Bayesian Modeling for Genetic Association in Case-Control Studies: Accounting for Unknown Population Substructure

Li Zhang, Bhramar Mukherjee*, Malay Ghosh and Rongling Wu
Department of Statistics, University of Florida, Gainesville, FL 32611

SUMMARY

A two-stage parametric Bayesian method is proposed to examine the association between a candidate gene and the occurrence of a disease after accounting for population substructure. This procedure, implemented via a Markov chain Monte Carlo numerical integration technique, first estimates the posterior probability of different unknown population substructures and then integrates this information into a disease-gene association model through the technique of Bayesian model averaging. The model relaxes certain assumptions of previous analyses and provides a unified computational framework to obtain an estimate of the log odds ratio parameter corresponding to the genetic factor after allowing for the allele frequencies to vary across subpopulations. The uncertainty in estimating the population substructure is taken into account while providing credible intervals for parameters in the disease-gene association model. Simulations on unmatched case-control studies that mimic an admixed Argentinean population are performed to demonstrate the statistical properties of our model. The method is also applied to a real dataset coming from a genetic association study on obesity.

KEY WORDS: Bayesian model averaging; Gene-disease association; Linkage equilibrium; Markov chain Monte Carlo; Obesity.

*Correspondence author: Assistant Professor, Department of Statistics, University of Florida P.O. Box 118545, Gainesville FL 32611-8545, USA. Phone: +1-352-392-1941 ext 241 Fax: +1-352-392-5175 Email: mukherjee@stat.ufl.edu

1 Introduction

The evaluation of the association between molecular markers and disease status can be used to study the genetic basis of common human diseases (Risch and Merikangas, 1996; Morton and Collins, 1998; Sullivan *et al.*, 2001). The basic principle for such so-called association studies arises from the dependence of allele frequencies at marker loci upon those of disease variants, that is, the linkage disequilibria between alleles from different genetic loci. A significant association detected between a marker and the disease can be considered as evidence for close physical linkage between the marker and a disease locus, given that the linkage disequilibrium between any two genes always decays exponentially with their genetic distance in a random mating idealized population (Lynch and Walsh, 1998).

In practice, however, there rarely exists an idealized population as a result of the action of various evolutionary forces (Lynch and Walsh, 1998). Evolutionary forces, such as population structure and population admixture, operating on a population can result in spurious associations between a phenotype and markers that are not linked to any causative loci. The presence of spurious association suggests that the detected statistical association does not necessarily imply the physical linkage between the disease phenotype and arbitrary markers that have no physical linkage to causative loci (Lander and Schork, 1994). A classic example of spurious association caused by population substructure is presented in Knowler *et al.* (1988). In this study, based on a sample of Native Americans of the Pima and Papago tribes, a very strong negative association between the Gm haplotype Gm3;5,13,14 and type 2 or non-insulin-dependent diabetes mellitus was detected. One might conclude from this observation that the absence of this haplotype, or the presence of a closely linked gene is a causal risk factor for the disease. However Gm3;5,13,14 is a marker for Caucasian admixture, and it is most likely that the presence of Caucasian alleles and decrease in Indian alleles led to lower susceptibility to type 2 diabetes, rather than the direct action of the haplotype or of a closely linked locus. This study demonstrates the effects of confounding due to population substructure, and the importance of considering genetic admixture while investigating the association between a disease and genetic markers.

In order to overcome the problem of spurious associations, many different genetic strategies have been proposed. Spielman *et al.* (1993) used the transmission disequilibrium test (TDT) to measure the association between a candidate gene and disease status by incorporating the genotypes of parents of affected individuals. This test has been instrumental in genetic association studies of human diseases (Spielman and Evens, 1998), but it is often limited because of difficulties with DNA sampling. For this reason, a simple case-control design that uses affected individuals and unrelated controls has recently received increased attention (Freedman *et al.*, 2004; Marchini *et al.*, 2004). A number of approaches have been developed to avoid the generation of spurious associations in case-control studies of disease-gene association. For a comprehensive recent review of admixture mapping for complex traits see McKeigue (2005).

Pritchard and colleagues used multilocus genotype data to estimate population substructure. They proposed a model-based clustering method to identify the population structure by genotyping samples at additional unlinked markers (Pritchard *et al.*, 2000a). This method was then extended to allow for the linkage between different markers (Falush *et al.*, 2003). A software package, STRUCTURE, has been written to implement their algorithms that consider both linked and unlinked markers. Pritchard *et al.* (2000b) proposed a two-stage procedure in which first the population structure is inferred by employing the method of Pritchard *et al.*(2000a), and then the tests of association within subpopulations are conducted conditional on the imputed substructure. However, this method does not develop a model for the probability of disease incidence and cannot be generalized easily to provide estimates of the odds ratio corresponding to the genetic risk factor. Hoggart *et al.* (2003, 2004) developed a combination of Bayesian and classical approaches for association studies based on the admixture between populations with different ancestries. Apart from STRUCTURE, two other softwares which employ Bayesian ideas for statistical modeling of genetic data from admixed population are ADMIXMAP (Hoggart *et al.*, 2003, 2004) and ANCESTRYMAP (Patterson *et al.*, 2004).

Different from the above treatments, Satten *et al.* (2001) provided a novel latent-class analysis to study the association between the disease and the candidate genes based on a series of additional markers that are in linkage equilibrium with each other and with the candidate genes within subpop-

ulations. Based on the Akaike information criterion (AIC), their method can estimate the number of subpopulations. But by either assuming the disease to be rare, or collapsing multiple genotypes into various binary genotypes, their method has not fully capitalized on the information about the multiple-genotype inheritance of the candidate gene.

In this article, we provide an alternative parametric Bayesian model for inferring on disease-gene association after accounting for population substructure. As in Satten *et al.* (2001), we use the latent-class approach to estimate the association parameters, while we account for the population substructure in a way similar to that of Pritchard *et al.* (2000a). However, unlike Satten *et al.* (2001), our analysis does not require the rare disease assumption or analyzing multicategory genotypes by several analyses using various possible binary genotypes of the candidate gene. Our model can also handle multi-allelic genotypes of the candidate genes, extending on earlier approaches for the genotypic analysis of only biallelic loci. The computational strategy followed in Satten *et al.* (2001) involved use of the E-M algorithm to estimate the parameters in the model, combined with a parametric bootstrap strategy to obtain standard error estimates. The Markov chain Monte Carlo strategy designed in this paper simplifies the computational complexity, with posterior standard deviation estimates and credible intervals being obtained from the random observations generated from the full conditional distributions of the parameters.

We should emphasize that in our Bayesian analysis, inference on the disease-gene association is not carried out on the basis of the particular imputed structure as done in Pritchard *et al.* (2000a). Instead, through use of model averaging (see for example, Madigan and Raftery (1994)), the association parameters are estimated by incorporating the uncertainty in estimating the substructure. In particular, instead of assuming the number of subpopulations I to be fixed, we put a prior on I and obtain the posterior distribution of I. For each possible value of I with positive posterior probability, we then estimate the association parameters in the disease-gene risk model. Finally we take the weighted average of these estimates, the weights being proportional to the posterior probabilities of the different values of I. The explicit model averaging formulas are given in Section 3.2. Our analysis thus combines the substructure estimation ideas of Pritchard *et al.* (2000a) using Bayesian clustering, and

the latent class disease risk models of Satten *et al.* (2001) posed in a purely frequentist framework, through a more general unified Bayesian approach. The paper presents a novel two-stage model with a clustering algorithm for inferring on cryptic population structure, followed by a logistic model for disease incidence, tied together through the technique of Bayesian model averaging.

The outline of the paper is as follows. Section 2 states both the statistical model and the genetic model, and briefly introduces the methods in Pritchard *et al.* (2000a) to estimate the number of subpopulations. Section 3 derives the underlying likelihood. We also introduce in this section the appropriate priors for the model parameters and obtain their estimates based on the posteriors. The posteriors are analytically intractable. So the Bayesian procedure is implemented by the Markov chain Monte Carlo numerical integration technique. In Section 4, we state our simulation strategy and provide results on simulated case-control studies under both a rare disease and a common disease assumption. Our simulation studies are conducted in the same setting as in Satten *et al.* (2001) and mimic an admixed Argentinean population as described in Sala *et al.* (1998, 1999). Under the rare disease assumption, we compare our results with those obtained in Satten *et al.* (2001). In Section 5, we apply our methods to real data collected in a genetic association study with obesity as the disease outcome and the β-adrenergic receptor β2AR as the candidate gene under investigation. Some concluding remarks are made in Section 6.

2 Model and Notation

2.1 Statistical Model

Let the binary variable D denote disease and let G be a (possibly vector-valued) genetic risk factor. We assume that the overall population of size N is comprised of I subpopulations, each having different frequencies of G and D. By the unmeasured covariate Z, we indicate the subpopulation to which an individual belongs. Thus, $D_j(= 1 \text{ or } 0)$ corresponds to the presence or absence of a disease for the jth individual with a genetic risk factor G_j, $j = 1, \cdots, N$.

We assume G_j to be a univariate discrete random variable, taking $M + 1$ values $g_0(= 0)$, $g_1, \cdots,$

g_M. We assume that the prospective conditional logistic distribution for the disease status is

$$\Pr(D_j = 1|G_j = g_m, Z = i) = H\{\beta_{0i} + \beta_{1m}\}, \quad m = 0, \cdots, M, \tag{1}$$

where $H(u) = \{1 + \exp(-u)\}^{-1}$. Here β_{0i} is a term representing the subpopulation effect on the probability of disease for individuals belonging to a particular subpopulation i, and β_{1m} is the coefficient corresponding to the genetic exposure variable in the above logistic regression model. For parameter identifiability, we set β_{10} = 0. The method can immediately be extended to a vector valued genetic risk factor $\boldsymbol{G}_j$ for individual j.

2.2 Genetic Model

Since different subpopulations may have different frequencies of other marker genes, we use a latent-class approach to infer about the population substructure by using information on those additional marker loci. Consider x_l^c as the allele at marker l on chromosome c =1, 2 (labeling of the two chromosomes in a given pair as 1 or 2 is arbitrary) and let X = (x_1^1, x_1^2, $\cdots$, x_L^1, x_L^2), where L is the number of marker loci under consideration.

First, we assume that the genes at the additional marker loci are unrelated to disease, that is

$$\Pr(D_j = 1|G_j, X_j, Z = i) = \Pr(D_j = 1|G_j, Z = i). \tag{2}$$

In the analysis that follows, we assume that Hardy-Weinberg equilibrium holds for each subpopulation. Human populations rarely show much divergence from the Hardy-Weinberg equilibrium once population substructure has been accounted for (Report of Committee on DNA Forensic Science 1996, pp. 104 and references cited therein).

Further, by choosing additional marker loci on different chromosomes from the chromosome where G is found, we first assume that the additional mutually independent marker genes are in

linkage equilibrium with the candidate gene G, so that

$$\Pr(G_j, X_j|Z = i) = \Pr(G_j|Z = i) \times \Pr(X_j|Z = i). \tag{3}$$

By Hardy-Weinberg equilibrium,

$$\Pr(X_j|Z = i) = \prod_{l=1}^{L} \prod_{c=1}^{2} p_{lix_l^c}, \tag{4}$$

where $p_{lix_l^c}$ is the proportion of persons in subpopulation i having allele x_l^c at marker loci l, L being the number of marker loci.

Suppose the candidate gene G has w alleles, e.g., $a_1, \cdots, a_w$, and the frequency of the allele a_u $(u = 1, \cdots, w)$ in the ith subpopulation is

$$\rho_{iu} = \Pr[G_l^c = a_u|Z = i].$$

Then by Hardy-Weinberg equilibrium the probabilities of the genotypes of G $(a_u a_v)$ $(u, v = 1, \cdots, w)$ are given by:

$$\Pr[G = a_u a_v|Z = i] = \begin{cases} \rho_{iu}^2, & u = v; \\ 2\rho_{iu}\rho_{iv}, & u \neq v. \end{cases} \tag{5}$$

2.3 Inference on I for the model with admixture

We consider the situation where we have multilocus genotype data from individuals sampled from a population with possibly unknown structure. Pritchard *et al.* (2000a) used the genotypes of a sample of individuals to identify the presence of population structure which is difficult to detect using visible characters, but may be significant in genetic terms. As Pritchard *et al.* (2000a) pointed out, the problem of inferring on the number of unknown populations, I, present in a data set is a very difficult task. In a Bayesian paradigm, with a suitably chosen prior distribution on I, one can base inference

for I on the posterior distribution:

$$P(I|\mathbf{X}) \propto P(\mathbf{X}|I)P(I), \tag{6}$$

where $\mathbf{X}$ denotes the vector of genotypes of the sampled individuals including the candidate gene G. Let $\mathbf{Z}$ denote the unknown population of origin of the individuals, $\mathbf{P}$ denote the unknown allele frequency vector in all populations, and $\mathbf{Q}$ denote the vector of admixture proportions for each individual. The harmonic mean estimator is one of the simplest ways of estimating $P(\mathbf{X}|I)$,

$$\frac{1}{P(\mathbf{X}|I)} = \int \frac{P(\mathbf{Z},\mathbf{P},\mathbf{Q}|\mathbf{X},I)}{P(\mathbf{X}|\mathbf{Z},\mathbf{P},\mathbf{Q},I)} d\mathbf{Z}d\mathbf{P}d\mathbf{Q} \approx \frac{1}{K}\sum_{k=1}^{K} \frac{1}{P(\mathbf{X}|\mathbf{Z}^{(k)},\mathbf{P}^{(k)},\mathbf{Q}^{(k)},I)}. \tag{7}$$

However this estimator is notoriously unstable, often having infinite variance, and thus poses severe computational challenges. Pritchard *et al.* (2000a) described an alternative approach which is a more *ad hoc* but effective approach based on the Bayesian deviance function

$$DV(\mathbf{Z},\mathbf{P},\mathbf{Q}) = -2\log P(\mathbf{X}|\mathbf{Z},\mathbf{P},\mathbf{Q}). \tag{8}$$

Let $k = 1, 2, \cdots$ denote the k-th iteration in the Markov chain. One estimates the conditional mean and variance of the deviance function DV given $\mathbf{X}$ as follows:

$$\begin{aligned} E(DV(\mathbf{Z},\mathbf{P},\mathbf{Q})|\mathbf{X}) &\approx \frac{1}{K}\sum_{k=1}^{K} -2\log P(\mathbf{X}|\mathbf{Z}^{(k)},\mathbf{P}^{(k)},\mathbf{Q}^{(k)}) = \hat{\mu}, \\ Var(DV(\mathbf{Z},\mathbf{P},\mathbf{Q})|\mathbf{X}) &\approx \frac{1}{K}\sum_{k=1}^{K} (-2\log P(\mathbf{X}|\mathbf{Z}^{(k)},\mathbf{P}^{(k)},\mathbf{Q}^{(k)}) - \hat{\mu})^2 = \hat{\sigma}^2. \end{aligned}$$

By assuming that the conditional distribution of the deviance function DV given $\mathbf{X}$ is normal, it follows from (7) that

$$-2\log P(\mathbf{X}|I) \approx \hat{\mu} + \hat{\sigma}^2/4. \tag{9}$$

An analytical explanation of this approximation is provided in Appendix A. An alternative interpreta-

tion of this method is that model selection is based on penalizing the mean of the Bayesian deviance by a quarter of its variance. Pritchard *et al.* (2000a) pointed out that replacing the assumption of normality with the assumption of the Bayesian deviance function being distributed as a Gamma random variable may be asymptotically more justifiable, but makes little or no difference in terms of estimation accuracy in practical applications.

One may use (9) to estimate $P(\mathbf{X}|I)$ for each I and then substitute the estimate into (6) to obtain approximate estimates of $P(I|\mathbf{X})$ (see Pritchard *et al.* 2000a, for a detailed algorithm). One would then impute the estimated substructure while conducting tests for disease-gene association. We will essentially follow the same technique for estimating $P(I|\mathbf{X})$ and embed the derived information into a disease risk model as described in the following section.

3 Likelihood and Priors

In this section, we derive the likelihood function, state our prior distributions and derive the posteriors. The key aspect of the modeling is in how we develop algorithms for estimating the model parameters and at the same time account for the population structure in our framework.

3.1 Likelihood

Because different subpopulations may have different frequencies of other marker genes, we make inference based on the marginal joint distribution of D, G and X, summing over all possible values of Z, the latent variate. Let $\Pr(Z = i) = q_i$, which is the proportion of persons in subpopulation i. Note that for subject j, G_j takes one of the values g_m, $m = 0, 1, \cdots, M$. By (3) and (4), for given I, the full likelihood L_I is factorized as follows:

$$\begin{aligned} L_I &= \prod_{j=1}^{N} \Pr(D_j, G_j, X_j) = \prod_{j=1}^{N} \sum_{i=1}^{I} \big[\Pr(Z = i) \times \Pr(G_j, X_j | Z = i) \times \Pr(D_j | G_j, Z = i) \big] \\ &= \prod_{j=1}^{N} \sum_{i=1}^{I} \Big[q_i \times \Big\{ \prod_{l=1}^{L} \prod_{c=1}^{2} p_{lix_l^c} \Big\} \times \Pr(G_j = g_m | Z = i) \times \frac{\exp\{D_j \times (\beta_{0i} + \beta_{1m})\}}{1 + \exp\{\beta_{0i} + \beta_{1m}\}} \Big]. \end{aligned} \quad (10)$$

where $\Pr(G_j|Z=i)$ is a function of ρ_{iu} $(u=1,\cdots,w)$ as described in (5), and L is the number of marker loci which are in linkage equilibrium with G.

We use a marginal likelihood rather than a conditional likelihood approach. The likelihood involves parameters of interest β_{1m} $(m=1,\cdots,M)$, and the nuisance parameters β_{0i}, ρ_{iu}, q_i and $p_{lix}(i=1,\cdots,I;\ \forall\ l$ and $\forall\ x)$, which grow in direct proportion to the number of subpopulations. This gives rise to the well known Neyman-Scott phenomenon where MLEs turn out to be inconsistent if I grows with sample size. Typically we deal with I between 1 through 7, and handling nuisance parameters is not a difficult issue in such scenarios. However, the marginal model does contain a large number of parameters, and we carry out Bayesian inference by introducing appropriate prior distributions for these parameters.

3.2 Priors and Posteriors

The main problem is to estimate the regression parameters $\beta_{1m}\ , m=1,\cdots,M$; we consider the following mutually independent normal priors:

$$\begin{aligned}\beta_{0i} &\sim \text{Normal}\ (\mu_{\beta_{0i}}, \sigma^2_{\beta_{0i}}), \quad i=1,\cdots,I;\\ \beta_{1m} &\sim \text{Normal}\ (\mu_{\beta_{1m}}, \sigma^2_{\beta_{1m}}), \quad m=1,\cdots,M.\end{aligned}$$

When inferring the number of subpopulations I, we consider a discrete uniform prior on the domain of I. The priors for $\mathbf{P}$ and $\mathbf{Q}$ correspondingly are the following:

$$\begin{aligned}(q_1,\cdots,q_I) &\sim \text{Dirichlet}(\alpha,\cdots,\alpha);\\ \rho_{iu} &\sim \text{Beta}(a_i,b_i);\\ (p_{li1},p_{li2},\cdots,p_{liX_l}) &\sim \text{Dirichlet}(\lambda_{p_{li1}},\lambda_{p_{li2}},\cdots,\lambda_{p_{liX_l}}).\end{aligned}$$

With the above model and prior specifications, one can obtain the full conditional distributions for the parameters β_{0i}, β_{1m}, ρ_{iu}, q_i and p_{lix}. The full conditionals are given in the Appendix B. None of the

conditionals has a standard distributional form.

For each given value of I, the parameters of interest can be estimated by generating random observations from the full conditionals using a Markov chain Monte Carlo numerical integration scheme and then taking averages of the generated observations. Corresponding to each value of I, we also have associated posterior probabilities $P(I|\mathbf{X})$ as discussed in section 2.3. Therefore, by setting $\theta = (\beta_{11}, \cdots, \beta_{1M})$, using a model-averaging technique, any generic parameter θ is estimated by the posterior mean

$$E(\theta|\mathbf{X}) = \sum_i E(\theta|\mathbf{X}, I = i) \Pr(I = i|\mathbf{X}) \tag{11}$$

with posterior variance

$$\begin{aligned} V(\theta|\mathbf{X}) &= \sum_i V(\theta|\mathbf{X}, I = i) \Pr(I = i|\mathbf{X}) \\ &+ \sum_i [E(\theta|\mathbf{X}, I = i)]^2 \Pr(I = i|\mathbf{X}) - \Big[\sum_i E(\theta|\mathbf{X}, I = i) \Pr(I = i|\mathbf{X})\Big]^2 \end{aligned} \tag{12}$$

Thus our posterior variance estimates for the parameters of interest account for uncertainty in the estimation of I. Our final point estimates are not byproducts of a single model with a fixed value of I, but averaged over possible models with weights proportional to the posterior probabilities $P(I|\mathbf{X})$.

3.3 Computational Details

1. Estimation of association parameters

None of the conditional distributions of the parameters has a standard distributional form and thus generating observations from the posterior distributions or calculating the posterior estimates is not automatic. We adopted a componentwise Metropolis-Hastings algorithm for each of the parameters.

Let η stand for a generic parameter, i.e., any of the β_{0i}, β_{1m}, ρ_{iu}, q_i and p_{lix} ($m = 1, 2$; $i = 1, \cdots, I$; $\forall \ \ l, x$). Let $L(\eta|\cdot)$ denote the full likelihood as given in (10) as a function of η given the data and all the other parameters. Let $\pi(\eta)$ be the prior distribution on η. In order to simulate observations from the full conditional distribution of η, namely $\pi(\eta|\cdot)$, we proceed as follows.

Step 1: Start with any reasonable initial value of η, say η_0. This is the current value of η.

Step 2: Generate a new value of η, say η^*, from a candidate density $g(\eta)$.

Step 3: Replace η_0 by η^* with probability min $\left\{1, \frac{\pi(\eta^*|\cdot)g(\eta_0)}{\pi(\eta_0|\cdot)g(\eta^*)}\right\}$. Retain the existing value of η_0 otherwise. Note that $\pi(\eta|\cdot) \propto \pi(\eta)L(\eta|\cdot)$. If the candidate density $\pi(\eta) = g(\eta)$, then the acceptance probability reduces to (after cancelation of the prior term with the identical candidate density term) min $\left\{1, \frac{L(\eta^*|\cdot)}{L(\eta_0|\cdot)}\right\}$.

2. Inference of the number of subpopulations I

The following algorithm (Pritchard *et al.*, 2000a) is used to sample from $\Pr(\mathbf{Z}, \mathbf{P}, \mathbf{Q})$. Starting with initial values of $\mathbf{Z}^{(0)}$, iterate the following steps for $k = 1, 2, \cdots$

Step 1. Sample $\mathbf{P}^{(k)}$ and $\mathbf{Q}^{(k)}$ from $\Pr(\mathbf{P}, \mathbf{Q}|\mathbf{X}, \mathbf{Z}^{(k-1)})$;

Step 2. Sample $\mathbf{Z}^{(k)}$ from $\Pr(\mathbf{Z}|\mathbf{X}, \mathbf{P}^{(k)}, \mathbf{Q}^{(k)})$;

Step 3. Update α using Metropolis-Hastings step (where we consider a uniform(0,10) prior to α).

Step 2 may be performed by simulating $z_l^{(j,c)}$ (population of origin of allele copy $x_l^{(j,c)}$), independently for each j, c and l from

$$\Pr(z_l^{(j,c)} = i|\mathbf{X}, \mathbf{P}) = \frac{q_i^{(j)} \Pr(x_l^{(j,c)}|\mathbf{P}, z_l^{(j,c)} = i)}{\sum_{i'=1}^{I} q_{i'}^{(j)} \Pr(x_l^{(j,c)}|\mathbf{P}, z_l^{(j,c)} = i')}, \tag{13}$$

where $\Pr(x_l^{(j,c)}|\mathbf{P}, z_l^{(j,c)} = i) = p_{il x_l^{(j,c)}}$.

4 Simulation

To illustrate our approach, we consider a scenario similar to the one in Satten *et al.* (2001) with an admixture of European and American Indian ancestry in Argentinean population. Sala *et al.* (1998, 1999) published allele frequency data on twelve short tandem repeat (STR) loci in Argentineans of European ancestry, as well as in three Argentinean American Indian aboriginal groups (Mapuche, Tehuelche, and Wichi) (Table 1). The Metropolitan population of Buenos Aires was studied and the population did not exhibit any significant difference from Hardy-Weinberg equilibrium. However, the STR allele frequency distributions are characterized by significant differences within and also between

different populations. We assume that Argentinean Europeans constituted 70% of a hypothetical target population and that each American Indian group constituted 10%.

We simulate a population such that all eleven additional mutually independent STR loci are in linkage equilibrium with the candidate gene for persons in the same subpopulation. Simulated data sets are constructed by using reasonable true values of the parameters. Specifically, by using the allele frequencies from Sala *et al.* (1999), we generate data on the candidate gene and other marker loci in a population that comprises four subpopulations. As in Satten *et al.* (2001), we select allele 3 of locus D6S366 as the disease-causing allele, with frequencies 0.277, 0.341, 0.446 and 0.557 in European, Mapuche, Tehuelche, and Wichi, respectively. Consider a biallelic candidate gene, i.e., a candidate gene with two alleles A (the disease-causing allele) and a (the non-disease-causing allele). The candidate gene G has 3 possible genotypes g_0, g_1 and g_2 corresponding to persons having zero (aa), one (Aa) and two (AA) copies of a disease-causing allele. If the frequency of the disease-causing allele in the ith subpopulation is

$$\rho_i = \Pr[G_l^c = A | Z = i] = 1 - \Pr[G_l^c = a | Z = i], \tag{14}$$

then by Hardy-Weinberg equilibrium, the probabilities of the genotypes of G are as the follows:

$$\begin{aligned} \Pr[G = g_0 | Z = i] &= (1 - \rho_i)^2; \\ \Pr[G = g_1 | Z = i] &= 2(1 - \rho_i)\rho_i; \\ \Pr[G = g_2 | Z = i] &= \rho_i^2. \end{aligned} \tag{15}$$

Finally, the disease status data that vary with changing frequencies of the disease-causing allele for each subpopulation are generated. As stated in Satten *et al.*(2001), persons who were homozygous for the disease-causing allele had an increased risk of disease corresponding to a log-odds ratio of 1.0 (relative risk = $\exp(1.0) = 2.72$); and persons who were heterozygous for the disease-causing allele had no increase in risk. This implies, in our notation, $\beta_{11} = 0$ and $\beta_{12} = 1.0$. The log odds of the rare disease (which implies that the control population mimics the whole population, and

$\Pr(G = g_m|D = 0, Z = i) \approx \Pr(G = g_m|Z = i)$) among persons with zero or one copy of the disease-causing allele was -5, -4, -3 and -3 in the European, Mapuche, Tehuelche, and Wichi populations, respectively. For the common disease with a higher prevalence rate, we assume that the log odds among persons with zero or one copy of the disease-causing allele was -2, -1.5, -1 and -1 in the European, Mapuche, Tehuelche, and Wichi populations, respectively.

The results we presented are based on a set of diffuse and mutually independent priors. We use $N(0, 9)$ prior on β_{0i} and β_{1m}, $Beta(0.5, 0.5)$ on ρ_i and a symmetric Dirichlet prior for the allele frequency parameters with all λ's being 0.5. For $(q_1, \cdots, q_I)$, we choose a Dirichlet$(\alpha, \cdots, \alpha)$ prior, with a U(0,10) hyperprior on α.

For each scenario, we generated 100 different data sets and obtained the parameter estimates by computing the model averaged posterior means for each simulated data set. In each replication of our simulation, we generated data for 125 (250) cases and 125 (250) controls from the above simulation strategy, followed by sampling the cases and controls from a larger random sample of subjects. For each replication, we ran multiple Markov chains, typically with $20000 - 30000$ iterations. The posterior means calculated for each replication were based on every tenth observation of the last 5000 observations in each chain, combined together to reduce auto-correlation. An estimate of the posterior variance was calculated based on the aggregate of the last 5000 values for each replication. We report average values for these quantities over the 100 replications. We also calculated an estimate of the mean squared error (MSE) corresponding to the estimates of each of our parameters of interest (say θ in general) based on the 100 replications. We considered this MSE, i.e., the squared deviations of the estimates from the true parameter, averaged over the 100 replications as a measure of performance of our method.

$$\text{MSE} = \frac{1}{100}\sum_{r=1}^{100}(\text{Posterior mean of } \theta \text{ in } r\text{-th replication} - \text{True value of } \theta)^2.$$

To examine the effect of the number of STR loci on the estimators, we analyzed the datasets with 250 subjects (125 cases and 125 controls) by (i) using all the additional loci and (ii) only the first six

additional loci. These two scenarios are labeled as X12 and X6 in Tables 2 and 4 respectively. By applying the methods stated in Section 2 (Pritchard *et al.* 2000a) and introducing a uniform prior for I ($I \in \{1, 2, 3, 4\}$), for each simulated dataset, first we obtain estimates of $P(I|\mathbf{X})$. For example, by (i), we obtain $P(I = 3|\mathbf{X}) = 0.2$ and $P(I = 4|\mathbf{X}) = 0.8$. Then the model averaged estimate of I is 0.2× 3+ 0.8× 4=3.8. The estimates of the association parameters are computed following (11) and (12). For the same dataset, the estimate of β_{12} is 1.09 for $I = 3$ and 1.02 for $I = 4$, thus the final model averaged estimate of β_{12} for that dataset is $1.09 \times 0.2 + 1.02 \times 0.8 = 1.034$. The results in Table 2 are obtained by averaging these estimates over the 100 simulated datasets, which shows that the posterior standard deviations of our model averaged estimates are typically smaller than the standard errors furnished by Satten *et al.* (2001) (we include the relevant numbers from Tables 2 and 3 of Satten *et al.* (2001) directly in Table 3 of the current paper). We realize that though our simulation settings are the same as of Satten *et al.* (2001), the two sets of estimates may not be exactly comparable as the two methods are not implemented on identical datasets, but still this might serve as a precursor for comparison purposes. Satten *et al.* (2001) do not provide MSE for their estimates over the replications. As a result we cannot compare the two procedures directly in terms of the MSE. As one might expect, when we increased the sample size to 500 (250 cases and 250 controls), adequate performance is achieved even with just the first six STR loci and the overall pattern of the results remain the same.

We also include the naive analysis completely ignoring additional multilocus information (denoted as X0 in Tables 2). One can note that the estimation results are much inferior if one ignores the genotypic information at a series of additional unlinked marker loci.

To show that our methods are not limited to the assumptions that either the disease is rare or the genotypes G are binary, we also analyzed a simulated dataset with 250 subjects (125 cases and 125 controls) and another with 500 subjects (250 cases and 250 controls) where the disease has a higher prevalence rate. The overall pattern of the results are fairly similar to the rare disease case. We note relatively smaller MSE's and posterior standard deviations for this common disease case as compared to the rare disease case. The results are presented in Table 4.

For analyzing the simulated data, we used the implicit prior belief that the source population may have 4 or less subpopulations, by putting a discrete uniform prior on $1, 2, 3, 4$ for I. However, we have also tried to put non-zero probability on a value of I greater than the true simulation value of 4, for instance, a discrete uniform prior on $1, \cdots, 8$. In this case, the estimates of the regression parameters β_{1m} appear to change very little even when I is estimated to be slightly greater than the true value used to generate the data (results are not provided). Pritchard *et al.* (2004) note that for situations where several values of I give similar estimates of $\log \Pr(\mathbf{X}|I)$, it is often the case that the smallest of these is 'correct'. In our practical implementation, we adopt a model selection perspective and try to obtain the smallest value of I that captures the major structure in the data.

5 Application to a real dataset

To illustrate our method, we apply our approach to explore genetic association of obesity and the β2AR candidate gene (for details of the study, please see Lin *et al.*, 2005). The β-adrenergic receptors (βAR) are known to play an important role in cardiovascular function and in response to drug. We analyze complete case data on 144 men and women who participated in this study and ignore the observations with missingness. Each of the participating subjects were genotyped for SNP markers at codon 16 within the β2AR gene, at codon 389 within the β1AR gene and at codon 492 within the α1A gene. The phenotypic information collected are weight and height of individuals, by which the body mass index (BMI) of each subject can be calculated. We define 'obese', i.e, $D = 1$ when BMI ≥ 30.0, and $D = 0$ otherwise. This leads to 85 undiseased and 59 diseased subjects in the dataset we consider.

Previous studies have detected possible association between polymorphism in the β2AR gene and obesity, the focus being particularly on codon 16 and codon 27 substitutions, but no association has been detected within β1AR gene or α1A gene (Johnson and Terra 2002, Lin *et al.* 2005, Takami *et al.* 1999). Therefore, we consider the β2AR gene as the candidate gene, denoted by G and the β1AR gene and the α1A gene as two other genes unrelated with the disease, denoted by $\mathbf{X} = (X_1, X_2)$. Note that

in this dataset, we only have the genotypic information regarding single polymorphisms in these three genes which have biallelic genotypes, generally expressed as $x = 0, 1, 2$. So the expression in (4) will be changed as $P(\mathbf{X}|Z = i) = \prod_{l=1}^{2} p_{lix}$, where p_{lix} is the proportion of persons in subpopulation i having genotype x ($x = 0, 1, 2$) corresponding to gene l.

We analyzed the data by considering genotypic information on all three genes (denoted by 'X2+G') and by only the candidate gene (denoted by 'X0+G'). Since in the real data, we do not know the true value of I, we should try to estimate the smallest value of I that captures the major substructure in the data, if any. To this end, we introduce a discrete uniform prior on $1, 2, \cdots, 15$ for I. We consider $(p_{li1}, p_{li2}, \cdots, p_{liI}) \sim$ Dirichlet$(0.5, 0.5, \cdots, 0.5)$, and for $(q_1, \cdots, q_I)$, we choose a Dirichlet$(\alpha, \cdots, \alpha)$ prior with a uniform hyperprior on α with range from 0 to 10. By applying the methods stated in Section 2, we first obtain inference on I. The principal findings are that with the inclusion of the two other genes, we detect some evidence of substructure with an estimate of I, as $\hat{I} = 3$, with $P(I = 3|\mathbf{X}) = 1$, whereas without these two genes and by only using G, we obtain $P(I = 1|\mathbf{X}) = 1$, implying $\hat{I} = 1$, i.e., no population substructure can be detected in the source population. In fact, the data came from a North American population with diverse ethnic composition of blacks, whites and others, so one could expect some latent population substructure in this data. The results of our analysis are presented in Table 5. In all the methods of analysis, the genetic factor does not appear to be a statistically significant risk factor. Our results suggest that codon 16 (Arg16Gly) polymorphisms of the β2AR gene is not a major contributing factor to obesity for this studied population. In fact, in Swedish Caucasians, Gln27Glu polymorphism at codon 27 of the β2AR gene was shown to be associated with obesity, but no such association was shown for Arg16Gly polymorphism at codon 16. None of the Gln27Glu and Arg16Gly polymorphisms of the β2AR gene were found to be a major contributing factor to obesity in Japanese men (Hayakawa *et al.* 2000). In the ordinary logistic regression model, with G as a categorical factor, we also find insignificance of G, (P-values 0.8591 and 0.1571 corresponding to G=1 and 2 respectively). Even after accounting for information in the other genes and population substructure, the effect of the candidate gene remains insignificant. Notice that the Bayesian HPD intervals are wider than the ordinary logistic model due to addition of

extra layer of uncertainty on I.

6 Discussion

In this article, we present an alternative Bayesian model for accounting for population substructure in genetic association studies. As compared to previous approaches, our model is advantageous in terms of the following aspects. First, it can estimate the number of subpopulations (I) that comprise the overall population. Although Satten *et al.* (2001) can also provide such an estimate, their approach is based on the grid procedure in which multiple different I's are fitted and the optimal one is then determined in terms of the minimum AIC. On the other hand, Pritchard *et al.* (2000b) estimated substructure and then conducted tests based on the imputed substructure. Based on marker and candidate gene information, our model estimates the posterior probabilities of I, which is then used in forming the final estimates of the relative risk parameters through model averaging. An additional advantage is that, unlike Satten *et al.*'s (2001) approach, our model does not rely on the assumption of the rare disease or the collapsing of multiple genotypes into binary genotypes, thus offers more power to study the genetic architecture of any type of diseases.

A new feature of our Bayesian analysis is the use of model averaging to estimate the regression coefficients. Rather than relying on one particular model with a fixed number of strata I, we have put a prior on I, and have estimated the regression parameters as the weighted average of their estimates for different values of I. The weights are proportional to the posterior probabilities of the different values of I. Thus we embed the substructure estimation together with inference on the association parameters in a unified Bayesian framework. The standard error of the relative risk estimates does incorporate the uncertainty in the estimation of I as reflected in (14). This is unlike the method proposed in Pritchard *et al.* (2000b) where the substructure is estimated first and tests are conducted based on the imputed substructure. Table 2 shows that our methods are comparable to those of Satten *et al.* (2001); however, since our set-up is different from that of Pritchard *et al.* (2000b), it is hard to compare the two methods directly in numerical sense. In principle, we do believe that combining inferences of the

substructure and association modeling will lend one more power in detecting association.

It should be pointed out that fewer additional markers are needed when the sample size is large. When additional marker loci are involved, the number of nuisance parameters (the allele frequencies of those loci for each subpopulation) in the model would increase, requiring more data to estimate them properly.

There remains the problem of handling marker loci in linkage disequilibrium with the candidate gene in our framework. According to Falush et al. (2003), there are three sources of linkage disequilibria (LD), mixture LD, admixture LD and background LD. The mixture LD arises from variation in individuals' ancestry and it can be measured by unlinked markers. The admixture LD occurs because of the correlation in ancestry among an extended genomic region. The background LD decays on a short scale and, therefore, occurs within a fine chromosomal structure. Pritchard et al. (2000a) modeled the mixture LD for association studies. In their "linkage" model, Falush et al. (2003) incorporated the "admixture LD" into the inference of population structure. The incorporation of the background LD is an interesting open question.

In summary, we have derived flexible Bayesian estimation techniques for disease-gene association in case-control studies by accounting for population structure. First, we applied Pritchard *et al.*'s (2000a) methods to infer population structure (i.e. estimating $P(I|\mathbf{X})$ and I) by using the genotypes of sampled individuals at a series of unlinked markers. Second, we propose a latent variable approach to estimate the association parameters, and account for population substructure using additional marker loci information as in Satten *et al.* (2001). The final results are calculated by the model averaging technique (as described in (11) and (12)) which combine inferences from the above two steps. Estimation results based on a simulated admixed population (mimicking the results presented in Sala *et al.* (1998)) show that the estimates of the relative risk parameters using additional mutilocus genetic information are superior to those when such information is not exploited. We also apply our method to a real dataset on obesity. The paper illustrates how the modeling tool of Bayesian model averaging can be effectively used to conduct posterior inference in an interesting application in human genetics.

ACKNOWLEDGMENTS

The authors will like to thank the editor, the associate editor and the referees for their careful reading of an earlier version of the manuscript and their constructive comments which led to substantial improvement in the content and presentation of the paper. We thank Dr. Julie Johnson for providing the dataset on obesity. The research of Bhramar Mukherjee was partially supported by NSA young investigator grant (H98230-06-1-0033). The research of Malay Ghosh was partially supported by NIH grant (R01 85414). Rongling Wu is grateful for the support from NSF grant (0540745) and NIH grant (R01 NS041670).

REFERENCES

Committee on DNA Forensic Science: An Update (1996) The evaluation of forensic DNA evidence. National Academy Press, Washington DC.

Freedman, M. L., Reich, D., Penney, K. L., McDonald, G. J., Mignault, A. A., Patterson, N., Gabriel, S. B., Topol, E. J., Smoller, J. W., Pato, C. N., Pato, M. T., Petryshen, T. L., Kolonel, L. N., Lander, E. S., Sklar, P., Henderson, B., Hirschhorn, J. N. and Altshuler, D. (2004). Assessing the impact of population stratification on genetic association studies. *Nature Genetics* **36**: 388–393.

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.

Hayakawa, T., Nagai, Y., Kahara, T., Yamashita, H., Takamura, T., Abe, T., Nomura, G. and Kobayashi, K. (2000) Gln27Glu and Arg16Gly polymorphisms of the beta2-adrenergic receptor gene are not associated with obesity in Japanese men. *Metabolism* **49**: 1215–8.

Hoggart, C. J., Parra, E. J., Shriver, M. D., Bonilla, C., Kittles, R. A., Clayton, D. G. and McKeigue, P.M. (2003) Control of confounding of genetic associations in stratified populations. *American Journal of Human Genetics* **72**: 1492-1504.

Hoggart, C. J., Shriver, M. D., Kittles, R. A., Clayton, D. G. and McKeigue, P. M. (2004) Design and analysis of admixture mapping studies. *American Journal of Human Genetics* **74**: 965-978.

Johnson, J. A., Terra, S.G. (2002) b-Adrenergic receptor polymorphisms: cardiovascular disease associations and pharmacogenetics. *Pharm Res* **19**: 17791787.

Knowler, W. C., Williams, R. C., Pettitt, D. J. and Steinberg, A. G. (1988). GM 3;5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. *American Journal of Human Genetics* **43**: 520–526

Lander, E. S., and Schork, N. J. (1994). Genetic dissection of complex traits. *Science* **265**: 2037-2048.

Lin, M., Aquilante, C., Johnson, J. A. and Wu, R. (2005) Sequencing drug response with HapMap

The Pharmacogenomics Journal **5**: 149156

Lynch, M., and Walsh, B. (1998) *Genetics and Analysis of Quantitative Traits*. Sinauer, Sunderland, MA.

Madigan, D. and Raftery, A. E. (1994). Model selection and model uncertainty in graphical models using Occam's Window. *J. Amer. Statist. Assoc.*, **89**: 1535-1546

Marchini, J., Cardon2, L. R., Phillips, M. S. and Donnelly, P., (2004). The effects of human population structure on large genetic association studies. *Nature Genetics* **36**: 512 - 517.

McKeigue, P. M. (2005). Prospects for admixture mapping of complex traits. *American Journal of Human Genetics* **76**:1-7.

Morton, N. E. and Collins, A. (1998). Tests and estimates of allelic association in complex inheritance. *Proc Nat Acad Sci, USA* **95**: 11389–11393.

Patterson, N., Hattangadi, N., Lane, B., Lohmueller, K. E., Hafler, D. A., Oksenberg, J. R., Hauser, S. L., Smith, M. W., OBrien, S. J., Altshuler, D., Daly, M. J. and Reich, D. (2004). Methods for High-Density Admixture Mapping of Disease Genes. *American Journal of Human Genetics.* **74**: 979–1000

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

Pritchard, J. K., Stephens, M., Rosenberg, N. A. and Donnelly, P. (2000b). Association mapping in structured populations. *American Journal of Human Genetics* **67**: 170–181.

Risch, N. and Merikangas, K. (1996). The future of genetic studies of complex diseases. *Science* **273**: 1516–517.

Sala, A., Penacino, G., Carnese, R., and Corach, D. (1999). Reference database of hypervariable genetic markers of Argentina: application for molecular anthropology and forensic casework. *Electrophoresis* **20**: 1733–739.

Sala, A., Penacino, G. and Corach, D. (1998). Comparison of allele frequencies of eight Loci from Argentinean Amerindian and European populations. *Human Biology* **70**: 937–947.

Satten, G. A., Flanders, W. D. and Yang, Q. (2001). Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. *American Journal of Human Genetics* **68**: 466–477.

Spielman, R.S. and Ewens, W.J. (1998) A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. *American Journal of Human Genetics* **62**: 450–458.

Spielman, R.S., McGinnis, R.E. and Ewens, W.J. (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *American Journal of Human Genetics* **52**: 506–516.

Sullivan, P. F., Eaves, L. J., Kendler, K. S. and Neale, M. C. (2001). Genetic case-control association studies in neuropsychiatry. *Archives of General Psychiatry* **58**: 1015–024.

Takami, S., Wong, Z. Y. H., Stebbing, M. and Harrap, S. B. (1999) Linkage analysis of glucocorticoid and b2-adrenergic receptor genes with blood pressure and body mass index *American Journal of Physiol Heart Circ Physiol* **276**: 1379–1384.

APPENDIX A

Estimation of $P(\mathbf{X}|I)$: An explanation to the approximation in equation (9)

From equation (7), by the strong law of large numbers,

$$\frac{1}{P(\mathbf{X}|I)} \stackrel{a.s.}{\to} E\Big(\frac{1}{Y}\Big),$$

where $Y = P(\mathbf{X}|\mathbf{Z}, \mathbf{P}, \mathbf{Q})$. This implies that,

$$-2\log P(\mathbf{X}|I) \stackrel{a.s.}{\to} 2\log E\Big(\frac{1}{Y}\Big), \tag{A.1}$$

Let $W = -2\log Y$=$[DV|\mathbf{X}]$, then

$$E\Big(\frac{1}{Y}\Big) = E_Y(\exp(-\log Y)) = M_W(\frac{1}{2}), \tag{A.2}$$

where $M_W(t)$ denotes the moment generating function of the distribution of W. By assuming that the deviance function $[DV|\mathbf{X}]$ is normal, i.e., $W \sim N(\mu, \sigma^2)$, by (A.2), we have,

$$E\Big(\frac{1}{Y}\Big) = \exp\big(\mu/2 + \sigma^2/8\big).$$

Hence by (A.1), and the fact that $\hat{\mu}$ and $\hat{\sigma}^2$ are consistent estimates of μ and σ^2, we have the approximation in (9).

Remark: Suppose we assume instead of normality of the deviance function, that $[DV|\mathbf{X}]$=$W = -2\log Y \sim Gamma(\mu^2/\sigma^2, \sigma^2/\mu)$, where μ and σ^2 are the mean and variance of W, and Gamma(a,b) denotes a Gamma distribution with shape parameter a and scale parameter b. Then by following steps exactly similar as above, one will obtain an analogue of (9) under the Gamma distributional assumption as,

$$-2\log P(\mathbf{X}|I) \approx -2\hat{\mu}^2/\hat{\sigma}^2 \log\Big(1 - \frac{\hat{\sigma}^2}{2\hat{\mu}}\Big), \quad \text{for } \frac{\hat{\sigma}^2}{\hat{\mu}} < 2.$$

APPENDIX B

The full conditional distributions of the parameters

Following the notations in Section 2, note that for subject j, G_j takes one of the values g_m, $m = 0, 1, \cdots, M$ and $\pi(\theta|\cdot) \propto \pi(\theta)L(\theta|\cdot)$. Therefore, the full conditional distributions for all the

parameters are given by :

$$\pi \quad (\beta_{0i}|\cdot) \propto \exp\Big\{-\frac{(\beta_{0i}-\mu_{\beta_{0i}})^2}{2\sigma^2_{\beta_{0i}}}\Big\} \times \Big\{\prod_{j=1}^{N}\sum_{i=1}^{I}\big[q_i \times \prod_{l=1}^{L}\prod_{c=1}^{2} p_{lix_l^c} \times \underbrace{\Pr(G_j=g_m|Z=i)}_{\text{functions of}\rho_{iu}\text{from (5)}}$$
$$\times \frac{\exp\{D_j \times (\beta_{0i}+\beta_{1m}\}}{1+\exp\{\beta_{0i}+\beta_{1m}\}}\big]\Big\}$$

$$\pi \quad (\beta_{1m}|\cdot) \propto \exp\Big\{-\frac{(\beta_{1m}-\mu_{\beta_{1m}})^2}{2\sigma^2_{\beta_{1m}}}\Big\} \times \Big\{\prod_{j=1}^{N}\sum_{i=1}^{I}\big[q_i \times \prod_{l=1}^{L}\prod_{c=1}^{2} p_{lix_l^c} \times \underbrace{\Pr(G_j=g_m|Z=i)}_{\text{functions of}\rho_{iu}\text{from (5)}}$$
$$\times \frac{\exp\{D_j \times (\beta_{0i}+\beta_{1k})\}}{1+\exp\{\beta_{0i}+\beta_{1k}\}}\big]\Big\}$$

$$\pi \quad (q_1,\cdots,q_I|\cdot) \propto \Big\{\prod_{i=1}^{I} q_i^{(\alpha-1)}\Big\} \times \Big\{\prod_{j=1}^{N}\sum_{i=1}^{I}\big[q_i \times \prod_{l=1}^{L}\prod_{c=1}^{2} p_{lix_l^c} \times \underbrace{\Pr(G_j=g_m|Z=i)}_{\text{functions of}\rho_{iu}\text{from (5)}}$$
$$\times \frac{\exp\{D_j \times (\beta_{0i}+\beta_{1m})\}}{1+\exp\{\beta_{0i}+\beta_{1m}\}}\big]\Big\}$$

$$\pi \quad (\rho_i|\cdot) \propto \Big\{\rho_i^{(a-1)} \times (1-\rho_i)^{(b-1)}\Big\} \times \Big\{\prod_{j=1}^{N}\sum_{i=1}^{I}\big[q_i \times \prod_{l=1}^{L}\prod_{c=1}^{2} p_{lix_l^c} \times \underbrace{\Pr(G_j=g_m|Z=i)}_{\text{functions of}\rho_{iu}\text{from (5)}}$$
$$\times \frac{\exp\{D_j \times (\beta_{0i}+\beta_{1m})\}}{1+\exp\{\beta_{0i}+\beta_{1m}\}}\big]\Big\}$$

$$\pi \quad (p_{li1},\cdots,p_{lix},\cdots,p_{liX_l}|\cdot) \propto \Big\{\prod_{x=1}^{X_l} p_{lix}^{(\lambda p_{lix}-1)}\Big\} \times \Big\{\prod_{j=1}^{N}\sum_{i=1}^{I}\big[q_i \times \prod_{l=1}^{L}\prod_{c=1}^{2} p_{lix_l^c}$$
$$\times \underbrace{\Pr(G_j=g_m|Z=i)}_{\text{functions of}\rho_{iu}\text{from (5)}} \times \frac{\exp\{D_j \times (\beta_{0i}+\beta_{1m})\}}{1+\exp\{\beta_{0i}+\beta_{1m}\}}\big]\Big\}.$$

Table 1: Allele frequencies for Twelve STR loci in the four Argentinean subpopulations, cited from Sala *et al.* (1998) and Satten *et al.* (2001).

Locus	Argentinian Europeans	Mapuche	Tehuelche	Wichi
D6S366	0.082	0.091	0.143	0
	0.204	0.114	0.071	0
	0.277	0.341	0.446	0.557
	0.119	0.136	0.036	0.086
	0.091	0.125	0.036	0.029
	0.183	0.159	0.143	0.200
	0.028	0.011	0.018	0.071
	0.015	0.023	0.107	0.057
FABP	0.589	0.683	0.732	0.485
	0.110	0.058	0.107	0.162
	0.300	0.260	0.161	0.353
CSF1PO	0.330	0.266	0.339	0.226
	0.313	0.282	0.232	0.194
	0.298	0.367	0.411	0.581
	0.059	0.085	0.018	0
F13A	0.151	0.222	0.357	0.173
	0.060	0.122	0.125	0.077
	0.202	0.122	0.054	0.346
	0.209	0.178	0.143	0.115
	0.325	0.344	0.304	0.288
	0.053	0.011	0.017	0
FESFPS	0.260	0.170	0.143	0.257
	0.420	0.500	0.714	0.543
	0.247	0.284	0.107	0.043
	0.073	0.045	0.036	0.157
THO1	0.233	0.526	0.286	0.132
	0.250	0.298	0.429	0.721
	0.105	0.009	0.018	0
	0.185	0.026	0.089	0.015
	0.226	0.140	0.179	0.132
HPRTB	0.032	0	0	0
	0.179	0.032	0.091	0
	0.317	0.323	0.227	0.357
	0.285	0.403	0.591	0.167
	0.137	0.242	0.091	0.357
	0.050	0	0	0.119
VWA	0.063	0.0096	0.036	0.014
	0.099	0.077	0.054	0.014
	0.294	0.577	0.429	0.514
	0.297	0.125	0.214	0.343
	0.246	0.212	0.268	0.114
D13S317	0.090	0.020	0	0
	0.160	0.240	0.15	0.464
	0.060	0.070	0.05	0.179
	0.290	0.120	0.15	0.089
	0.250	0.260	0.3	0.089
	0.100	0.180	0.225	0.179
	0.040	0.110	0.125	0
D7S820	0.156	0.070	0.050	0
	0.115	0.050	0.050	0.070
	0.276	0.220	0.175	0.125
	0.245	0.420	0.525	0.450
	0.159	0.210	0.200	0.250
	0.046	0.030	0	0.105
D16S539	0.156	0.110	0.225	0.125
	0.100	0.130	0.075	0.232
	0.294	0.240	0.100	0.321
	0.252	0.370	0.550	0.250
	0.195	0.150	0.050	0.071
RENA4	0.772	0.728	0.881	0.690
	0.074	0.229	0.023	0
	0.153	0.041	0.095	0.310

Table 2: The results of simulated rare-disease data with marker loci in linkage equilibrium with the candidate gene D6S366. Ratio of the sample sizes of cases to controls is $125/125$ and $250/250$. X12 and X6, represent that the parameters were estimated by using the twelve and the first six additional marker loci, respectively. X0 is the analysis without using any additional marker loci. Here Mean and posterior standard deviation refers to the average of the Bayes estimates and posterior standard deviations obtained in 100 replications, whereas MSE is the estimated mean squared error based on 100 replications.

Sample size	Model		β_{11}	β_{12}	I
		True value	0.0000	1.0000	4
125/125	X12	Mean	-0.0475	1.1093	3.8178
		MSE	0.1497	0.0765	0.1802
		Post. std. dev.	0.3126	0.2638	0.3854
	X6	Mean	-0.1095	1.1028	3.6403
		MSE	0.2005	0.0986	0.3540
		Post. std. dev.	0.3277	0.3127	0.4763
	X0	Mean	-0.3380	0.8855	4.0000
		MSE	1.2277	0.4982	
		Post. std. dev.	1.5982	1.0677	
250/250	X12	Mean	0.0005	1.0966	3.7873
		MSE	0.0546	0.0551	0.2107
		Post. std. dev.	0.2704	0.1592	0.4089
	X6	Mean	0.0051	1.1035	3.5415
		MSE	0.0631	0.0582	0.4572
		Post. std. dev.	0.3127	0.1952	0.4994
	X0	Mean	-0.2766	0.9489	4.0000
		MSE	1.2603	0.4330	
		Post. std. dev.	1.4152	0.9236	

Table 3: The results of simulated rare-disease data with marker loci in linkage equilibrium with the candidate gene D6S366 which are analyzed by Satten *et al.* (2001). $125/125$ and $250/250$ denote ratio of the sample sizes of cases to controls. X12 and X6 represent that the parameters were estimated by using the twelve and the first six of the additional marker loci, respectively. Here Mean and standard error refers to the average of the estimates and standard errors obtained in 500 replications.

Sample Size	Model		β_{11}	β_{12}	I
		True value	0.000	1.000	4
125/125	X12	Mean	0.061	1.006	3.53
		Std. err.	0.293	0.453	0.76
	X6	Mean	0.023	0.883	3.32
		Std. err.	0.865	1.718	0.69
	Crude Analysis*	Mean	0.366	1.760	1.00
		Std. err.	0.285	0.370	
250/250	X6	Mean	0.023	0.962	3.37
		Std. err.	0.226	0.394	0.61

* Ignore stratification and analyze data without additional marker loci.

Table 4: The results of simulated common-disease data with marker loci in linkage equilibrium with the candidate gene D6S366. Ratio of the sample sizes of cases to controls is $125/125$ and $250/250$. X12 and X6, represent that the parameters were estimated by using the twelve and the first six additional marker loci, respectively. X0 is the analysis without using any additional marker loci. Here Mean and posterior standard deviation refers to the average of the Bayes estimates and posterior standard deviations obtained in 100 replications, whereas MSE is the estimated mean squared error based on 100 replications.

Sample size	Model		β_{11}	β_{12}	I
		True value	0.0000	1.0000	4
125/125	X12	Mean	-0.0062	1.1116	3.8492
		MSE	0.1106	0.1005	0.1456
		Post. std. dev.	0.3152	0.1607	0.3523
	X6	Mean	0.0017	1.1299	3.6279
		MSE	0.1173	0.1371	0.3634
		Post. std. dev.	0.3488	0.2766	0.4766
250/250	X12	Mean	0.0023	1.0928	3.9331
		MSE	0.0600	0.0551	0.0461
		Post. std. dev.	0.2165	0.1806	0.2412
	X6	Mean	0.0191	1.1051	3.6228
		MSE	0.0408	0.0470	0.3748
		Post. std. dev.	0.2627	0.1991	0.4846

Table 5: The results of real data analysis with the posterior mean (Estimate), posterior standard deviation and 95% highest posterior density (HPD) interval (MLE and confidence interval (CI) for the ordinary logistic regression model).

Model		β_{11}	β_{12}	I
X2+G	Estimate	-0.0895	0.7165	3
	Post std.dev.	0.3997	0.5201	$\cdot^*$
	HPD	(-0.8619,0.6831)	(-0.2996,1.7259)	
X0+G	Estimate	-0.1206	0.7433	1
	Post std.dev.	0.4515	0.5602	$\cdot^*$
	HPD	(-1.0028,0.7865)	(-0.3339,1.8303)	
Ordinary logistic regression with only G as covariate	Estimate	-0.0668	0.7143	
	Std.err.	0.3765	0.5048	
	CI	(-0.8047,0.6711)	(-0.2751,1.7037)	

*:All of the posterior probability concentrated on a single value of I, thus we are unable to obtain estimates of posterior variance.