of the mixture beta distribution, however, we could easily calculate Bayesian estimate of $\mathrm{E}(Num/Denom)$ by plugging posterior estimate of p_{ik} into (7) for each locus. A comparison with Bayesian estimate of θ based on (4) will reveal more about how the correlation affects the F_{ST} analysis in a Bayesian framework.

BAYESIAN MODEL

Now we develop our Bayesian model using the above mixture model. For different types of genetic data, the likelihood functions are slightly different. For clarity, we specify the Bayesian model for allele frequency data and codominant data separately. Data directly on allele frequency may be available from haploid organisms or the haploid stage of diploid organisms, and from diploid organisms by simulation under genetic models. More often, data are available as the number

of different phenotypes and calculation of allele frequencies varies with the dominance types and presence or absence of inbreeding.

Modeling allele frequencies

For loci with two allele types, the number of each type at any locus is assumed to follow a binomial distribution. Consider a diploid population with N individuals, the total number of alleles at each locus is $2N$. We use $2\mathbf{N}_{A1}$ and $2\mathbf{N}_{A2}$, both $I \times K$ matrices, to denote the number of allele type A_1 and A_2 in K populations at I locus, e.g., $2\mathbf{N}_{A1,ik}$ and $2\mathbf{N}_{A2,ik}$ be the number of allele A_1 and A_2 at locus i in population k respectively. Also let $\mathbf{x}$ and $\mathbf{y}$ denote the collection of x_{ik} and y_i, $i = 1, \cdots, I$, $k = 1, \cdots, K$, respectively. If we assume that magnitude of gametic disequilibrium within populations are negligible, which is equivalent to assuming independent loci, then the likelihood is given by

$$\begin{aligned} P(2\mathbf{N}_{A1}, 2\mathbf{N}_{A2}|\mathbf{x}, \mathbf{y}, w) &\propto \prod_{i=1}^{I}\prod_{k=1}^{K} p_{ik}^{2\mathbf{N}_{A1,ik}}(1-p_{ik})^{2\mathbf{N}_{A2,ik}} \\ &\propto \prod_{i=1}^{I}\prod_{k=1}^{K} (wx_{ik}+(1-w)y_i)^{2\mathbf{N}_{A1,ik}}(1-wx_{ik}-(1-w)y_i)^{2\mathbf{N}_{A2,ik}}. \end{aligned} \tag{8}$$

To complete model specification, we use (2) as the prior distributions for x_{ik} and y_i and denote them as $P(x_{ik}|\theta^x, \pi_k)$ and $P(y_i|\theta^y, \pi)$ respectively. Let $P(\cdot)$ denote the prior distribution for any of other parameters and hyperparameters. We use uniform(0,1) for $P(\cdot)$ throughout this paper though $P(\cdot)$ could also be specified by using information from previous comparable studies, if available, like power prior (Ibrahim & Chen, 2000). Let $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_K)$, then the full conditional posterior

distribution for $\mathbf{x}$, $\mathbf{y}$, $\boldsymbol{\pi}$, w, θ^x, θ^y and π is given by

$$P(\mathbf{x},\mathbf{y},\boldsymbol{\pi},w,\theta^x,\theta^y,\pi|2\mathbf{N}_{A1},2\mathbf{N}_{A2}) \quad \propto$$
$$P(2\mathbf{N}_{A1},2\mathbf{N}_{A2}|\mathbf{x},\mathbf{y},w)\left\{\prod_{i=1}^{I}\left\{\prod_{k=1}^{K}P(x_{ik}|\pi_k,\theta^x)\right\}P(y_i|\theta^y,\pi)\right\}$$
$$\times\left\{\prod_{k=1}^{K}P(\pi_k)\right\}P(w)P(\theta^x)P(\theta^y)P(\pi). \tag{9}$$

Modeling codominant markers

Again let us assume that the sample consists of data on genetic variation in K diploid populations at I loci, each locus having two alleles A_1 and A_2. All three phenotypes A_1A_1, A_1A_2 and A_2A_2 are observable when A_1 and A_2 are codominant. Let $\mathbf{N}_{A11}$, $\mathbf{N}_{A12}$ and $\mathbf{N}_{A22}$, all $I\times K$ matrices, denote the number of phenotypes A_1A_1, A_1A_2 and A_2A_2 in the sample respectively and similarly, $\boldsymbol{\gamma}_{A11}$, $\boldsymbol{\gamma}_{A12}$ and $\boldsymbol{\gamma}_{A22}$, the frequency of three phenotypes. The numbers of different phenotypes at locus i of population k are usually described by a multinomial distribution. If we assume that phenotypes are sampled at random across loci, which is equivalent to assuming that magnitudes of gametic and identity disequilibrium within populations are negligible, the likelihood of the sample is given as:

$$P(\mathbf{N}_{A11},\mathbf{N}_{A12},\mathbf{N}_{A22}|\boldsymbol{\gamma}_{A11},\boldsymbol{\gamma}_{A12},\boldsymbol{\gamma}_{A22}) \propto \prod_{i=1}^{I}\prod_{k=1}^{K}\boldsymbol{\gamma}_{A11,ik}^{\mathbf{N}_{A11,ik}}\boldsymbol{\gamma}_{A12,ik}^{\mathbf{N}_{A12,ik}}\boldsymbol{\gamma}_{A22,ik}^{\mathbf{N}_{A22,ik}}, \tag{10}$$

where

$$\begin{aligned}
\boldsymbol{\gamma}_{A11,ik} &= p_{ik}^2 + f p_{ik}(1-p_{ik}) \quad , \\
\boldsymbol{\gamma}_{A12,ik} &= 2p_{ik}(1-p_{ik})(1-f) \quad , \\
\boldsymbol{\gamma}_{A22,ik} &= (1-p_{ik})^2 + f p_{ik}(1-p_{ik}) \\
&= 1 - \boldsymbol{\gamma}_{A11,ik} - \boldsymbol{\gamma}_{A22,ik}
\end{aligned} \tag{11}$$

and $p_{ik} = wx_{ik} + (1-w)y_i$ is the allele frequency at locus i in population k and f is the inbreeding coefficient. We assume a common f across all loci (compare Holsinger *et al.*, 2002). For codominant markers, f can be precisely estimated.

Again the prior distributions for x_{ik} and y_i are based on (2), and $P(\cdot)$ is used to denote the prior distribution for the rest of the parameters and hyperparameters. The full conditional posterior distribution for $\mathbf{x}$, $\mathbf{y}$, $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_K)$, w, f, θ^x, θ^y and π could written as

$$\begin{aligned}
P(\mathbf{x},\mathbf{y},\boldsymbol{\pi},w,f,\theta^x,\theta^y,\pi|\mathbf{N}_{A11},\mathbf{N}_{A12},\mathbf{N}_{A22}) &\propto P(\mathbf{N}_{A11},\mathbf{N}_{A12},\mathbf{N}_{A22}|\boldsymbol{\gamma}_{A11},\boldsymbol{\gamma}_{A12},\boldsymbol{\gamma}_{A22}) \\
&\times \left\{\prod_{i=1}^{I}\left\{\prod_{k=1}^{K} P(x_{ik}|\pi_k,\theta^x)\right\} P(y_i|\theta^y,\pi)\right\} \\
&\times \left\{\prod_{k=1}^{K} P(\pi_k)\right\} P(w)P(f)P(\theta^x)P(\theta^y)P(\pi).
\end{aligned} \tag{12}$$

If we assume that A_1 is dominant to A_2 at each locus, we have dominant markers and the number and frequency of dominant phenotype are the sum of the number and frequency of phenotypes of A_1A_1 and A_1A_2 respectively. It is then straightforward to build a Bayesian model for dominant markers using the mixture beta model.

Analytical expression for the posterior distributions of the parameters and hyperparameters derived from (9) and (12) are not available in closed form. The posterior inference is achieved through Markov chain Monte Carlo(MCMC) simulation and we use the Metropolis-Hasting algorithm (Gilks *et al.*, 1996) in MCMC implementation.

Test for goodness of fit

Goodness of fit of the mixture model to genetic data are evaluated from two aspects. One is to evaluate whether the mixture beta model provides a good approximation for the allele frequency itself and the other is whether the mixture model appropriately incorporates the correlation among allele frequencies. Note that for the beta model (4), while it neglects the correlation among populations, there is no need to check whether it provides a good approximation for the allele frequency since the allele frequency p_{ij} itself is modeled directly. In the mixture model, the allele frequency is modeled as the weighted sum of two beta variables, thus the accuracy of the estimated allele frequency, is compromised at the expense of incorporating correlation. Even so, we are going to show that this compromise occurs to a very minor degree and the mixture model provides an adequate approximation for the allele frequency.

To check whether the mixture beta model provides a good approximation for the allele frequency in our Bayesian model, we apply the chi-squared statistic to test goodness of fit. For (9), conditioning on x_{ik}, y_i and w, the statistic is defined in the following way:

$$\chi^2 = \sum_{i=1}^{I}\sum_{k=1}^{K}\left(\frac{(\mathbf{N}_{A1,ik} - \hat{\mathbf{N}}_{A1,ik})^2}{\hat{\mathbf{N}}_{A1,ik}} + \frac{(\mathbf{N}_{A2,ik} - \hat{\mathbf{N}}_{A2,ik})^2}{\hat{\mathbf{N}}_{A2,ik}}\right). \tag{13}$$

Here $\mathbf{N}_{A1,ik}$ and $\mathbf{N}_{A2,ik}$ are the observed numbers of allele A_1 and A_2. $\hat{\mathbf{N}}_{A1,ik} = (\mathbf{N}_{A1,ik} + \mathbf{N}_{A2,ik})(\hat{w}\hat{x_{ik}} + (1-\hat{w})\hat{y_i})$ and $\hat{\mathbf{N}}_{A2,ik} = (\mathbf{N}_{A1,ik} + \mathbf{N}_{A2,ik})(1 - \hat{w}\hat{x_{ik}} - (1-\hat{w})\hat{y_i})$, where $\hat{w}$, $\hat{x_{ik}}$ and $\hat{y_i}$ are values sampled from the posterior distribution, and treated as the expected values of A_1 and A_2 from the model. Strictly, the Bayesian estimates of allele frequencies are not unbiased. However, with a large amount of data and using Uniform(0,1) as the prior specification, the likelihood dominates the posterior inference and we expect the bias is negligible. So if the mixture model is a good approximation of the allele frequency, the test statistics approximately has a chi-square distribution with degrees of freedom $I \times K - 1$ and χ^2 in (13) divided by the degrees of freedom should be close to 1.

Under the finite island model, for loci with two alleles, the stationary variance (σ^2) and correlation (ρ) for allele frequencies are given by (Fu *et al.*, 2003):

$$\begin{aligned} \sigma^2 &= \frac{u - u^2}{2N - (2N-1)(1-\mu_{21}-\mu_{21})^2\,(1 - r(m,K) + r(m,K)\rho)} \\ \rho &= \frac{\frac{r(m,K)}{K-1}(\mu_{11}-\mu_{21})^2}{1-(1-\mu_{21}-\mu_{21})^2\left(1-\frac{r(m,K)}{K-1}\right)} \end{aligned} \tag{14}$$

where $u = \mu_{21}/(\mu_{12}+\mu_{21})$ is the stationary mean of allelle frequency, μ_{21} is the rate of mutation from A_2 to A_1 and μ_{12}, from A_1 to A_2. In addition, $r(m,K) = 2m - m^2K/(K-1)$, m is the migration rate and K is the number of populations; finally N is the number of diploid individuals in each population. The performance of mixture model is further evaluated by comparing the moments estimated from the mixture model with true values based on (14). We expect little bias in posterior estimates of variance and correlation, too.

Fitting codominant markers to (12), the chi-squared statistic is defined similarly but slightly

different from (13):

$$\chi^2 = \sum_{i=1}^{I}\sum_{k=1}^{K}\left(\frac{(\mathbf{N}_{A11,ik} - \hat{\mathbf{N}}_{A11,ik})^2}{\hat{\mathbf{N}}_{A11,ik}} + \frac{(\mathbf{N}_{A12,ik} - \hat{\mathbf{N}}_{A12,ik})^2}{\hat{\mathbf{N}}_{A12,ik}} + \frac{(\mathbf{N}_{A22,ik} - \hat{\mathbf{N}}_{A22,ik})^2}{\hat{\mathbf{N}}_{A22,ik}}\right). \qquad (15)$$

Write $\mathbf{N}_{ik} = \mathbf{N}_{A11,ik} + \mathbf{N}_{A12,ik} + \mathbf{N}_{A22,ik}$, we calculate $\hat{\mathbf{N}}_{A11,ik} = \mathbf{N}_{ik}\boldsymbol{\gamma}_{\hat{A11}}$, $\hat{\mathbf{N}}_{A12,ik} = \mathbf{N}_{ik}\boldsymbol{\gamma}_{\hat{A12}}$ and $\hat{\mathbf{N}}_{A22,ik} = \mathbf{N}_{ik}\boldsymbol{\gamma}_{\hat{A22}}$ where $\boldsymbol{\gamma}_{\hat{A11}}$, $\boldsymbol{\gamma}_{\hat{A12}}$ and $\boldsymbol{\gamma}_{\hat{A22}}$ are obtained by plugging in the corresponding posterior parameter estimates to (11) and $p_{ik} = wx_{ik} + (1-w)y_i$. Again if the mixture model is a good approximation of the allele frequency, the test statistics approximately has a chi-square distribution with degrees of freedom $2(I \times K - 1)$ for this model and χ^2 in (15) divided by $2(I \times K - 1)$ should be close to 1.

MODELING DATA WITH CLUSTERS

In practice, it is very common to have genetic data from different geographical regions. The human microsatellite data analyzed later in this paper is one such example. Populations within same geographical region may be relatively homogeneous but populations from different geographical regions could be considerably different. More generally, we may just have a set of populations in which some populations are more similar to one another than to others. In such a case, the set of populations could be clustered into different groups naturally by geographical regions or by an appropriate clustering method (e.g., Pritchard *et al.*, 2000). It is reasonable to assume populations within the same cluster are more correlated than populations belonging to different clusters. The within-cluster correlation might also be quite different from one another. To incorporate this characteristic into the model, we assume a different set of parameters for each cluster. Suppose

there are J clusters in the sample, then for each cluster j, we specify

$$p_{ik_j} = w_j x_{ik_j} + (1 - w_j) y_{i_j}, \qquad j = 1, \cdots, J, \tag{16}$$

where

$$\begin{aligned} x_{ik_j} &\sim \text{Beta}(\frac{1-\theta_j^x}{\theta_j^x}\pi_{k_j}, \frac{1-\theta_j^x}{\theta_j^x}(1-\pi_{k_j})), \\ y_{i_j} &\sim \text{Beta}(\frac{1-\theta_j^y}{\theta_j^y}\pi_j, \frac{1-\theta_j^y}{\theta_j^y}(1-\pi_j)). \end{aligned} \tag{17}$$

Again x_{ik_j}'s are assumed to be independent and so are y_{i_j}'s. They are also assumed to be independent from each other.

The above formulation completely neglects the possible correlation among clusters. In fact, by assuming either a common w or a common θ_x or a common θ_y (or equivalently y_i) across clusters, altogether we are interested in comparing the following six models:

(I) common w, common θ^x and common θ^y;

(II) different w, common θ^x and common θ^y;

(III) common w, different θ^x and common θ^y;

(IV) different w, different θ^x and common θ^y;

(V) common w, different θ^x and different θ^y;

(VI) different w, different θ^x and different θ^y.

Note that there are two other possible models, one is common w and common θ^x and different θ^y and the other is different w and common θ^x and different θ^y. We will not include these two models in the comparison. In our view, it is not sensible to assume a common parameter across clusters for the individual component of the mixture model but a different parameter for the common component .

Model (VI) is the same model specified by (16) and (17). Model (V) assumes that the distributions of the two beta variates are different from cluster to cluster but the weight used to mix the two distribution is the same across clusters. Model (I) treats the whole set of the populations as equally correlated and there is no cluster effect. Models (II), (III) and (IV) attempt to build correlation among populations both within clusters and between clusters by sharing a common y. For example, given locus i, it follows from model (IV) that

$$\begin{aligned} \mathrm{Var}(x_{ik_j}) &= w_j^2\theta_j^x\pi_{k_j}(1-\pi_{k_j}) + (1-w_j)^2\theta^y\pi(1-\pi), \\ \mathrm{Cov}(x_{ik_j}, x_{ik'_j}) &= (1-w_j)^2\theta^y\pi(1-\pi), \\ \mathrm{Cov}(x_{ik_j}, x_{ik'_{j'}}) &= (1-w_j)(1-w_{j\prime})\theta^y\pi(1-\pi), \end{aligned} \tag{18}$$

which gives the covariance among populations from within clusters and among clusters. For human populations, by treating populations from each geographical region as one cluster, populations from different clusters are obviously correlated because of shared history and gene flow due to migration. Our goal is to model both correlations simultaneously. However, one drawback of the above approach is that both correlations are built through the same set of variables y_i, thus the magnitude of correlation among any two clusters is greatly affected by the correlation within each

cluster. This may be true in some special cases but may not be desirable in general. We recognize the limitation of this approach to investigate correlations from both within and among clusters and concentrate on the inference of correlation within clusters. Further, we will perform model selection by using quadratic loss L measure(Ibrahim and Laud, 1994; Gelfand and Ghosh, 1998; Ibrahim *et al.*, 2004). Given any sample, if the above approach is not the appropriate way to specify correlation among clusters, we expect the model performance is not as good as some alternative models and would not be selected.

Model Comparison

Models (I) - (VI) all assume some correlation among populations. Besides comparing models (I) - (VI), we are also interested in comparing models (I) - (VI) with four beta models that do not incorporate correlation among populations. One of the models is (4) and the other three are modified versions of (4) with cluster effects incorporated to different extents. Equation (4) assumes no cluster effect, i.e., a common θ and π_i for the whole set of populations. For notation, we refer (4) as model (i). The other three models assume a common θ and cluster-specific π_i (model (ii)), or a cluster-specific θ and common π_i (model (iii)), or both θ and π_i cluster-specific (model (iv)). Model (iv) is given by

$$p_{ik_j} \sim \text{Beta}(\frac{1-\theta_j}{\theta_j}\pi_{i_j}, \frac{1-\theta_j}{\theta_j}(1-\pi_{i_j})) \tag{19}$$

where $\text{E}(p_{ik_j}) = \pi_{i_j}$, $\text{Var}(p_{ik_j}) = \pi_{i_j}(1-\pi_{i_j})\theta_j$ and π_{i_j} represents the mean of allele frequency across all populations in cluster j at locus i. A common θ implies that the genetic differentiation among populations, if not correlated, is the same across clusters, though the mean and variance of allele frequency might be different across clusters.

F_{ST} analysis will be compared between models with correlation incorporated or not. They will also be compared based on quadratic loss L measure (Ibrahim and Laud, 1994; Gelfand and Ghosh, 1998; Ibrahim *et al.*, 2004). Quadratic loss L measure is a decision-theoretic approach to model choice based on expected losses on replicate data sets. For codominant data, the number of phenotype follows a multinomial distribution and we use the multivariate version developed by Ibrahim *et al.* (2004) to estimate L Measure. Let $\mathbf{y}$ denote observed data and $\mathbf{z} = (z_1, \cdots, z_n)$ denote future response, and each z_i is a vector. Let

$$L(\mathbf{y}, \nu) = \sum_{i=1}^{n} \mathrm{Cov}(z_i|\mathbf{y}) + \nu \sum_{i=1}^{n} (\mathrm{E}(z_i|\mathbf{y}) - y_i)(\mathrm{E}(z_i|\mathbf{y}) - y_i)' \tag{20}$$

then the quadratic L measure criterion is given as the trace of $L(\mathbf{y}, \nu)$. The expectation for variance-covariance matrix and mean of z is taken with respect to the posterior predictive distribution, which is defined as:

$$p(z|\mathbf{y}) = \int_{\mathbf{B}} p(z|\boldsymbol{\beta}) p(\boldsymbol{\beta}|\mathbf{y}) d\boldsymbol{\beta} \tag{21}$$

where $p(\boldsymbol{\beta}|\mathbf{y})$ is the posterior distribution of $\boldsymbol{\theta}$ given observed data. The quantity ν is a number between 0 and 1, and usually decided by the investigator. Equation (20) indicates that L measure could be considered as a weighted sum of expected variance and estimated variance of future response. When $\nu = 0$, L measure depends only on the expected variance of the future observations. For reasons discussed in the section of Test for goodness of fit (allele frquency is modelled as a sum of two beta variates), the estimated allele frequency from the mixture model is not as accurate as that from the beta model, thus the mixture model would produce a larger estimated variance of future response than the beta model. With this in mind, we are more interested in small values of

ν and testing whether the mixture model could provide better (or smaller) expected variance for future observation. For L measure in general, a smaller value indicates a better model.

IMPLEMENTATION

Finite island model simulation

Finite island model is one of the most important theoretical models in population genetics. The defining assumption of finite island model is equal migration rate among populations. Substantial results have been derived from this widely studied model (e.g., Crow and Aoki, 1984; Cockerham and Weir, 1987; Fu *et al.*, 2003). Under this model, the correlation among populations is induced by migration (gene exchange), but the number of populations involved in gene exchange and mutation rate affects the magnitude of correlation.

To evaluate the performance of the mixture beta model, we consider loci with two alleles and simulate data from the finite island model with different combinations of mutation rate (μ) and migration rate (m), population size ($2N$) and number of populations (K). In particular, there are four different sets of simulations. The first set of simulation is equivalent to exhaustive sampling, and each simulated data set consist of allele frequencies from N individuals in each of all K populations. In this simulation, we use symmetric mutation rate between the two alleles and the mutation rate μ is set to be one of (1.0×10^{-3}, 1.0×10^{-4}, 1.0×10^{-5}, 1.0×10^{-6}). The migration rate m is one of (0.001, 0.01, 0.1) and population size $2N$, one of (1000, 2000, 20000). We choose the number of populations K to be one of (2, 10, 25, 100). The above combinations of population process parameters would result in the mean allele frequency being 0.5 in each combination, variance

varying from 1.0×10^{-4} to 0.20 and correlation varying from 0.02 to 0.999. In the second set of simulation, we only consider the cases with a relatively large population size of $2N = 20000$. For each of the K populations, allele frequencies of $n = 100$ individuals were sampled from the whole population with a subset of mutation and migration rates chosen from the set above. In the third set of simulation, the mutation rate is set to be asymmetric and the number of populations is one of (100, 250, 500, 1000). Then the combinations of mutation rate, migration rate, population size are chosen such that the mean allele frequency is 0.4, the correlation is about 0.28 and the variance is within the range of (0.01,0.04). In each combination, $2n = 40$ is sampled from each population and $K' = 52$ populations are sampled. These values are chosen to approximate those from the full human data set analyzed below. For the fourth set of simulation, the number of populations K is one of (25, 50, 100, 250, 500, 1000) and population size, one of (500, 1000, 2000). Again, we use symmetric mutation rates between two alleles and choose mutation and migration rates such that the correlation is about 0.67 and variance varys in the range of (0.002,0.06). In each combination, $2n = 40$ is sampled from each population and $K' = 2$ or $K' = 10$ populations are sampled. This simulation evaluates the performance of mixture beta model when data are available only from a small number of populations. The variance and correlation of this simulation are comparable to those from the human data of different geographical regions. At last, in the first three sets of simulations, for each combination of process parameters, we sample allele frequency from 50 independent generations from the stationary distribution, which is equivalent to 50 independent loci evolving under the finite island model. In the fourth set of simulation, we sample allele frequency from both 50 and 377 loci (there are 377 loci in the human dataset). All the simulated data are fitted to model (9).

For the first set of simulation, Table 1 ($K = 25$) and Table 2 ($K = 100$) show the variance and correlation among allele frequencies estimated from mixture model, compared with true values from finite island model and values calculated from the simulated data as summary statistics. With data from exhaustive sampling, the mixture model provides accurate estimates of variance and correlation for all simulated cases when the migration rate is 0.1 and for some cases when the migration rate is smaller ($m = 0.01$ and $\mu \leq 1.0 \times 10^{-4}$; $m = 0.001$ and $\mu \leq 1.0 \times 10^{-3}$ for the given $2N$ and K). The variance and correlation are very close to the true values, whether the true correlation is 0.04 or 0.99. In some other cases when $m = 0.01$ ($K = 25, \mu = 1.0 \times 10^{-5}, 2N = 1000$ or $2N = 2000$; $K = 25, \mu = 1.0 \times 10^{-6}, 2N = 2000$) or 0.001 ($K = 25, \mu = 1.0 \times 10^{-6}, 2N = 20000$), the mixture model provides good estimates of variance and correlation, which are close to the true values. Only when both mutation and migration rates are small (e.g., $m = 0.001$ and $\mu = 1.0 \times 10^{-5}$ or 1.0×10^{-6}) along with relatively small population size ($2N = 1000$ or 2000), does the mixture model fail to estimate the correlation accurately but estimates of variance are still fairly good.

In population genetics, the distribution of allele frequency are commonly described in terms of $2Nm$ and $2N\mu$. For the mixture model, as shown from the current simulations, the performance is satisfactory other than when both $2Nm$ and $2N\mu$ are very small ($2Nm = 1$ or 2 and $2N\mu$ is less than 1). However, obtaining the explicit conditions when the performance of the mixture model is satisfactory/unsatisfactory through simulation has not proven feasible because it is determined by the complex interaction among the four process parameters. On the other hand, the values of process parameters are usually not known in reality and it is almost impossible to judge the fit of the mixture model from the values of $2Nm$ and $2N\mu$. Instead, by looking at the data, whenever the estimates of mixture model is unsatisfactory, the values of allele frequencies have a predominantly

amount of 0's and 1's. Although the posterior estimate of allele frequencies will not be 0's and 1's due to input form the uniform prior, the mixture beta model or (any beta model) is deemed not be the best description of the data.

When $K = 2$ or $K = 10$, the results are similar (not shown here). Both variance and correlation could be accurately estimate unless the values of allele frequencies have a predominantly amount of 0's and 1's. The parameter estimates are associated with a wider credible interval.

The last columns of Table 1 and Table 2 give the values of χ^2/df for the mixture model. In some earlier simulations, the values of χ^2 are not calculated and represented by " / ". The values of χ^2/df are very close to 1 in all the calculated cases, indicating the mixture model could provide a good approximation of allele frequency whether the correlation is incorporated or not.

Table 3 shows the comparison of moments from the second set of simulations where allele frequencies from $2n = 200$ are sampled from $2N = 20000$. Values of χ^2/df are close 1 and the estimated variance and correlation are very close to the true values from finite island model in all simulated cases. Table 4 shows the results from the third set of simulations with $K' = 52$ and $2n = 40$, similar to the results from the second set of simulations. Allele frquency is well approximated and both variance and correlation well estimated. For the fourth set of simulations with $K' = 2$ or $K' = 10$ from $K = (25, 50, 100, 250, 500, 1000)$, Table 5 selectively shows estimates with$K' = 2$. For both $I = 50$ and $I = 377$, correlation could be well estimated and with $I = 377$, correlation is estimated with better precision. Estimates when $K' = 10$ is omitted here. For the last three sets of simulations, there are few 1's and 0's in the data. Also for sampled data, the mixture data usually provides better estimates for variance and correlation than estimates as summary statistics directly from simualated data.

So in general, these results suggests that the mixture model works well for a broad range of data, unless the data consist of a predominantly amount of allele frequencies being 0's and 1's. Both allele frequency and the variance-covariance structure are accurately estimated. Model parameter estimates of θ^x, θ^y, w and π are not reported in Fu (2003). The mean allele frequency could always be estimated whether the mutation rates between the two alleles are symmetric or not. Estimates of θ^x and θ^y indicate that the mixture model could be two unimodal betas, a U-shaped beta and unimodal beta as well as two U-shaped betas.

An example of human Data

Cann *et al.*. (2002) reported a HGDP-CEPH Human Genome Diversity Cell Line Panel. The data set includes genotypes at 377 autosomal microsatellite loci in 1056 individuals from 52 worldwide populations. Rosenberg *et al.* (2002) clustered the 52 populations into 5 groups, which correspond to 5 geographical areas by using the Bayesian clustering method proposed by Pritchard *et al.* (2000). In the data set, there are multiple allele types for each loci. To make this data set to be able to fit to our mixture model, we designate the most frequent allele type as A_1 and group all the other allele types into a pseudo-allele type A_2. Since the numbers of each of the three genotypes A_1A_1, A_1A_2 and A_2A_2 are known from the data set, this is an example of codominant data. We also use the 5 clusters presented by Rosenberg *et al.* (2002) and fit the data to models (i) - (iv) and models (I) - (VI). The five groups are EuroAsia (21 populations), African (6 populations), EastAsia (18 populations), American (5 populations) and Oceania (2 populations).

Results of L Measure for each model are shown in Table 6. Among models (I) - (VI), model (VI) produces the smallest L Measure consistently for the different values of ν thus is considered

Table 1: Comparison of Moment Estimates for data simulated from finite island model with different combinations of m, μ and $2N$ when $K = 25$.

			Moment estimates						
			Mixture Model		Finite Island Model		Simulated Data		
			Var	Corr	Var	Corr	Var	Corr	χ^2/df
(μ)	(m)	$2N$	Mean	Mean			Mean	Mean	Mean
($2N\mu$)	($2Nm$)		(95% CI)	(95% CI)					(95% CI)
1.0×10^{-3}	0.001	2000	0.0189	0.0268	0.0194	0.0203	0.0189	0.0303	1.001
(2)	(2)		(0.0176,0.0203)	(0.00614,0.0542)					(0.925,1.088)
		20000	0.00215	0.0471	0.00208	0.0203	0.0212	0.0329	1.002
(20)	(20)		(0.00198,0.00233)	(0.0188,0.0883)					(0.930,1.077)
1.0×10^{-3}	0.01	2000	0.00581	0.171	0.00598	0.171	0.00579	0.148	1.002
(2)	(20)		(0.00525,0.00644)	(0.112,0.248)					(0.930,1.086)
		20000	0.000591	0.203	0.000611	0.171	0.00590	0.186	1.002
(20)	(200)		(0.00529,0.000672)	(0.133,0.293)					(0.926,1.082)
1.0×10^{-3}	0.1	2000	0.00142	0.609	0.00184	0.663	0.00132	0.541	1.018
(2)	(200)		(0.00105,0.00192)	(0.440,0.759)					(0.941,1.095)
		20000	0.000204	0.732	0.000185	0.663	0.000206	0.700	1.019
(20)	(2000)		(0.000152,0.000285)	(0.654,0.808)					(0.945,1.109)
1.0×10^{-4}	0.001	2000	0.0463	0.0428	0.0490	0.172	0.0497	0.157	1.003
(0.2)	(2)		(0.0436,0.0490)	(0.0296,0.0562)					(0.929,1.078)
1.0×10^{-4}	0.01	1000	0.0339	0.693	0.0318	0.675	0.0345	0.709	0.995
(0.1)	(10)		(0.0291,0.0392)	(0.640,0.738)					(0.918,1.079)
		20000	0.00194	0.664	0.00181	0.675	0.00184	0.650	1.003
(2)	(200)		(0.00149,0.00261)	(0.568,0.765)					(0.921,1.087)
1.0×10^{-4}	0.1	1000	0.0206	0.955	0.0238	0.952	0.0202	0.945	1.003
(0.1)	(100)		(0.0150,0.0275)	(0.939,0.968)					(0.933,1.068)
		2000	0.0115	0.954	0.0125	0.952	0.0109	0.943	1.021
(0.2)	(200)		(0.00809,0.0160)	(0.936,0.969)					(0.946,1.101)
		20000	0.00152	0.966	0.00130	0.952	0.00142	0.957	1.021
(2)	(2000)		(0.00104,0.00220)	(0.951,0.977)					(0.938,1.099)
1.0×10^{-5}	0.001	1000	0.149	0.00142	0.148	0.676	0.148	0.713	0.944
(0.01)	(1)		(0.139,0.155)	(0.000679,0.00236)					(0.852,1.075)
		2000	0.0891	0.0316	0.105	0.676	0.104	0.677	1.011
(0.02)	(2)		(0.0821,0.0939)	(0.0257,0.0383)					(0.937,1.100)
1.0×10^{-5}	0.01	1000	0.0996	0.845	0.128	0.954	0.140	0.964	/
(0.01)	(10)		(0.0902,0.108)	(0.824,0.863)					/
		2000	0.0715	0.907	0.0859	0.954	0.0735	0.944	/
(0.02)	(20)		(0.0607,0.0815)	(0.889,0.922)					/
		20000	0.0149	0.962	0.0124	0.954	0.0141	0.961	1.044
(0.2)	(200)		(0.0106,0.0206)	(0.947,0.973)					(0.965,1.119)
1.0×10^{-5}	0.1	1000	0.124	0.997	0.125	0.995	0.143	0.996	/
(0.01)	(100)		(0.106,0.141)	(0.996,0.998)					/
		2000	0.0732	0.996	0.0836	0.995	0.0651	0.994	/
(0.02)	(200)		(0.0578,0.0913)	(0.995,0.997)					/
1.0×10^{-6}	0.001	2000	0.168	0.0132	0.210	0.954	0.193	0.946	/
(0.002)	(2)		(0.151,0.177)	(0.00937,0.0169)					/
		20000	0.0487	0.780	0.0859	0.954	0.0608	0.929	1.018
(0.02)	(20)		(0.0427,0.0544)	(0.741,0.817)					(0.941,1.097)
1.0×10^{-6}	0.01	2000	0.136	0.894	0.209	0.995	0.224	0.997	/
(0.002)	(20)		(0.129,0.141)	(0.884,0.903)					/

Table 2: Comparison of Moment Estimates for data simulated from finite island model with different combinations of m, μ and $2N$ when $K = 100$.

			Moment estimates						
			Mixture Model		Finite Island Model		Simulated Data		
			Var	Corr	Var	Corr	Var	Corr	χ^2/df
(μ)	(m)	$2N$	Mean	Mean			Mean	Mean	Mean
($2N\mu$)	($2Nm$)		(95% CI)	(95% CI)					(95% CI)
1.0×10^{-3}	0.01	2000	0.00510	0.0375	0.00535	0.0477	0.0323	0.00518	1.000
(2)	(20)		(0.00491,0.00529)	(0.0214,0.0596)					(0.962,1.041)
1.0×10^{-4}	0.001	1000	0.0754	0.000217	0.0758	0.0480	0.0768	0.0440	/
				($4.359 \times 10^{(-9)}$,					
(0.1)	(1)		(0.0719,0.0777)	$5.010 \times 10^{(-4)}$)					/
		2000	0.0433	0.00603	0.0446	0.0480	0.0439	0.0330	1.001
(0.2)	(2)		(0.0417,0.0448)	(0.00392,0.00853)					(0.962,1.044)
1.0×10^{-4}	0.01	1000	0.0175	0.366	0.0171	0.334	0.00174	0.358	/
(0.1)	(10)		(0.0159,0.0196)	(0.302,0.434)					/
		2000	0.00824	0.305	0.00884	0.334	0.00822	0.297	1.001
(0.2)	(20)		(0.00742,0.00926)	(0.237,0.385)					(0.961,1.043)
1.0×10^{-4}	0.1	1000	0.00536	0.808	0.00733	0.827	0.00533	0.765	/
(0.1)	(100)		(0.00399,0.00745)	(0.744,0.865)					/
		2000	0.00274	0.814	0.00372	0.827	0.00269	0.770	1.018
(0.2)	(200)		(0.00191,0.00385)	(0.749,0.872)					(0.977,1.058)
1.0×10^{-6}	0.001	1000	0.184	0.00163	0.1874	0.835	0.185	0.878	/
(0.001)	(1)		(0.164,0.194)	(0.00122,0.00211)					
1.0×10^{-6}	0.01	1000	0.103	0.818	0.180	0.981	0.147	0.970	/
(0.001)	(10)		(0.0951,0.110)	(0.800,0.834)					/
1.0×10^{-6}	0.1	1000	0.171	0.998	0.179	0.998	0.175	0.998	/
(0.001)	(100)		(0.156,0.185)	(0.9982,0.9986)					/

Table 3: Comparison of Moment Estimates for data simulated from finite island model with different combinations of m, μ and $K = 25$ with $2n = 200$ sampled from $2N = 20000$.

			Moment estimates						
			Mixture Model		Finite Island Model		Simulated Data		
			Var	Corr	Var	Corr	Var	Corr	χ^2/df
K	(μ)	(m)	Mean	Mean			Mean	Mean	Mean
	$(2N\mu)$	$(2Nm)$	(95% CI)	(95% CI)					(95% CI)
25	1.0×10^{-4}	0.001	0.00553	0.158	0.00594	0.172	0.00670	0.119	1.015
	(2)	(20)	(0.00497,0.00627)	(0.0895,0.239)					(0.936,1.095)
25	1.0×10^{-4}	0.01	0.00165	0.711	0.00181	0.675	0.00288	0.379	1.024
	(2)	(200)	(0.00120,0.00123)	(0.592,0.824)					(0.945,1.105)
25	1.0×10^{-5}	0.01	0.0134	0.968	0.0124	0.954	0.0140	0.875	1.084
	(0.2)	(200)	(0.00886,0.0198)	(0.947,0.984)					(1.005,1.159)
25	1.0×10^{-5}	0.1	0.0127	0.999	0.0120	0.995	0.0128	0.907	1.015
	(0.2)	(2000)	(0.00857,0.0187)	(0.9994,0.9999)					(0.999,1.035)
25	1.0×10^{-6}	0.01	0.0732	0.998	0.0836	0.995	0.0741	0.982	1.093
	(0.02)	(200)	(0.0816,0.859)	(0.997,0.999)					(1.032,1.157)
25	1.0×10^{-6}	0.1	0.0815	0.999	0.0834	0.999	0.0885	0.991	0.961
	(0.02)	(2000)	(0.0639,0.101)	(0.9999,0.9999)					(0.940,0.985)
100	1.0×10^{-3}	0.001	0.00205	0.00233	0.00207	0.00501	0.00341	-0.00163	1.047
	(20)	(20)	(0.00188,0.00219)	(0.000899,0.00814)					(1.004,1.090)
100	1.0×10^{-3}	0.01	0.000440	0.0249	0.000545	0.0477	0.00180	0.00378	1.052
	(20)	(200)	(0.000313,0.000530)	(0.00474,0.0560)					(1.011,1.094)
100	1.0×10^{-4}	0.001	0.00527	0.0590	0.00531	0.0480	0.00650	0.0420	1.017
	(2)	(20)	(0.00500,0.00557)	(0.0360,0.0909)					(0.9772,1.056)
100	1.0×10^{-4}	0.01	0.000798	0.395	0.000913	0.334	0.00213	0.136	1.051
	(2)	(200)	(0.000623,0.000984)	(0.287,0.521)					(1.013,1.091)

Table 4: Comparison of moments for data simulated from finite island model with different combinations of m, μ ,K and $2N$: $K' = 52$, $2n = 40$.

				Moment estimates						
				Mixture Model		Finite island model		Simulated Data		
				Var	Corr	Var	Corr	Var	Corr	χ^2/df
v_{12} $2N\mu_{12}$	v_{21} $2N\mu_{21}$	m $2Nm$	$2N$	Mean (95% CI)	Mean (95% CI)			Mean	Mean	Mean (95% CI)
$K = 100$										
0.00015 (0.075)	0.0001 (0.05)	0.01 (5)	500	0.0278 (0.0245,0.0311)	0.297 (0.229,0.371)	0.0288	0.287	0.0338	0.231	1.030 (0.976,1.085)
0.00037 (0.075)	0.00025 (0.05)	0.025 (5)	200	0.0299 (0.0264,0.0336)	0.330 (0.254,0.407)	0.0292	0.287	0.0358	0.255	1.040 (0.987,1.095)
(0.185)	(0.125)	(12.5)	500	0.0116 (0.0102,0.0133)	0.276 (0.197,0.367)	0.0126	0.278	0.0177	0.158	1.052 (0.995,1.108)
0.00075 (0.075)	0.0005 (0.05)	0.05 (5)	100	0.0262 (0.0230,0.0299)	0.299 (0.224,0.381)	0.0294	0.282	0.0322	0.208	1.041 (0.988,1.098)
(0.15)	(0.1)	(10)	200	0.0129 (0.0114,0.0149)	0.282 (0.206,0.372)	0.0156	0.282	0.0188	0.146	1.043 (0.985,1.100)
0.00115 (0.0575)	0.00075 (0.0375)	0.075 (3.75)	50	0.0333 (0.0300,0.0367)	0.274 (0.210,0.341)	0.0380	0.277	0.0408	0.229	1.040 (0.988,1.097)
(0.115)	(0.075)	(7.5)	100	0.0192 (0.0165,0.0228)	0.381 (0.290,0.486)	0.0205	0.277	0.0248	0.225	1.037 (0.982,1.094)
0.0015 (0.075)	0.001 (0.05)	0.01 (5)	50	0.0243 (0.0214,0.0276)	0.284 (0.208,0.370)	0.0302	0.277	0.0303	0.201	1.028 (0.934 1.083)
(0.15)	(0.1)	(10)	100	0.123 (0.0106,0.0146)	0.328 (0.241,0.427)	0.0160	0.277	0.0185	0.171	1.059 (1.004,1.114)
$K = 250$										
0.00015 (0.0225)	0.0001 (0.015)	0.025 (3.75)	150	0.0352 (0.0313,0.0393)	0.306 (0.241,0.376)	0.0379	0.284	0.0418	0.214	1.043 (0.987,1.097)
(0.0375)	(0.025)	(6.25)	250	0.0227 (0.0199,0.0258)	0.324 (0.247,0.410)	0.0241	0.284	0.0284	0.213	1.032 (0.973,1.088)
0.0003 (0.03)	0.0002 (0.02)	0.05 (5)	100	0.0253 (0.0224,0.0287)	0.278 (0.207,0.362)	0.0299	0.281	0.0317	0.187	1.041 (0.989,1.095)
(0.06)	(0.04)	(10)	200	0.0136 (0.0115,0.0159)	0.308 (0.222,0.411)	0.0159	0.281	0.0198	0.187	1.050 (0.995,1.105)
$K = 500$										
0.000075 (0.01125)	0.00005 (0.0075)	0.025 (3.75)	150	0.0374 (0.0334,0.0412)	0.298 (0.234,0.366)	0.0381	0.284	0.0473	0.299	1.071 (1.015,1.129)
(0.01875)	(0.0125)	(6.25)	250	0.0248 (0.0218,0.0279)	0.315 (0.240,0.393)	0.0243	0.284	0.0303	0.214	1.040 (0.982,1.095)
0.00015 (0.015)	0.0001 (0.01)	0.05 (5)	100	0.0250 (0.0222,0.0277)	0.244 (0.183,0.310)	0.0301	0.281	0.0308	0.181	1.033 (0.982,1.079)
(0.03)	(0.02)	(10)	200	0.0136 (0.0117,0.0158)	0.326 (0.242,0.428)	0.0160	0.281	0.0193	0.163	1.048 (0.991,1.107)
0.00225 (0.1125)	0.0015 (0.075)	0.075 (3.75)	50	0.0337 (0.0297,0.0376)	0.279 (0.214,0.351)	0.0391	0.278	0.0415	0.214	1.045 (0.992,1.095)
(0.225)	(0.15)	(7.5)	100	0.0202 (0.0177,0.0230)	0.360 (0.282, 0.443)	0.0211	0.278	0.0260	0.244	1.033 (0.977,1.089)
$K = 1000$										
0.000115 (0.0575)	0.000075 (0.00375)	0.075 (3.75)	50	0.0309 (0.0280,0.0336)	0.192 (0.145,0.250)	0.0389	0.275	0.0379	0.184	1.030 (0.973,1.084)
(0.0115)	(0.0075)	(7.5)	100	0.0182 (0.0159,0.0212)	0.321 (0.242,0.415)	0.0210	0.276	0.0241	0.206	1.035 (0.980,1.095)
0.00015 (0.0075)	0.0001 (0.005)	0.1 (5)	50	0.0278 (0.0245,0.0313)	0.329 (0.253,0.404)	0.0309	0.276	0.0338	0.217	1.046 (0.990,1.101)
(0.015)	(0.01)	(10)	100	0.0123 (0.0128,0.0141)	0.263 (0.183,0.355)	0.0164	0.276	0.0184	0.132	1.053 (0.999,1.110)

Table 5: Comparison of Moment Estimates for data simulated from finite island model with different combinations of m, μ, K and $2N$ with $K' = 2$ and $2n = 40 : I = 50$ and $I = 377$.

				Moment estimates						
				Mixture Model		Finite Island Model		Simulated Data		
				Var	Corr	Var	Corr	Var	Corr	χ^2/df
K	μ	m	$2N$	Mean	Mean			Mean	Mean	Mean
	$(2N\mu)$	$(2Nm)$		(95% CI)	(95% CI)					(95% CI)
$I = 50$										
25	1.0×10^{-4}	0.01	500	0.0617	0.519	0.0564	0.675	0.0564	0.644	1.033
	(0.05)	(5)		(0.0499,0.0739)	(0.313,0.689)					(0.776,1.355)
			1000	0.0302	0.413	0.0318	0.675	0.0363	0.378	1.031
	(0.1)	(10)		(0.0219,0.0403)	(0.0975,0.676)					(0.792,1.318)
			2000	0.0162	0.617	0.0170	0.675	0.0218	0.457	1.034
	(0.2)	(20)		(0.0109,0.0234)	(0.272,0.866)					(0.752,1.339)
100	2.5×10^{-5}	0.01	500	0.0453	0.579	0.0575	0.668	0.0491	0.556	1.040
	(0.013)	(5)		(0.0338,0.0589)	(0.375,0.747)					(0.770,1.377)
			1000	0.0359	0.523	0.0325	0.668	0.0411	0.522	1.002
	(0.025)	(10)		(0.0263,0.0470)	(0.270,0.733)					(0.749,1.296)
			2000	0.0165	0.516	0.0173	0.668	0.0218	0.407	1.036
	(0.05)	(20)		(0.0109,0.0248)	(0.159,0.805)					(0.789,1.349)
500	2.5×10^{-5}	0.05	500	0.0112	0.727	0.0143	0.661	0.0166	0.446	1.055
	(0.013)	(25)		(0.00672,0.0173)	(0.317,0.996)					(0.773,1.360)
			1000	0.0103	0.432	0.00734	0.661	0.0155	0.287	1.020
	(0.025)	(50)		(0.00616,0.0157)	(0.0321,0.847)					(0.752,1.324)
			2000	0.00769	0.383	0.00372	0.661	0.0130	0.159	1.036
	(0.05)	(100)		(0.00405,0.0128)	(0.00994,0.813)					(0.789,1.349)
$I = 377$										
25	1.0×10^{-4}	0.01	500	0.0552	0.665	0.0564	0.675	0.0637	0.654	1.042
	(0.05)	(5)		(0.0507,0.0601)	(0.594,0.724)					(0.940,1.152)
			1000	0.0321	0.638	0.0318	0.675	0.0378	0.548	1.011
	(0.1)	(10)		(0.0288,0.0358)	(0.562,0.706)					(0.919,1.110)
			2000	0.0159	0.669	0.0170	0.675	0.0578	0.652	1.002
	(0.2)	(20)		(0.0139,0.0184)	(0.564,0.757)					(0.908,1.100)
100	2.5×10^{-5}	0.01	500	0.0511	0.669	0.0575	0.668	0.0594	0.640	1.024
	(0.013)	(5)		(0.0470,0.0554)	(0.607,0.723)					(0.924,1.133)
			1000	0.0326	0.655	0.0325	0.668	0.0366	0.569	1.013
	(0.025)	(10)		(0.0291,0.0365)	(0.584,0.721)					(0.916,1.110)
			2000	0.0177	0.647	0.0173	0.668	0.0233	0.481	1.011
	(0.05)	(20)		(0.0154,0.0202)	(0.550,0.733)					(0.918,1.108)
500	2.5×10^{-5}	0.05	500	0.0137	0.701	0.0143	0.661	0.0195	0.485	1.006
	(0.013)	(25)		(0.0115,0.0158)	(0.588,0.800)					(0.914,1.108)
			1000	0.00767	0.776	0.00734	0.661	0.0137	0.424	1.009
	(0.025)	(50)		(0.00626,0.00917)	(0.634,0.919)					(0.905,1.125)
			2000	0.00329	0.730	0.00372	0.661	0.00948	0.241	1.016
	(0.05)	(100)		(0.00230,0.00438)	(0.427,0.994)					(0.915,1.122)

as our best model to incorporate correlation among populations for the human data. Model (VI) recognizes substantial differences among clusters but ignores the correlation among clusters. Since none of models (II) - (IV) gives a performance as good as model (VI), we take this as evidence that these models could not appropriately incorporate correlation among clusters. So, we focus our inference on each cluster. Values of χ^2/df (Table 7) from each model are all very close to 1 and models (I) - (VI) provide an adequate approximation to allele frequency p_{ik}.

Among models (i) - (iv), model (iv) (and model (VI)) outperforms the other three models based on L Measure. Model VI produces smaller L measure than model (iv) when $\nu < 0.3$, which indicates that model (VI) provides a better estimate of expected variance for future response than the model (iv). Also, the estimated χ^2/df and the corresponding 95% credible interval from model (VI) is 1.040(1.027, 1.054). This value is not significantly different from the corresponding estimates from model (iv), which is 1.028(1.014, 1.046). In other words, not only does model (VI) produces a smaller expected variance for future responses and incorporate correlation into account, it costs little accuracy in estimating allele frequency for the mixture mdoel to do so. Actually, when comparing model (i) and model (I) using data simulated from finite island model without clusters, we get similar results on L Measure and χ^2/df.

Table (7) shows the parameter estimates from models (I) - (VI). For Model (VI), posterior estimate of inbreeding coefficient f is 0.015. The small value is expected since in human populations, inbreeding occurs very infrequently and the closest degree of inbreeding usually encountered in most societies is first-cousin mating. Also with codominant data, f is precisely estimated as shown by the narrow credible interval. Estimates of θ^x, θ^y and w are quite different, indicating it is appropriate to fit data from each cluster to separate mixture models. Figure (1) shows the kernel density

Table 6: Quadratic L Measure estimated from different models for model comparison.

	Models of non-correlated populations				Models of correlated populations					
	(i)	(ii)	(iii)	(iv)	(I)	(II)	(III)	(IV)	(V)	(VI)
$\nu = 0.0$	294449	282362	276809	267969	292720	280584	283624	282368	267417	266095
$\nu = 0.1$	307996	296909	291988	283959	307494	295988	299170	297307	284232	282968
$\nu = 0.2$	321543	311456	307167	299949	322269	311392	314715	312246	301046	299842
$\nu = 0.3$	335090	326004	322347	315938	337043	326796	330261	327185	317861	316715
$\nu = 0.4$	348637	340551	337526	331928	351817	342200	345806	342124	334676	333589
$\nu = 0.5$	362184	355099	352705	347918	366592	357604	361352	357064	351490	350462

estimation of parameters from each cluster based on model (VI).

Estimated variance of allele frequency and correlation among populations within each cluster from model (VI) are given in Table 8. Since there are only very few 0's and 1's in the dataset, based on results from simulated finite island data, we believe that these estimates of correlation among populations are reliable and accurate. In general, the populations are highly correlated and it would not be sensible to ignore such correlation in the analysis. Particularly, the cluster of EastAsia has the highest correlation of 0.885 followed by African, 0.767; EuroAsia, 0.747 and Oceania, 0.614. The populations in American has the least correlation of 0.420. These values reflect the common belief that the history of humans is shorter in American than that in the other continents. The most recent claim by Seielstad *et al.* (2003) is that humans first entered America within the last 15000 years . Hence the populations in American have not had the time long enough to get mixed and related, contrary to the other continents.

Values of F_{ST} (Bayesian estimate of $\mathrm{E}(Num/Denom)$ shown in Table 8 are obtained by plugging posterior estimates of allele frequency p_{ik} from model (VI) for each loci then taking average over populations within each cluster. For all but one cluster, estimates of F_{ST} are small (i.e., ≤ 0.05) and indicate little genetic differentiation among populations within each cluster; and the cluster of

Table 7: Model parameter estimates of the mixture models.

	EuroAsia	African	EastAsia	American	Oceania
	$K_1 = 21$	$K_2 = 6$	$K_3 = 18$	$K_4 = 5$	$K_5 = 2$
Parameters	Mean (95% CI)	Mean (95% CI)	Mean (95% CI)	Mean (95% CI)	Mean (95% CI)
Model (I)					
θ^x	Common across clusters:			0.121(0.117,0.125)	
θ^y	Common across clusters:			0.415(0.373,0.461)	
w	Common across clusters:			0.730(0.718,0.741)	
π	Common across clusters:			0.387(0.354,0.426)	
f	Common across clusters:			0.0166(0.0135,0.0198)	
χ^2/df				1.050(1.039,1.063)	
Model (II)					
θ^x	Common across clusters:			0.237(0.227,0.246)	
θ^y	Common across clusters:			0.0958(0.0840,0.110)	
w	0.276 (0.264,0.288)	0.656 (0.629,0.682)	0.428 (0.408,0.447)	0.999 (0.995,0.999)	0.804 (0.763,0.851)
π	Common across clusters:			0.264(0.242,0.284)	
f	Common across clusters:			0.0160(0.0127,0.0190)	
χ^2/df				1.042(1.029,1.053)	
Model (III)					
θ^x	0.0499 (0.0469,0.0531)	0.155 (0.144,0.168)	0.0967 (0.0911,0.103)	0.393 (0.371,0.412)	0.245 (0.221,0.268)
θ^y	Common across clusters:			0.387(0.347,0.427)	
w	Common across clusters:			0.722(0.710,0.732)	
π	Common across clusters:			0.352(0.321,0.382)	
f	Common across clusters:			0.0163(0.0130,0.0196)	
χ^2/df				1.042(1.031,1.056)	
Model (IV)					
θ^x	0.0757 (0.0695,0.0818)	0.129 (0.119,0.139)	0.177 (0.148,0.206)	0.272 (0.260,0.284)	0.253 (0.223,0.290)
θ^y	Common across clusters:			0.196(0.167,0.231)	
w	0.629 (0.606,0.651)	0.866 (0.841,0.889)	0.385 (0.349,0.420)	0.994 (0.979,0.999)	0.726 (0.677,0.775)
π	Common across clusters:			0.333(0.308,0.363)	
f	Common across clusters:			0.0155(0.0128,0.0186)	
χ^2/df				1.029(1.019,1.042)	
Model (V)					
θ^x	0.0899 (0.0764,0.0100)	0.178 (0.154,0.203)	0.0815 (0.0704,0.0942)	0.553 (0.517,0.588)	0.390 (0.325,0.470)
θ^y	0.121 (0.105,0.139)	0.208 (0.182,0.239)	0.245 (0.214,0.275)	0.433 (0.393,0.478)	0.298 ((0.250,0.353)
w	Common across clusters:			0.426(0.411,0.441)	
π	0.386 (0.355,0.417)	0.302 (0.274,0.330)	0.406 (0.374,0.440)	0.376 (0.341,0.410)	0.374 (0.298,0.437)
f	Common across clusters:			0.0158(0.0129,0.0192)	
χ^2/df				1.038(1.025,1.052)	
Model (VI)					
θ^x	0.186 (0.132,0.248)	0.712 (0.549,0.784)	0.773 (0.746,0.795)	0.377 (0.347,0.411)	0.493 (0.322,0.765)
θ^y	0.0816 (0.0668,0.0988)	0.124 (0.106,0.144)	0.106 (0.0932,0.120)	0.622 (0.554,0.697)	0.271 (0.204,0.364)
w	0.271 (0.228,0.325)	0.161 (0.145,0.186)	0.114 (0.105,0.121)	0.609 (0.578,0.643)	0.370 (0.264,0.479)
π	0.322 (0.299,0.348)	0.217 (0.200,0.235)	0.351 (0.333,0.367)	0.435 (0.385,0.484)	0.353 (0.274,0.439)
f	Common across clusters:			0.0159(0.0126,0.0193)	
χ^2/df				1.040(1.027,1.054)	

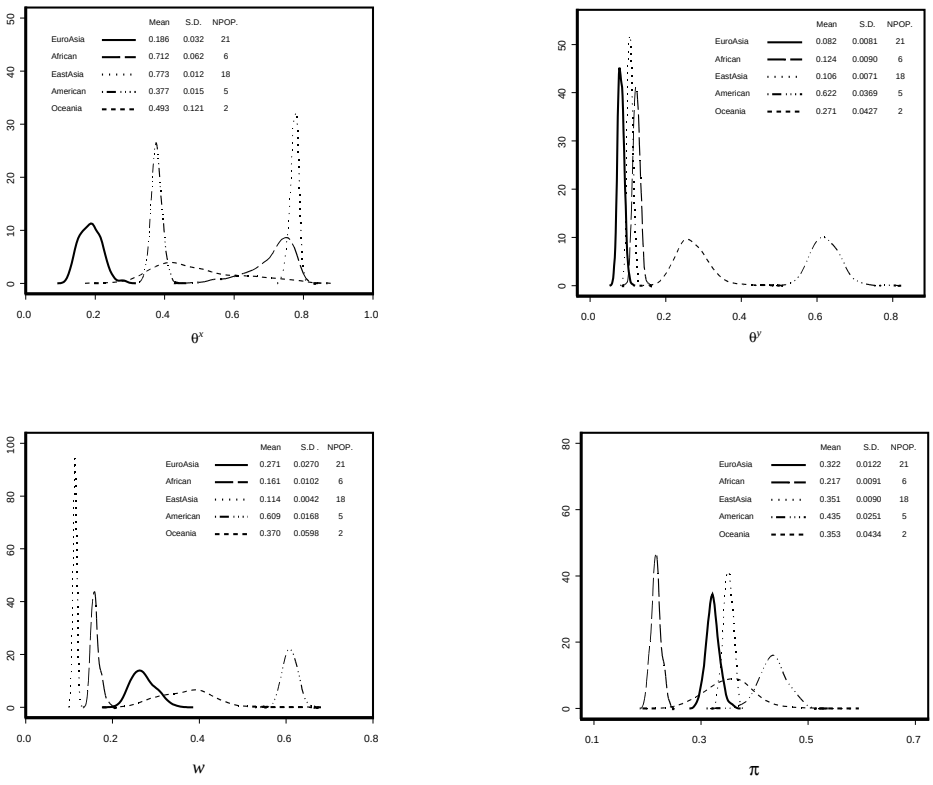

Figure 1: Comparison of kernel density estimation of parameter estimates from model (VI) for each cluster.

American shows evidence for moderate genetic differentiation among populations with an estimate of F_{ST} larger than 0.1. It is interesting to observe that when the numbers of population are similar between two clusters (e.g., African ($K = 6$) vs. American ($K = 5$); EuroAsia ($K = 21$) vs. EastAsia ($K = 18$)), a larger estimate of F_{ST} is always associated with a smaller correlation. The cluster of EastAsia yields the smallest estimate of F_{ST} and the largest estimate of correlation, and for the cluster of American, it produces the larger estimate of F_{ST} and smallest estimate of correlation. This intuitively makes sense since larger estimate of F_{ST} (greater genetic differentiation) means

more variation thus less similarity and correlation among allele frequencies.

To compare F_{ST} (Bayesian estimate of E($Num/Denom$) with θ from model (iv), estimates of θ from each cluster are uniformly greater than estimates of F_{ST}, and when the number of populations are small (i.e., clusters of America and Oceania), significantly smaller (Table 8). For cluster of America, estimate of θ is about 30% greater than estimate of F_{ST}; and for cluster of Oceania, estimate of θ is almost twice of estimate of F_{ST}. These results are consistent with Fu *et al.* (2003) that when populations are small, genetic differentiation is signicantly overestimated if correlation is not accounted for. There are less populations in America and Oceania to exchange genes. (*Dr. Kent, is this a fair statement? Or is there a reference?).* The small number of populations sampled from America and Oceania may also play a role in the difference. Rosenberg *et al.* (2002) used Weir and Cockerham's method (Weir and Cockerham, 1984) to estimate allele frequency differentiation for each cluster. These results are denoted as θ^c also shown in Table 8 for comparison purpose. It is not surpurising that their estimates are similar to estimates of θ's based on model (iv) but greater than our estimates as both estimates neglect correlation among populations.

In this analysis, the multiallelic microsatellite data are reduced to biallelic data by designating the most frequest allele type as A_1. To test the sensitivity of our estimates to this assumption, we assign the second most frequest allele type as A_1 and fit the converted data to Model (VI) and also report the results in Table 8. The estimates of F_{ST} and correlation are very comparable to the estimates when most frequest allele type is treated as A_1. It remains the same that the cluster of EastAsia has the highest correlation(0.851 vs. 0.885) followed by African and EuroAsia, and American has the least correlation. Estimates of F_{ST} are also similar, so are estimtes of f (0.0147(0.0117, 0.0181)) and χ^2/df (1.033(1.015,1.055)). So our estimates are quite robust.

Table 8: Comparison of estimated variance, correlation and F_{ST} for each geographical region.

	EuroAsia	African	EastAsia	American	Oceania
	$K_1 = 21$	$K_2 = 6$	$K_3 = 18$	$K_4 = 5$	$K_5 = 2$
Parameters	Mean (95% CI)	Mean (95% CI)	Mean (95% CI)	Mean (95% CI)	Mean (95% CI)
Model (VI): the most frequent allele type as A_1					
Variance	0.0127	0.0191	0.0214	0.0555	0.0401
	(0.0114,0.0143)	(0.0168,0.0215)	(0.0190,0.0240)	(0.0525,0.0589)	(0.0357,0.0463)
Correlation	0.747	0.767	0.885	0.420	0.614
	(0.645,0.839)	(0.726,0.828)	(0.870,0.900)	(0.374,0.465)	(0.416,0.773)
F_{ST}	0.0144	0.0236	0.0116	0.101	0.0334
	(0.0135,0.0155)	(0.0208,0.0268)	(0.0100,0.0131)	(0.0972,0.106)	(0.0288,0.0378)
Model (iv): the most frequent allele type as A_1					
θ	0.0158	0.0298	0.0141	0.129	0.0618
	(0.0146,0.0170)	(0.0261,0.0337)	(0.0124,0.0158)	(0.120,0.138)	(0.0516,0.0730)
Rosenberg *et al.* (2002)					
θ^c	0.0150	0.0310	0.0130	0.116	0.0640
	(0.0140,0.0160)	(0.0290,0.0330)	(0.0110,0.0140)	(0.110,0.123)	(0.0570,0.0720)
Model (VI): the second most frequent allele type as A_1					
Variance	0.00717	0.0112	0.0119	0.0337	0.0222
	(0.00643,0.00803)	(0.00933,0.0130)	(0.0104,0.0137)	(0.0282,0.0398)	(0.0191,0.0258)
Correlation	0.655	0.668	0.851	0.259	0.590
	(0.613,0.699)	(0.591,0.772)	(0.801,0.894)	(0.191,0.332)	(0.488,0.688)
F_{ST}	0.0137	0.0221	0.0109	0.0948	0.0273
	(0.0128,0.0146)	(0.0196,0.0248)	(0.00926,0.0126)	(0.0907,0.0987)	(0.0236,0.0317)

Furthermore, the fact that estimates of θ from model (iv) are very similar to estimates from Rosenberg *et al.* (2002) but the former are based on converted biallelic data and the latter based on multiallelic data, also indicates this conversion has little effect.

DISCUSSION

In population genetics, probability models have been used to describe allele frequency and make inference about population structure. The beta model developed by Balding and Nichols (1995) and its multiallelic version may have been the most commonly used one. Nicholson *et al.* (2002) proposed a truncated normal model for single nucleotide polymorphism allele frequencies. The beta distribution is usually justified as the equilibrium distribution under several genetic models of interest and the rational for the truncated normal distribution is in terms of modeling the transient states of allele frequency. However, the assumption of equilibrium or transient states is effectively impossible to check. In most cases, it is more practical to select the model which gives the best fit of data. The marginal distribution of allele frequency may well be approximated by a beta distribution empirically unless multimodal. Our mixture model is based on beta distribution. When there is a lot of allele frequencies of 0 or 1, the truncated normal distribution could be a better choice since it has mass at 0 or 1.

Correlation among populations has not been treated adequately in probability models for allele frequency, even it affects the estimate of population structure. The mixture model we present here explicitly incorporates the correlation among population. Based on evaluation using simulated data from the finite island model, the mixture model provides a good approximation of allele frequency and an accurate estimate of correlation in general unless there is a large proportion of

allele frequencies being 0s and 1s. This model could be easily applied to allele frequency data and dominant/codominant phenotype data.

Although (function of) our model parameters do not bear the interpretation as a measure of population structure, Bayesian estimate of traditional measure based on relative reduction of heterozygosity could be calculated by using the posterior estimate of allele frequency to quantify genetic differentiation, in which effect of correlation among populations is also included. Our results indicate that the amount of genetic differentiation is overestimated when effect of correlation among populations is not accounted for. It would take more studies to investigate to what extent this correlation affects the estimation of population structure. In the future, we will pursue how to modeling correlated allele frequency with a large proportion of 0s and 1s. One possibility may use a mixture of truncated normal distribution (Nicholson *et al.* (2002)) since it has mass at 0 and 1. Our current model is appropriate for multilocus genotype data with two allele types and we will seek to extend our model to multiallelic data. We also extend our approach to model data with clusters and attempt to model correlation both within and among clusters. Effects of correlation on estimates of effective size, migration rate and other quantities of interest will also be the topic of future research.

REFERENCES

Balding,D.J. and Nichols,R.A. (1995) A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica*, **96**, 3–12.

Balding,D.J. and Nichols,R.A. (1997) Significant genetic correlations among Caucasians at forensic DNA loci. *Heredity*, **78**, 583–589.

Beerli and Felsenstein. (1999) Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalesent approach. *Genetics*, **152**, 763–773.

Cann,H.M., de Toma,C., Cazes,L., Legrand,M.F., Morel,V., Piouffre,L., Bodmer,J., Bodmer,W.F., Bonne-Tamir,B., Cambon-Thomsen,A., Chen,Z., Chu,J., Carcassi,C., Contu,L., Du,R., Excoffier,L., Ferrara,G.B., Friedlaender,J.S., Groot,H., Gurwitz,D., Jenkins,T., Herrera,R.J., Huang,X., Kidd,J., Kidd,K.K., Langaney,A., Lin,A.A., Mehdi,S.Q., Parham,P., Piazza,A., Pistillo,M.P., Qian,Y., Shu,Q., Xu,J., Zhu,S., Weber,J.L., Greely,H.T., Feldman,M.W., Thomas,G., Dausset,J., Cavalli-Sforza,L.L. (2002) A human genome diversity cell line panel. *Science*, **296**, 261–262.

Cockerham,C.C. and Weir,B.S. (1987) Correlations, descent measures: drift with migration and mutation. *Proceedings of the National Academy of Sciences USA*, **84**, 8512–8514.

Crow,J.F. and Aoki,K. (1984) Group selection for a polygenic behavioral trait: estimating the degree of population subdivision. *Proceedings of the National Academy of Sciences USA*, **81**, 6073–6077.

Diaconis,P. and Ylvisaker,D. (1985) Quantifying prior opinion. In Bernardo,J.M., Degroot,M.H., Lindley,D.V. and Smith,A.F.M. (eds.) *Bayesian Statistics 2. Proc. 2nd Valencia Int'l Meeting, 9-83.* North-Holland, Amsterdam, pp. 133–156 .

Ewens,W.J. (1979) *Mathematical Population Genetics.* Springer-Verlag, New York, NY.

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.

Fu,R., Gelfand,A.E. and Holsinger,K.E. (2003) Exact moment calculations for genetic models with

migration, mutation and drift. *Theorectical Population Biology*, **63**, 231-243.

Fu,R. 2003. Probabilistic Structure and Statistical Inference for Nonexplicit Population Models of Allele Frequency. Ph.D. Dissertation, University of Connecticut, U.S.A.

Gelfand,A.E. and Ghosh,S.K. (1998) Model choice: A minimum posterior predictive loss approach. *Biometrika*, **85**, 1-13.

Gilks,W.R., Richardson,S., Spiegelhalter,D.J. (1996) Introducing Markov chain Monte Carlo. In Gilks,W.R., Richardson,S. and Spiegelhalter D.J. (eds), *Markov Chain Monte Carlo in Practice.* Chapman & Hall, London, pp. 1-19.

Holsinger,K.E. (1999) Analysis of genetic diversity in geographically structured populations: a Bayesian perspective. *Hereditas*, **130**, 245–255.

Holsinger,K.E., Lewis,P.O. and Dey,D.K. (2002) Bayesian analysis of population structure with dominant marker data. *Molecular Ecology*, **11**, 1157–1164.

Ibrahim,J.G. and Laud,P.W. (1994) A predictive approach to the analysis of designed experiments. *Journal of American Statistical Associaton*, **89**, 309–319.

Ibrahim,J.G. and Chen,M.-H. (2000) Power prior distributions for regression models. *Statistical Science*, **15**, 46–60.

Ibrahim,J.G., Chen,M.-H. and Sinha,D. (2004) Bayesian methods for joint modeling of longitudinal and survival data with applications to cancer vaccine studies. *To appear in Statistica Sinica.*

Nicholson,G., Smith,A.V, Jónsson,F., Gústaffson,Ó., Steánsson,K. and Donnelly,P. (2002) Assessing population differentiation and isolation from single-nucleotide polymorphism data. *Journal of Royal Statistical Society B*, **64**, 695-716.

Pritchard,J.K., Stephens,M. and Donnelly,P. (2000) Inference of population structure using multi-

locus genotype data. *Genetics*, **155**, 945–959.

Roeder,K., Escobar,M., Kadane,J.B. and Balazs,I. (1998) Measuring heterogeneity in forensic databases using hierarchical Bayes models. *Biometrika*, **85**, 269–287.

Rosenberg,N.A., Pritchard,J.K., Weber,J.L., Cann,H.M., Kidd,K.K., Zhivotovsky,L.A. and Feldman,M.W. (2002) Genetic structure of human populations. *Science*, **298**, 2381–2385.

Seielstad,M., Yuldasheva,N., Singh,N., Underhill,P., Oefner,P., Shen,P. and Wells,R.S. (2003) A Novel Y-Chromosome Variant Puts an Upper Limit on the Timing of First Entry into the Americas. *American Journal of Human Genetics* , **73**, 700–705.

Weir,B.S. and Cockerham,C.C. (1984) Estimating F-statistics for the analysis of population structure. *Evolution*, **38**, 1358–1370.

Wright,S. (1943) Isolation by distance. *Genetics*, **28**, 114–138.

Wright,S. (1951) The genetical structure of populations. *Annals of Eugenics*, **15**, 323–354.

Wright,S. (1969) *Evolution and the Genetics of Populations. Volume 2, The Theory of Gene Frequencies.* University of Chicago Press, Chicago, IL.