

A Bayesian Approach for Incorporating Variable Rates of Heterogeneity in Linkage Analysis

Swati Biswas* and Shili Lin**

*Department of Biostatistics, School of Public Health, University of North Texas Health Sciences Center, Fort Worth, TX 76107.

**Department of Statistics, The Ohio State University, Columbus, OH 43210.

Address for correspondence: Swati Biswas, PhD

Department of Biostatistics, School of Public Health
University of North Texas Health Sciences Center
3500 Camp Bowie Blvd.
Fort Worth, TX 76107-2699
Tel: (817) 735-5442
Fax: (817) 735-2314
Email: sbiswas@hsc.unt.edu

Swati Biswas is Assistant Professor, Department of Biostatistics, University of North Texas Health Sciences Center, Fort Worth, TX 76107. (email: sbiswas@hsc.unt.edu); and Shili Lin is Professor, Department of Statistics, The Ohio State University, Columbus, OH 43210 (email: shili@stat.ohio-state.edu). This work was supported in part by NSF grants DMS-9971770 and DMS-0366800, and NIH grant 1R01HG002657-01A1. The authors sincerely thank Dr. Chris Amos for providing the lung cancer dataset of Genetic Epidemiology of Lung Cancer Consortium, which is supported by U.S. Public Health Service National Cancer Institute grant R01 CA76293. The authors are grateful to the editor, an associate editor and two anonymous reviewers for constructive comments and suggestions, which led to improvement and clearer presentation of the manuscript.

Abstract

A widely used approach for dealing with locus heterogeneity in linkage analysis is based on mixture likelihood, in which a single mixing (heterogeneity) parameter represents the probability that each family is of linked type. However, in general, different types of families exhibit different heterogeneity levels. To incorporate this variability, we propose a new approach, wherein each family has its own heterogeneity parameter representing the probability that it is of linked type. These parameters are nuisance parameters while the main parameter of interest is the location of the disease gene, if there is any. We model the problem in the Bayesian framework and implement it using the Markov chain Monte Carlo (MCMC) methodology. In particular, we utilize the reversible jump MCMC sampler to move between the two models: linkage and no linkage. We first estimate the posterior probability of linkage on a chromosome and the corresponding Bayes factor. If linkage is inferred, the location of the disease gene along with its credible set is estimated. The asymptotic joint distribution of the estimators is derived. We show that this approach is more powerful than the currently used approach in detecting linkage while the two approaches have comparable false positive rates. The proposed method was applied to a lung cancer dataset of Genetic Epidemiology of Lung Cancer Consortium and an asthma dataset consisting of three samples from Genetic Analysis Workshop 12. Since both lung cancer and asthma are complex traits with heterogeneous genetic predisposition, they provide suitable applications for the proposed method.

Keywords

heterogeneity; admixture approach; Markov chain Monte Carlo; reversible jump; lung cancer; asthma

1 INTRODUCTION

Locus heterogeneity is one of the major reasons for limited success of linkage analysis in mapping genes that influence complex genetic traits. This phenomenon refers to the situations when a disease is caused in some families by one gene while in other families it is caused by some other gene and/or non-hereditary factors. The most popular approach currently used for incorporating heterogeneity is the admixture approach (Ott 1999). It uses a single heterogeneity (mixing) parameter, α , to model the probability that the disease-causing gene of a family is linked to a reference map of markers, i.e., a family is of linked type. A mixture likelihood based on the two parameters, α , and the location of disease gene, is formed. The overall LOD score (\log_{10} of likelihood ratio) based on this mixture likelihood, referred to as the heterogeneity LOD (HLOD) score, is then maximized with respect to the parameters (Ott 1999). Notwithstanding its wide usage, several limitations of the admixture approach have been pointed out in the literature (see e.g., Janssen, Halley, and Sandkuijl 1997; Vieland, Wang, and Huang 2001; Whittemore and Halpern 2001; Vieland and Logue 2002). In particular, we have shown recently that a single heterogeneity parameter is insufficient to capture the variable rates of heterogeneity across different types of families, and hence can lead to inconsistent estimators (Lin and Biswas 2004; Biswas and Lin 2004).

In this article, we propose a new approach by assigning to each family its own heterogeneity (α) parameter that denotes the probability that it is of linked type. However, the large number of parameters involved in such a model poses a big challenge that cannot be handled by the traditional approaches. A viable alternative to circumvent this problem is provided by the Markov chain Monte Carlo (MCMC) methodology. These methods have significantly broadened the scope of modeling the complexities in many applications (Gilks, Richardson, and Spiegelhalter 1996). To take advantage of the MCMC techniques, we cast our problem in a Bayesian framework. The α parameters are nuisance parameters while the main parameter of interest is the location of the disease gene. Being in the Bayesian setting, we focus on obtaining the posterior distributions of the parameters. These distributions are estimated by drawing a large number of dependent samples via realizations of a Markov chain whose stationary distribution is the joint posterior distribution of all parameters. Due to the

need of switching between subspaces of different dimensionality created by the linked and unlinked states, we employ the reversible jump MCMC method (Green 1995). The Bayesian approach has been evaluated and compared with the usual two-parameter admixture approach through simulations. We also performed a sensitivity analysis on the specifications of the priors. In addition, the robustness of this approach to model misspecification is studied by analyzing the Genetic Analysis Workshop (GAW) 13 simulated dataset, which was modeled after the Framingham Heart Study (Almasy et al. 2003). We applied the proposed Bayesian approach to two real datasets: a lung cancer dataset of Genetic Epidemiology of Lung Cancer Consortium (GELCC; Bailey-Wilson et al. 2004) and the GAW 12 asthma dataset (Wijsman et al. 2001).

Lung cancer has a severe mortality rate reflected in the fact that it claims more lives than breast, colon, and prostate cancers combined. While smoking and other behavioral and environmental factors play a major role in lung cancer, numerous studies point to involvements of genetic factors in familial lung cancer (FLC). Bailey-Wilson et al. (2004) reviewed these studies and conducted a genome-wide linkage analysis to identify susceptibility genes. They used the data collected by the FLC recruitment sites of GELCC and computed HLOD. No significant linkage signal was detected when the full dataset was analyzed. This suggests that perhaps there are varying rates of heterogeneity among the families in the full dataset, and HLOD, being unable to account for it, suffers power loss. Since our approach is designed to explicitly account for variable levels of heterogeneity among different families, we anticipate that it may have the power to detect linkage in the whole dataset.

Asthma is one of the most common chronic childhood diseases in the developed countries. It is a complex disease caused by the interplay of multiple genetic and/or environmental factors. Although several studies have been carried out to dissect the genetic mechanism of asthma, many of them have produced contradictory results; this undesirable situation is largely attributed to the presence of heterogeneity (Wijsman et al. 2001). Of the three asthma samples that are analyzed in the current paper, two comprise Caucasian families while the other one has African-American families. Eerdewegh et al. (2001) found that an overall HLOD obtained by pooling the three samples suffers substantial power loss, causing its failure to find linkage signals. The most plausible explanation is the presence of different

ethnicities, which often exhibit different levels of heterogeneity. Indeed, a common conclusion from various investigations of the three samples is that, unlike the two Caucasian samples, the African-American sample consists mostly of unlinked type of families for chromosome 6 (Eerdewegh et al. 2001; Wang, Huang, Logue, and Vieland 2001; Biswas and Lin 2004). The Bayesian paradigm lends a natural way to incorporate this a priori information about heterogeneity across ethnicities into the analysis. We expect that such a strategy will help in identifying linkage signals, thereby demonstrating the flexibility of the method.

2 METHODS

2.1 Model

We focus on one chromosome at a time to search for disease gene(s). Suppose there are k families in the sample. Let α_j be the probability that the j th family is of linked type, $j = 1, \dots, k$, and d be the position of the disease gene on the chromosome. Also let x_j and $L_j(d|x_j)$ denote the observed data and the corresponding (homogeneity) likelihood, respectively, of the j th family, $j = 1, \dots, k$. Denote $\mathbf{x} = (x_1, x_2, \dots, x_k)$ and $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_k)$. Here $\boldsymbol{\alpha}$ is a set of nuisance parameters. Then the likelihood is given by

$$L(\boldsymbol{\alpha}, d|\mathbf{x}) = \prod_{j=1}^k [\alpha_j L_j(d|x_j) + (1 - \alpha_j) L_j(\infty|x_j)].$$

Here $L_j(\infty|x_j)$ denotes the likelihood of the data when the disease gene is at distance “ ∞ ” from the marker map, i.e., it is unlinked to the chromosome. Suppose there are N distances at which the LOD scores are calculated on the chromosome. We label the indexes of these N distances as $1, \dots, N$. Let I_d denote the index of distance d . Then $I_d \in \{1, \dots, N\}$. Note that computing LOD scores at a pre-specified grid of points across the chromosome is a usual practice in linkage analysis since analytical forms of likelihoods are usually not available except for a few simple pedigree structures (Ott 1999).

Prior distribution for d consists of two components: one when there is a disease gene present on the chromosome of interest ($d < \infty$) and another when there is no such disease gene ($d = \infty$), with prior probabilities $\pi_{d < \infty}$ and $\pi_{d = \infty}$, respectively. Further, for $d < \infty$,

there is a probability distribution of d (location of disease gene) denoted by $\pi_d(I_d)$. This is a probability mass function on the N distances at which the LOD scores, and hence the likelihoods, are calculated. Next, let the prior distribution of α_j ($j = 1, \dots, k$) be $\pi_j(\alpha_j)$. These distributions may not be identical but they are assumed to be independent of each other and are also independent of the distribution of d when $d < \infty$. Note that α_j 's are meaningless when $d = \infty$. So, now we can write the joint posterior distribution of $(\boldsymbol{\alpha}, d)$ as

$$\begin{aligned} \pi(\boldsymbol{\alpha}, d|\mathbf{x}) \propto & \left[\prod_{j=1}^k \pi_j(\alpha_j) \right]^{I(d < \infty)} \left[\pi_{d < \infty} \pi_d(I_d) I(I_d \in \{1, \dots, N\}) + \pi_{d = \infty} I(d = \infty) \right] \\ & \times \prod_{j=1}^k [\alpha_j L_j(d|x_j) + (1 - \alpha_j) L_j(\infty|x_j)], \end{aligned}$$

where the likelihood factor reduces to $\prod_{j=1}^k L_j(\infty|x_j)$ when $d = \infty$, independent of the α parameters. The goal is to obtain the posterior distributions of α_j 's and d . In the next section, we propose an MCMC scheme that generates a Markov chain whose stationary distribution is the joint posterior distribution. To be more precise, we note that there are, in fact, two posterior distributions depending on whether $d < \infty$ or $d = \infty$ (point mass), corresponding to the linked and unlinked models, respectively. Notice that $\pi_{d < \infty}$ and $\pi_{d = \infty}$ are the prior probabilities of these two models. An initial part of the chain (burn-in period), after which the chain is believed to have navigated into the target posterior distribution, is discarded. The chain is then allowed to run for a sufficiently large number of iterations so that it covers the target distribution reasonably well.

2.2 MCMC Sampling Scheme

The values of $d < \infty$ (linked subspace) and $d = \infty$ (unlinked subspace) lead to two different models with different numbers of parameters. The linked subspace (L) consists of $k + 1$ parameters $(\boldsymbol{\alpha}, d)$ while the unlinked subspace (U) has no parameters, since $d = \infty$ renders the α 's meaningless. So, we need to use a sampler that allows moves to be made between different models with different number of parameters. The sampling scheme described here uses a reversible jump MCMC sampler (Green 1995; Richardson and Green 1997) to enable moves to be made from L to U and vice versa. For updating the α_j 's and d , when the

chain is currently in the L subspace, we use the Gibbs sampler (Geman and Geman 1984) and Metropolis-Hastings algorithm (Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller 1953; Hastings 1970), respectively.

Suppose the probability of staying in the L subspace given that the current state is in L, denoted by $P(L|L)$, is λ_1 . Similarly, let $P(U|U)$ be λ_2 . Hence $P(U|L)$ and $P(L|U)$ are $1 - \lambda_1$ and $1 - \lambda_2$, respectively. At iteration t , first we randomly choose whether a move to the other subspace is proposed or not, according to one of the above probabilities. Depending on the current and the proposed states, there are four possible types of moves: $L \rightarrow L$, $L \rightarrow U$, $U \rightarrow U$, and $U \rightarrow L$, where the letter on the left (right) side of \rightarrow is the subspace of the current (proposed) state. We note that the moves as described here resemble a random walk scheme and are slightly different from the sweeping moves described in some of the reversible jump algorithm literature (e.g., see Richardson and Green 1997). However, there is a correspondence between the two: the moves $U \rightarrow L$ and $L \rightarrow U$, which alter the dimension of the parameter space, would be called birth and death of a gene, respectively, in the usual terminology of the reversible jump algorithm; the other two moves, $L \rightarrow L$ and $U \rightarrow U$ form the so-called updating of parameters moves. Detailed description of these four move types is given in Appendix A.

Starting with some initial values for $(\boldsymbol{\alpha}, d)$, we run the Markov chain for a burn-in period of B iterations. After this, the chain is run for another T iterations. The posterior distributions are estimated based on these T iterations.

2.3 Inference Using Posterior Distribution

Our interest lies in drawing inference about the disease gene location, so we focus on the marginal posterior distribution of d . It can be viewed as a mixture of point mass at ∞ and a distribution for the location of the disease gene on the chromosome, under linkage, i.e.,

$$\pi(d|\mathbf{x}) = P(d = \infty|\mathbf{x})I(d = \infty) + P(d < \infty|\mathbf{x})P(d|d < \infty, \mathbf{x})I(d < \infty).$$

$P(d = \infty|\mathbf{x})$ and $P(d < \infty|\mathbf{x})$ can be interpreted as the posterior probability of no linkage and linkage on the chromosome, respectively. Let t_i denote the number of times the distance with index i is in the MCMC output of the T iterations, $i = 1, \dots, N$, thus $\sum_i t_i$ is the

number of times that the chain stays in the L subspace. We also denote by t_∞ the number of times the chain is in the U subspace. Then, the estimate of the posterior distribution of d as obtained from the MCMC output can be written as

$$\begin{aligned}\hat{\pi}(d|\mathbf{x}) &= \frac{t_\infty}{T}I(d = \infty) + \frac{T - t_\infty}{T} \sum_{i=1}^N \frac{t_i}{T - t_\infty} I(I_d = i) \\ &= \hat{P}(d = \infty|\mathbf{x})I(d = \infty) + \hat{P}(d < \infty|\mathbf{x}) \sum_{i=1}^N \hat{P}(d = d_i|d < \infty, \mathbf{x})I(I_d = i),\end{aligned}$$

where d_i is the distance with index i , $i = 1, \dots, N$. The first term in the above expression is a point mass at $d = \infty$. From the above representation, it is clear that $\hat{p} = (T - t_\infty)/T$ (proportion of times the chain is in the L subspace) is the estimated posterior probability of linkage. The larger the value of \hat{p} , the greater the signal for linkage on that chromosome. Under linkage, the mean of the estimated posterior distribution of d when the chain is in the L subspace, given by

$$\hat{m} = \frac{\sum_{i=1}^N t_i d_i}{\sum_{i=1}^N t_i} = \frac{\sum_{i=1}^N t_i d_i}{T - t_\infty},$$

is an estimate of the location of a disease gene.

From the above discussion, we see that the estimators that summarize the posterior distribution of d are (\hat{p}, \hat{m}) . Note that \hat{m} is meaningful only under linkage. At this point, it is worth noting that although the Markov chain output consists of the sequence $\{(\boldsymbol{\alpha}_{(t)}, d_{(t)}), t = 1, 2, \dots, T\}$, for some t , $\boldsymbol{\alpha}_{(t)}$ vanishes and $d_{(t)} = \infty$ (when the chain moves to the U subspace). Further, since d is the parameter of interest, the estimators (\hat{p}, \hat{m}) are based only on the $\{d_{(t)}, t = 1, 2, \dots, T\}$ component of the MCMC output. Although this may seem a bit peculiar in the usual MCMC scenario, this peculiarity alleviates when we view it from the perspective of model determination, where the two models are linkage and no linkage, each having a different number of parameters. One way of comparing between the linkage and no linkage models is via the Bayes Factor (BF), which can be estimated by

$$\widehat{BF} = \frac{\hat{p}/(1 - \hat{p})}{\pi_{d < \infty}/\pi_{d = \infty}} = \frac{\hat{p}/(1 - \hat{p})}{\pi_{d < \infty}/(1 - \pi_{d < \infty})}.$$

An attractive feature of BF is that it is theoretically independent of the prior, $\pi_{d < \infty}$ (Richardson and Green 1997); inference under BF does not need to reference the prior used. For a

fixed $\pi_{d < \infty}$, \widehat{BF} is a strictly increasing function of \hat{p} , and so \widehat{BF} exceeding a pre-specified threshold BF_0 can be used to conclude linkage.

It can be shown that the joint asymptotic distribution of (\hat{p}, \hat{m}) is Bivariate Normal (BVN) with means (p, m) , where p is the true posterior probability of linkage and m is the true location of the disease gene *under linkage*. The proof of this result along with the expressions for variances and correlation are in Appendix B.

In addition to estimating the disease gene location under linkage (m) by \hat{m} , we also obtain its interval estimates. We explore two ways of getting such an estimate. First, we could get a classical confidence interval by using the asymptotic normality property of \hat{m} and estimating the SE of \hat{m} by batch means method. In fact, we can get more than this - if we use the asymptotic bivariate normality property of (\hat{p}, \hat{m}) , then we can get simultaneous confidence intervals for (p, m) . Another way of obtaining an interval estimate for m is to compute a Bayesian credible set. A level $(1 - \gamma)\%$ credible set is given by the interval covered between $\gamma/2$ th and $(1 - \gamma)/2$ th quantiles of the posterior distribution of d under linkage.

2.4 Diagnostics

For drawing valid inference from an MCMC procedure, we need to ensure that mixing of the chain is good and it converges to its stationary distribution. Here we use two diagnostic tools. The first one is a simple graph wherein we plot the estimators \hat{p} and \hat{m} obtained from dispersed starting points at every 100th iteration; this allows us to visually monitor the chain.

The second one, a more formal tool, is to test independence of batch means. This also serves the purpose of checking the assumption of independent batch means needed for the batch means approach of computing variances. Divide the total chain (excluding the burn-in) into R batches, each of length L . Let $d_r^{(l)}$ denote the l th value of distance in the r th batch, $l = 1, 2, \dots, L$, $r = 1, 2, \dots, R$. For batch means $y_r = \sum_{l=1}^L d_r^{(l)} I(d_r^{(l)} < \infty) / L$, $r = 1, \dots, R$, we compute the Von Neumann statistic (Kleijnen 1987) given by

$$q = \frac{\sum_{r=1}^{R-1} (y_r - y_{r+1})^2}{\sum_{r=1}^R (y_r - \bar{y})^2},$$

where $\bar{y} = \sum_{r=1}^R y_r / R$. Its standardized version is

$$Z_y = \frac{q - 2}{\sqrt{\frac{4(R-2)}{R^2-1}}},$$

which has approximately $N(0, 1)$ distribution under the null hypothesis of independent batch means. So the null hypothesis is rejected at α level of significance if $|Z| > z_{\alpha/2}$. More details can be found in Kleijnen (1987). Similarly we define the statistic Z_p for batch means $p_r = \sum_{l=1}^L I(d_r^{(l)} < \infty) / L$, $r = 1, \dots, R$. We carry out the test of independence by dividing the MCMC output into 100 batches and computing Z_y and Z_p while the batch means SE is obtained using 30 batches, following Kleijnen (1987) and Schmeiser (1982).

3 SIMULATION RESULTS

The following is a description of the data and the simulation models used in all subsections of this section. We simulated datasets along the lines of the GAW13 simulated data (Almasy et al. 2003) and used GENEHUNTER (Kruglyak, Daly, Reeve-Daly, and Lander 1996) to compute the LOD scores that are needed as input to our procedure. A sample consists of 274 GAW13 pedigrees (nuclear and extended families) which include the pedigree structures that GENEHUNTER can analyze as a whole and some trimmed pedigrees. The family sizes range from 7 to 19 members (mean = 11.2). For all simulations, except the homogeneity model to be described later, we randomly chose the probability that a family is of a linked type from the Beta(3,2) distribution. In addition to the pedigree structure, the observed data for each family consist of marker genotypes and phenotypes (affection statuses) for family members. Biswas, Papachristou, Irwin, and Lin (2003) labeled the affection status of about 45% of the people to be missing for one of the replicates of the GAW13 data; we kept those people's affection status missing in our simulated samples also. Unless stated otherwise, we used the chromosome 21 markers with the same allele frequencies and inter-marker distances as provided in GAW13. Specifically, there are six markers on chromosome 21 at locations 0, 10.02, 22.74, 36.20, 40.07 and 59.53 cMs (adjusted such that the first marker is at 0 cM) with respective heterozygosities of 0.63, 0.73, 0.77, 0.74, 0.88 and 0.69. For most of the simulations, we used the following disease model, where D is the disease

allele and “Aff” denotes affected with the disease under study: disease allele frequency $P(D) = 0.2$, and the penetrances for the three genotypes DD, Dd , and dd are $P(\text{Aff} | DD) = 0.7$, $P(\text{Aff} | Dd) = 0.5$, and $P(\text{Aff} | dd) = 0.05$. This model lies in-between dominant and recessive models, the two classical Mendelian disease models, and we refer to it as a “intermediate” model. We also carried out some simulations with a “dominant” model: $P(D) = 0.05$, $P(\text{Aff} | DD) = P(\text{Aff} | Dd) = 0.5$, $P(\text{Aff} | dd) = 0.05$, and a “recessive” model: $P(D) = 0.1$, $P(\text{Aff} | DD) = 0.7$, $P(\text{Aff} | Dd) = P(\text{Aff} | dd) = 0.05$. In addition, we also simulated under a “homogeneity” model, i.e., all families being of linked type. For this simulation, we used the recessive disease model described above.

We use the following prior distributions: $\pi_j(\alpha_j)$ is taken to be $U(0, 1)$, $j = 1, \dots, k$, $\pi_{d < \infty} = 1/22$, $\pi_{d = \infty} = 21/22$, and $\pi_d(I_d) = 1/N$, $I_d = 1, \dots, N$. The rationale for the chosen prior for d is as follows. Humans have 22 pairs of autosomes and one pair of sex chromosomes. If we believe that the disease is not caused by a gene on the sex chromosomes, then the probability that a randomly chosen chromosome has the disease gene is $1/22$. Since we compute LOD scores at N points on a chromosome, we evenly distribute the probability mass at those N points. In some real data applications, the investigator may have some a priori information about d and/or α 's; such knowledge can be readily incorporated into informative priors for the parameters. We used the non-informative priors for simulations (except in the sensitivity study) to avoid influence of any particular prior on our results and subsequent comparison with the admixture approach. Nevertheless, as noted earlier, changing the prior $\pi_{d < \infty}$ is unlikely to affect BF as BF explicitly takes $\pi_{d < \infty}$ into account. We conclude linkage if $\widehat{BF} \geq BF_0 = 25$. According to Raftery (1996), a BF of at least 25 is considered to be a strong evidence for discriminating between two models.

3.1 Preliminary Analysis

We set the burn-in period of the Markov chain to be $B = 10,000$ iterations and after burn-in, the chain is run for $T = 300,000$ iterations. To see if this chain length is adequate and gives satisfactory mixing and convergence, we performed the diagnostics described in Section 2.4. For a given sample, we ran the chain from seven dispersed starting points. After excluding the burn-in period, we calculated \hat{p} and \hat{m} at every 100th iteration using the d values up to

that point. Typical plots are shown in Figure 1. We note that the chains forget about their starting points quickly and show nice convergence. We have also constructed the seven final 95% confidence intervals (CIs) computed using the batch means method for estimating SEs. All seven CIs for both parameters overlap, although they are extremely narrow (this aspect is discussed after the next paragraph). Similar pattern is seen in the plots for several other samples generated with the disease gene location varying over the chromosome. Further, in all the samples, the final estimates, \hat{p} and \hat{m} obtained from all seven starting points, are very close to each other.

We carried out the test of independence of batch means for several samples. For each sample, the number of batches is 100, so the batch length is 3,000. Table 1 shows the two test statistics, Z_y and Z_p for five representative samples. These statistics show evidence of independence of batch means at the 5% level for all samples, as none of the Z_y or Z_p values fall into the rejection region. So, overall the diagnostic tools indicate that $B = 10,000$ and $T = 300,000$ give satisfactory mixing and convergence of the chain.

Since all the \widehat{BF} values exceed the cutoff of 25 for declaring linkage, the disease gene locations are estimated (\hat{m}) and their corresponding interval estimates are also obtained, as shown in Table 1. There are two interval estimates: the 95% CIs obtained using the batch means method and the 95% Bayesian credible set (CS). As can be seen from the table, all the CSs contain their corresponding disease gene locations while none of the 95% CIs does so. Since the same results were obtained for all the other samples that we have considered, it seems that the batch means estimator of SE is a severe underestimate of the true SE. This is a known limitation of batch means estimators, although it is still used routinely to report Monte Carlo errors, perhaps due to the fact that the extent of underestimation is problem-dependent. We further increased the chain length but it did not alleviate the problem. Although the true SE is unknown, we can estimate it from simulations (see next section) and it leads us to conclude that batch means estimators of SEs would not be useful in our case. So, in what follows, we do not consider batch means estimators any further.

3.2 Evaluation of Proposed Approach and Comparison with Admixture Approach

3.2.1 Bayesian versus Admixture

We carried out extensive simulations by placing the disease gene at various positions (5, 13, 25.64, 38, 49.77; in cM) on the chromosome and for each position we generated 500 samples under the intermediate (Int) disease model. For each sample, we analyzed it using our Bayesian approach, and the usual admixture approach with a single α parameter. For the Bayesian approach, we estimated the posterior probability of linkage (\hat{p}), the corresponding Bayes Factor (\widehat{BF}), the location of disease gene under linkage (\hat{m}), and 95% CS for the disease gene location. Summary statistics across all 500 samples are reported in the first segment of Table 2 (before the first set of double horizontal lines). In particular, we report the power (the percentage of the samples showing linkage signals, i.e., $\widehat{BF} > 25$), and the mean and SD of the \hat{m} 's over the samples with linkage signals. Furthermore, we also present the mean 95% CS among those samples with $\widehat{BF} > 25$ and the percentage of these CSs that contain the true value of the disease gene location. For the admixture approach, we report the power (the percentage of samples in which HLOD is greater than 3), the mean and SD of the estimated disease gene locations (positions at which maximum HLOD occurs) over the samples with HLOD > 3 . We use HLOD > 3 as a linkage signal for the admixture approach because this cutoff is customarily used in linkage analysis (Ott 1999), although it has been now accepted that when one accounts for heterogeneity via admixture model, the cutoff should be higher (Abreu, Hodge, and Greenberg 2002). In traditional linkage analysis, a confidence interval (or support interval as it is sometimes referred to) may be obtained by taking all the points whose LOD scores are within 1 unit of the maximum LOD score when it exceeds 3 (Ott 1999). However, properties of such an interval is unclear in heterogeneity analysis and hence we do not report them. We also compared the performances of the two methods under the dominant (Dom), recessive (Rec), and homogeneity (Hom) models at the disease gene position of 25.64 cM. The results are shown in the second segment of Table 2 (between the two sets of double horizontal lines).

From the table, we see that the proposed Bayesian approach is more powerful than the

admixture approach. The relative increases in power range from 18% to 120%. Note that if we had used the correct (higher) cutoff for HLOD, the power of the admixture approach would have gone down further. Moreover, there are no samples for which the admixture approach can pick up a linkage signal while the Bayesian cannot. The mean and SD of the estimated disease gene location are similar for both approaches. In general, it seems that the Bayesian approach has a reasonably good power given the fact that there is so much missing data. Missing data may also be a factor in the resulting 95% CSs being wide, although the probabilities that these CSs contain the true value d is very high. Note that a negative lower limit of a CS means that the CS contains positions beyond (below) the first marker whose position is set to 0 cM, a usual practice in linkage analysis. Compared to the others, the position $d = 49.77$ gives the worst results by both methods. This is most likely due to the fact that the two flanking markers are farthest apart (19.46 cM) among all pairs of consecutive markers.

3.2.2 Fine Mapping

We further evaluated the Bayesian approach in a fine mapping setting, which is the natural next step after preliminary linkage is established. Equally importantly, this further investigation should shed light on whether the seemingly wide CSs are partially due to the coarse map, in addition to the missing data factor as discussed above. We focused on mapping the gene at 25.64 cM, flanked by the two central markers. We saturated the interval (12, 39.5) (slightly overcovering the mean 95% CS reported in Table 2) with 12 new markers with 2.5 cM separation. Each new marker has four equally likely alleles, i.e., heterozygosity of 0.75, which is approximately the average heterozygosity of the markers in the original map. In addition to these new markers, the two original markers in this region are retained, leading to a total of 14 markers for this study. Other than using this finer map of markers, the generation of the data and their analysis were carried out exactly as before under the intermediate model. The results are shown in the last segment of Table 2. Compared to the results using the original map (third row of the first segment of the same table), we observe that, for the Bayesian approach, both the SD and the mean 95% CS are cut by approximately half. Further, the location estimate is much closer to the truth, and the power has also increased.

However, for the admixture approach, although the power also rises sharply, the reduction in SD is less impressive, resulting in its SD been considerably higher than that of the Bayesian approach.

3.2.3 False Positive Rate

To gauge the false positive rates, we simulated 500 samples with no disease gene present on the chromosome. The marker map is the same as the one used in all the other simulations (except for fine mapping). Both Bayesian and admixture approaches gave 1 false positive. So, the two approaches have comparable false positive rates, at least for the setting considered.

3.3 Sensitivity Study

We have varied the values of the proposal parameters, λ_1 , λ_2 , and s (in an $L \rightarrow L$ move, the updated value of d is chosen from the s neighboring distances of the current value of d ; see Appendix A) and found that they do not change the results qualitatively. This seems to indicate that the proposed MCMC scheme is quite robust to the chosen values of these proposal parameters. Also, the algorithm is insensitive to the chosen proposal distribution for d in the $U \rightarrow L$ move (denoted by P_i in Appendix A), since either setting all the P_i 's to be equal or generating the P_i 's from a discretized, truncated normal distribution does not make a difference in the parameter estimates. In conclusion, the results of these sensitivity analyses regarding the proposal distributions reaffirm the comment of Richardson and Green (1997) that it is rarely worth fine tuning the proposal distribution in Metropolis-Hastings methods.

We also studied the sensitivity of the results to the chosen prior distributions. First, we compare the results obtained by using the Beta(3,2) and Beta(1,1) distributions as priors for all α 's. Note that the Beta(1,1) is just the $U(0,1)$ distribution and we used it as priors for the α 's in all of our earlier simulations. Also, recall that the Beta(3,2) is the distribution from which the α values (probability that a family is of linked type) were actually generated in all simulations. Both distributions gave comparable results, indicating that the simulated data are informative enough to overwhelm the effect of prior distributions. We

also examined the sensitivity of the prior for d . Although we use a non-informative prior by setting $\pi_{d<\infty} = 1/22$, another possible non-informative prior is to let $\pi_{d<\infty} = (\text{length of the chromosome})/(\text{total genome length excluding sex chromosomes})$. For chromosome 21, this is approximately 0.017. Apart from these two non-informative priors, we also considered an informative prior of $\pi_{d<\infty} = 0.1$ (relevant for applications in which some a priori knowledge is available). The results from the three priors in terms of BF were the same, due to the fact that BF is independent of $\pi_{d<\infty}$ (Richardson and Green 1997), reaffirming its value as a model selection tool.

4 EFFECT OF DISEASE MODEL MISSPECIFICATION

In real applications of any parametric approach in linkage analysis (as in our real data analyses to be described in the next section), since the true disease model is unknown, usually an approximate model is used. So a natural question is what is the effect of model misspecification. We investigated this issue for the proposed Bayesian approach by analyzing the simulated data of GAW13, consisting of 100 replicates that have been simulated to mimic the real data from the Framingham Heart Study (Almasy et al. 2003). The model used for simulation is extremely complex with about 50 trait genes that interact via complicated relationships to produce various traits. We considered the simulated trait “high blood pressure (HBP)”. In the simulating model, HBP is affected by several genes over several chromosomes. We particularly focused on chromosome 21 as it has three genes, b37, s12, and s10, located at (sex-averaged distance) 25.31, 25.64, and 49.77 cM, respectively, that directly affect blood pressure. The phenotypic data for each person were taken over a range of time. So to label whether a person is affected by HBP, we combined the longitudinal data on blood pressure and another closely related variable, hypertensive treatment using the criterion of Biswas et al. (2003). The marker data are as described in the section Simulation Results. Note that b37 and s12 lie between markers 3 and 4 while s10 lies between markers 5 and 6.

We analyzed each of the 100 replicates using the following two models, where D is the

disease allele. Model I is the intermediate model used in the simulation study; it is the kind of incomplete penetrance model that one might use as an approximation to the true but unknown complex model. Model II has $P(D) = 0.3$, $P(\text{Aff} | DD) = 0.8$, $P(\text{Aff} | Dd) = 0.4$, and $P(\text{Aff} | dd) = 0$. It corresponds to the average of the disease gene s10's simulating models for the systolic and diastolic blood pressures. Figure 2 shows a scatter plot of pairs of \widehat{BF} values (one from each of the two models) in the log scale, and a side-by-side plot of the 95% CSs and associated \hat{m} values from the the two models. The results are similar under the two models. Out of the 100 replicates, there is evidence for linkage in 81 of the replicates under model I and in 82 under model II. There are three replicates for which model I gives signals but model II does not, and vice versa is true in four other replicates. A closer look at the top plot reveals that the $\log(\widehat{BF})$ values for model II tend to be slightly larger than their counterparts for model I, reflected in having more points cluttered below the diagonal line. Also, there is a little less variability associated with model II as represented by the shorter CSs compared to model I. Since model II specifies zero phenocopy rate and is closer to the true models of s10, it seems that it performs slightly better in terms of showing stronger linkage signals and yielding \hat{m} values and CSs closer to the true location of s10. However, more CSs under model I capture the other two disease genes than those under model II. This observation is worth further discussion. In the majority of the replicates, the values of \hat{m} lie between markers 5 (at 40.07 cM) and 6 (at 59.53 cM), i.e., the gene s10 is detected. However, the actual values of \hat{m} in many of the replicates are off from the true location of s10; rather, they are pulled towards the center of the chromosome. This seems to be the effect of the other two genes, which are detected in a few of the replicates, especially under model I. These observations are, in fact, consistent with the simulating model in which the effects of b37 and s12 are much lower (low penetrances) on blood pressure than s10. The fact that model II is closer to the true model for gene s10 makes it less likely to be able to detect the other two genes. Finally, returning back to our primary question, we note that results from both models are very similar, leading us to conclude that this approach is reasonably robust to model misspecification.

5 APPLICATIONS TO TWO DATASETS

5.1 Lung Cancer

The GELCC dataset consists of 52 extended pedigrees. Bailey-Wilson et al. (2004) analyzed these data using an autosomal dominant model with a disease allele frequency of 0.01, and penetrances of 10% and 1% for disease allele carriers and non-carriers, respectively. Using the full dataset, they found a maximum HLOD score of 2.79 on chromosome 6, which is a bit short of the traditional cutoff of 3 to conclude linkage.

We considered the same 52-pedigree dataset and applied the Bayesian approach to the chromosome 6 data. There are 18 markers on chromosome 6 with an average heterozygosity of 0.72 and an average inter-marker distance of 10.47 cM. The LOD scores are obtained from SIMWALK2 (Sobel and Lange 1996) using the same disease model as that used by Bailey-Wilson et al. (2004). We used the same non-informative priors for d and α 's as used in the simulation study. This yielded $\widehat{BF} = 31.94$, exceeding the threshold of 25 for declaring linkage by a substantial margin. The point estimate, \hat{m} , and the 95% CS for the location of the disease gene are 157.56 cM (near marker D62436), and (146, 169) cM, respectively. The CS covers three markers, namely, C6S1848, D6S2436, and D6S1035, and overlaps a genomic region on the q arm of chromosome 6 that exhibits allelic loss in non-small-cell lung carcinoma (Bailey-Wilson et al. 2004). As to be elaborated in the Discussion section, the apparent failure of the admixture approach was due to the variable rates of heterogeneity in the dataset, which reduces its power to detect linkage. In contrast, by accounting for this variability across different families, the Bayesian approach is able to extract the information contained in the dataset to detect the linkage signal.

5.2 Asthma

Of the three asthma samples, two are from the Collaborative Study on the Genetics of Asthma (CSGA), namely Caucasians (S1) and African Americans (S2), and the third is a German sample (S3). S1, S2, and S3 consist of 112, 113, and 97 families, respectively. Eerdewegh et al. (2001) computed several heterogeneity statistics using these datasets.

HLOD did not give any linkage signal. HLOD-C (Vieland et al. 2001), an extension of HLOD, wherein each of the three samples are assigned a separate α parameter, showed some signal on chromosome 6. However, it is not clear whether this can be considered as a significant evidence since HLOD-C involves more parameters than HLOD, and hence needs a higher cutoff than HLOD (see Biswas (2003) for details). A major reason for diminished linkage signal is that the sample S2 has virtually no linked families, which washes out moderate signals contained in S1 and S3 (Eerdewegh et al. 2001; Wang et al. 2001; Biswas and Lin 2004).

Based on the above findings, we decided to analyze chromosome 6 data using the disease model used in Eerdewegh et al. (2001): dominant model with a disease allele frequency of 10.56%. The two CSGA samples and the German sample consist of 21 and 19 markers on chromosome 6, respectively, with only a few markers in common. The markers in these two types of samples have almost the same average heterozygosities, 0.77 and 0.79, and are separated by 9.7 cM on the average. The LOD scores were calculated using GENEHUNTER (Kruglyak et al. 1996). The prior distributions for the α parameters in the S1, S2, and S3 samples were set to be Beta(1,1), Beta(1,8), and Beta(1,1), respectively. These priors incorporates our a priori information (knowledge provided by the data collectors and through their and others' prior analyses) that S2 consists of mostly unlinked families. The disease gene location parameter, d , was assigned the same non-informative prior as in the simulation study. We obtained $\widehat{BF} = 102.5$. Less informative priors for the α 's of the S2 families were also entertained, including Beta(1,5), Beta(2,10), and Beta(1,4); all yielded strong evidence for linkage, with an estimated BF of at least 31.5. The estimated gene location and 95% CS are rather invariant to the priors used, and they are around 46 cM and (38, 52) cM, respectively. This CS contains the estimates reported by Eerdewegh et al. (2001) and Wang et al. (2001) as well as the point and interval estimates in Biswas and Lin (2004), suggesting that we may have narrowed down the disease gene to a correct interval.

6 DISCUSSION

The problem of heterogeneity in human genetics is very complicated but important for mapping genes responsible for complex traits. We address this problem through a Bayesian approach by taking advantage of the general MCMC methodology and one of its more recent developments on tackling model selection issues, the reversible jump algorithm. We have shown that the proposed Bayesian approach can be much more powerful than the currently used admixture approach with a single α parameter. This holds true also for data generated under complicated disease models not considered here, such as the GAW14 simulated data, which are generated under an epistasis model involving four disease genes and two modifying loci (Biswas, Lin, and Berry 2005). Notably, this power gain does not lead to an increase in false positive rates. In addition, interval estimates of parameters are readily available with known properties under the stated assumptions. Today, there is a much greater need for reliable interval estimates in linkage analysis as the focus in genetic studies has been shifted from testing whether a chromosome harbors a disease gene to narrowing the gene down to a small chromosomal region for fine mapping and association studies. We found that this approach is sensitive neither to proposal parameters involved in the MCMC moves nor to disease model misspecification.

In regard to the confidence interval estimates, we found that the ones obtained from the ordinary batch means approach are too narrow to capture the true value of d . This may serve as a cautionary note on relying on batch means' SE to estimate the true SE in other applications. A potential alternative method might be to use the fixed-width approach (Jones, Haran, Caffo, and Neath 2006) which can guard against under estimation of SE, although it is not clear whether the method can be easily adapted to the setting here. We found that the Bayesian credible sets provide reasonable interval estimates. With missing data and coarse marker map, they may be a little wider than what one would like. However, our simulations under the fine mapping setting show that the CSs given by the Bayesian approach localizes the disease gene in reasonably narrow intervals. With complete data, further narrowing of the CSs are observed. Another alternative interval estimate is the highest posterior density (HPD) set. An HPD set may be narrower than the usual credible

sets but it may consist of disjoint intervals. For the point estimate of the location of the disease gene, we have also explored posterior median and mode, besides mean, and we found the three statistics to give similar results in our simulations.

For the 52-pedigree lung cancer dataset, the ability of the Bayesian method to detect the linkage signal while the admixture approach failed to do so, reaffirms our conclusion of increased power of the Bayesian approach drawn from our extensive simulation study. Since our simulation results indicate that the increase in power for the Bayesian approach does not come at the expense of inflated false positive rate, we are quite confident that the identified linkage region harbors a true disease locus. Our conclusion is further strengthened by a subset analysis performed in Bailey-Wilson et al. (2004). By focusing on a subset of 23 pedigrees that were believed to have minimal heterogeneity, a significant signal was detected. Their point estimate is very close to ours and is contained in our 95% CS. Two aspects of our investigation are noteworthy. First, in addition to the point estimate, the Bayesian approach also provides an interval estimate for the gene location with known statistical properties, a feat unmatched by the subset analysis. Second, there are known shortcomings of subset analysis, the most notable being power loss unless the stratification factor(s) adequately represents heterogeneity (Leal and Ott 2000).

In the GAW12 asthma application, we saw how a priori information can be used to form informative priors, which in turn can increase the power to identify linkage. In fact, if we use non-informative $U(0,1)$ priors for all the α 's in the three samples, no linkage signal is obtained. So we recommend gaining some information about the α parameters and incorporating it via informative priors. An important difference between the sample S2 and the other two samples, S1 and S3, is ethnicity. Such factors, which may discriminate between various heterogeneity levels, can be used to form groups and then HLOD-C can be applied to get some ideas about the α 's. Another way of grouping is according to distributions of families as discussed in Lin and Biswas (2004), but this grouping scheme may not be realistic in most applications other than experimental crosses. Alternatively, one may simply examine the homogeneity LOD score curves for each family to get a sense of the probabilities that they are of linked type.

Another level of flexibility can be added to the Bayesian approach when covariate infor-

mation that can help discriminate between families is available. For example, if race (age of disease onset) is a discriminating factor in determining the linked type of a family, as in the GAW12 dataset (breast cancer), then hierarchical priors for the α 's can be reasonably introduced. More specifically, we can let the α parameters follow different distributions with hyper priors determined by their covariate values.

In this paper, we have presented a Bayesian approach in the context of localizing one disease gene at a time, i.e., marginal analysis. We note that the reversible jump MCMC mechanism employed here readily allows for treating the number of disease genes as an unknown parameter, and simultaneous mapping of the disease genes (joint analysis). This type of analysis will be especially useful in mapping more than one gene on the same chromosome. The basic framework of this general case has been outlined elsewhere (Biswas 2003). Nonetheless, since marginal analysis is still pre-dominant in real data analysis, and joint analysis would certainly be more intricate, it is more logical to first investigate the method for marginal analysis. In practice, such a single gene analysis may still shed light on the existence of multiple disease loci. A strong linkage signal coupled with a wide CS and a disjoint HPD set may be indicative of the presence of more than one gene on the chromosome. Such a predicament was observed in our analysis of the GAW13 data, which contain two distinguishable genes (by linkage analysis) on chromosome 21. For joint analysis, which is the subject of a future research project, hierarchical modeling as discussed above will be much more important as the number of α parameters will go up dramatically.

Appendix A

Four move types: $\mathbf{L} \rightarrow \mathbf{L}$, $\mathbf{L} \rightarrow \mathbf{U}$, $\mathbf{U} \rightarrow \mathbf{U}$, and $\mathbf{U} \rightarrow \mathbf{L}$

In an $\mathbf{L} \rightarrow \mathbf{L}$ move, we update each of the parameters in $\boldsymbol{\alpha}$ using Gibbs sampler by sampling from the conditional distribution of $\pi(\alpha_j | \boldsymbol{\alpha}_{(-j)}, d, \mathbf{x})$, $j = 1, \dots, k$, where $\boldsymbol{\alpha}_{(-j)}$ is the remaining set of elements of $\boldsymbol{\alpha}$ excluding α_j . To do this, we note that

$$\pi(\alpha_j | \boldsymbol{\alpha}_{(-j)}, d, \mathbf{x}) \propto \pi_j(\alpha_j) \left[\alpha_j \frac{L_j(d|x_j)}{L_j(\infty|x_j)} + (1 - \alpha_j) \right], \quad 0 \leq \alpha_j \leq 1.$$

For the special case of uniform priors for all α_j 's, i.e., $\pi_j(\alpha_j) = 1$, $0 \leq \alpha_j \leq 1$, for all j , we can find the normalizing constant for the above distribution and hence we use the simple inverse cdf method to sample from $\pi(\alpha_j | \boldsymbol{\alpha}_{(-j)}, d, \boldsymbol{x})$. This case may arise in many applications where there is no prior information about the α_j 's and so this is a natural non-informative prior to use. For the general case of any other prior distribution $\pi_j(\alpha_j)$, $\pi(\alpha_j | \boldsymbol{\alpha}_{(-j)}, d, \boldsymbol{x})$ may not be available in closed form and so inverse cdf method cannot be used. In this case, we use rejection sampling. A detailed description of both cases can be found in Biswas (2003).

After updating $\boldsymbol{\alpha}$, in an L \rightarrow L move, we update d using a Metropolis-Hastings algorithm. Note that

$$\pi(d | \boldsymbol{\alpha}, \boldsymbol{x}) \propto \pi_{d < \infty} \pi_d(I_d) \prod_{j=1}^k \left[\alpha_j \frac{L_j(d | x_j)}{L_j(\infty | x_j)} + (1 - \alpha_j) \right] = g(d), \quad d < \infty.$$

Let $d_{(t)}$ be the value of d at iteration t and s be a pre-specified integer. At $(t+1)$ th iteration, we select I_{d^*} from the set of integers in the range $(\max\{1, I_{d_{(t)}} - s\}, \min\{I_{d_{(t)}} + s, N\})$, with equal probability of selecting each integer in this set. So, the chosen d^* can be $d_{(t)}$ itself or any of the s neighboring distance values of $d_{(t)}$, unless $d_{(t)}$ is near the beginning or the end of the chromosome. Let us denote this probability mass function by $f(d | d_{(t)})$. We accept d^* as the next value of d , i.e., $d_{(t+1)} = d^*$, with probability $\min \left\{ 1, \frac{g(d^*)f(d_{(t)} | d^*)}{g(d_{(t)})f(d^* | d_{(t)})} \right\}$. If d^* is not accepted, $d_{(t+1)} = d_{(t)}$.

An U \rightarrow U move does not involve updating of any parameters, so we simply set $d_{(t+1)} = \infty$.

When an U \rightarrow L move is proposed, we need to generate $(\boldsymbol{\alpha}, d)$ since the U state has no parameters. Each α_j is chosen randomly from a distribution, $p_j(\alpha_j)$, and I_d is selected from the set $\{1, \dots, N\}$ with respective probabilities $\{P_1, \dots, P_N\}$. We accept this move with probability $\min\{1, A(\boldsymbol{\alpha}, d)\}$, where

$$A(\boldsymbol{\alpha}, d) = \frac{\pi_{d < \infty} \pi_d(I_d) \prod_{j=1}^k \pi_j(\alpha_j)}{\pi_{d = \infty}} \times \frac{\prod_{j=1}^k \left[\alpha_j \frac{L_j(d | x_j)}{L_j(\infty | x_j)} + (1 - \alpha_j) \right]}{1} \\ \times \frac{1 - \lambda_1}{(1 - \lambda_2) \cdot P_{I_d} \cdot \prod_{j=1}^k p_j(\alpha_j)} \times |1|.$$

Note that $A(\boldsymbol{\alpha}, d)$ is the product of prior ratio, likelihood ratio, proposal ratio, and Jacobian of the transformation. If this move is accepted, the generated $\boldsymbol{\alpha}$ and d are taken as $\boldsymbol{\alpha}_{(t+1)}$ and $d_{(t+1)}$, respectively, otherwise the chain stays in the U state, i.e., $d_{(t+1)} = \infty$.

The move $L \rightarrow U$ is opposite of the move $U \rightarrow L$. So, it is accepted with the probability $\min\{1, A^{-1}(\boldsymbol{\alpha}_{(t)}, d_{(t)})\}$. If it gets accepted, $(\boldsymbol{\alpha}_{(t)}, d_{(t)})$ are discarded and $d_{(t+1)} = \infty$. Otherwise the chain stays in the L state with $(\boldsymbol{\alpha}_{(t+1)}, d_{(t+1)}) = (\boldsymbol{\alpha}_{(t)}, d_{(t)})$.

We use $\lambda_1 = \lambda_2 = 0.5$ and $s = 5$. In the $U \rightarrow L$ move, $p_j(\alpha_j)$ is taken to be $U(0, 1)$ for all j , and $P_i, i = 1, \dots, N$, are calculated using a discretized and truncated normal distribution.

Appendix B

Asymptotic Joint Distribution of Estimators (\hat{p}, \hat{m})

Divide the total realizations for d (excluding the burn-in period) into R batches, each of length L . Let $d_r^{(l)}$ denote the l th value of the distance in the r th batch, $l = 1, 2, \dots, L, r = 1, 2, \dots, R$. Define two functions, $f(d) = dI(d < \infty)$ and $g(d) = I(d < \infty)$. Also, let $y_r = \sum_{l=1}^L f(d_r^{(l)})/L$ and $p_r = \sum_{l=1}^L g(d_r^{(l)})/L, r = 1, 2, \dots, R$. Then the following results are apparent by applying a standard Markov chain Central Limit Theorem (Chung 1967). For large $L, \left\{ \sqrt{L}(y_r - m^*), r = 1, 2, \dots, R \right\}$ and $\left\{ \sqrt{L}(p_r - p), r = 1, 2, \dots, R \right\}$ are approximately iid samples from $N(0, \sigma_f^2)$ and $N(0, \sigma_g^2)$, respectively, where $m^* = Ef(d) = E(y_r)$ and $p = Eg(d) = E(p_r)$.

Now, consider the estimators \hat{p} and \hat{m} . We can rewrite them in terms of the f and g functions as

$$\hat{m} = \frac{\sum_{r=1}^R \sum_{l=1}^L f(d_r^{(l)})}{\sum_{r=1}^R \sum_{l=1}^L g(d_r^{(l)})} = \frac{\sum_{r=1}^R y_r/R}{\sum_{r=1}^R p_r/R} = \frac{\bar{y}}{\bar{p}}, \quad \text{for } \bar{p} \neq 0,$$

$$\hat{p} = \frac{1}{R} \sum_{r=1}^R \frac{1}{L} \sum_{l=1}^L g(d_r^{(l)}) = \frac{1}{R} \sum_{r=1}^R p_r = \bar{p}.$$

By invoking the bivariate Central Limit Theorem and the bivariate Delta Method (Lehmann 1999),

$$\left(\sqrt{R}(\hat{p} - p), \sqrt{R}(\hat{m} - m) \right) \sim BVN(0, 0, \tau_{11}, \tau_{22}, \tau_{12}), \quad \text{approximately,}$$

where

$$\begin{aligned}
\tau_{11} &= \left[\left(\frac{\partial h}{\partial \bar{y}} \right)^2 \text{Var}(y_r) + 2 \left(\frac{\partial h}{\partial \bar{y}} \right) \left(\frac{\partial h}{\partial \bar{p}} \right) \text{Cov}(y_r, p_r) + \left(\frac{\partial h}{\partial \bar{p}} \right)^2 \text{Var}(p_r) \right]_{\bar{p}=p, \bar{y}=m^*} \\
&= \text{Var}(p_r), \\
\tau_{22} &= \left[\left(\frac{\partial k}{\partial \bar{y}} \right)^2 \text{Var}(y_r) + 2 \left(\frac{\partial k}{\partial \bar{y}} \right) \left(\frac{\partial k}{\partial \bar{p}} \right) \text{Cov}(y_r, p_r) + \left(\frac{\partial k}{\partial \bar{p}} \right)^2 \text{Var}(p_r) \right]_{\bar{p}=p, \bar{y}=m^*} \\
&= \left(\frac{m^*}{p} \right)^2 \left\{ \frac{\text{Var}(y_r)}{m^{*2}} - \frac{2\text{Cov}(y_r, p_r)}{m^*p} + \frac{\text{Var}(p_r)}{p^2} \right\}, \\
\tau_{12} &= \left[\frac{\partial h}{\partial \bar{y}} \frac{\partial k}{\partial \bar{y}} \text{Var}(y_r) + \left(\frac{\partial h}{\partial \bar{y}} \frac{\partial k}{\partial \bar{p}} + \frac{\partial h}{\partial \bar{p}} \frac{\partial k}{\partial \bar{y}} \right) \text{Cov}(y_r, p_r) + \frac{\partial h}{\partial \bar{p}} \frac{\partial k}{\partial \bar{p}} \text{Var}(p_r) \right]_{\bar{p}=p, \bar{y}=m^*} \\
&= \frac{1}{p} \text{Cov}(y_r, p_r) - \frac{m^*}{p^2} \text{Var}(p_r).
\end{aligned}$$

Hence, \hat{p} and \hat{m} are consistent estimators of p and m , respectively (when both $L \rightarrow \infty$ and $R \rightarrow \infty$). We note that these results are not the standard ones. Although each batch consists of a fixed number (L) of realizations, the number of realizations in each batch that correspond to the linked model ($d < \infty$) are random and \hat{m} is based on the realizations from the linked model only. Finally, we also note that the simultaneous $(1-\alpha)\%$ confidence intervals for (m, p) can also be obtained, if so desired.

References

- Abreu, P. C., Hodge, S. E., and Greenberg, D. A. (2002), ‘‘Quantification of Type I Error Probabilities for Heterogeneity LOD Scores,’’ *Genetic Epidemiology*, 22, 156-169.
- Almasy, L., Amos, C. I., Bailey-Wilson, J. E., Cantor, R. M., Jaquish, C. E., Martinez, M. and Neuman, R. J., Olson, J. M., Palmer, L. J., Rich, S. S., Spence, M. A., and MacCluer, J. W. (2003), ‘‘Genetic Analysis Workshop 13: Analysis of Longitudinal Family Data for Complex Diseases and Related Risk Factors,’’ *BMC Genetics*, 4, S1.
- Bailey-Wilson, J. E., Amos, C. I., Pinney, S. M., Petersen, G. M., de Andrade, M., Wiest, J. S., Fain, P., Schwartz, A. G., You, M., Franklin, W., Klein, C., Gazdar, A., Rothschild, H., Mandal, D., Coons, T., Slusser, J., Lee, J., Gaba, C., Kupert, E., Perez, E., Zhou,

- X., Zeng, D., Liu, Q., Zhang, Q., Seminara, D., Minna, J., and Anderson, M. W. (2004), "A Major Lung Cancer Susceptibility Locus Maps to Chromosome 6q23-25," *American Journal of Human Genetics*, 75, 460-474.
- Biswas, S. (2003), "On Incorporating Heterogeneity in Linkage Analysis," PhD dissertation, The Ohio State University, Department of Statistics. Available at <http://www.ohiolink.edu/etd/view.cgi?osu1070468056>.
- Biswas, S., and Lin, S. (2004), "Evaluations of Maximization Procedures for Estimating Linkage Parameters under Heterogeneity," *Genetic Epidemiology*, 26, 206-217.
- Biswas, S., Lin, S., and Berry, D. A. (2005), "A new Bayesian Approach Incorporating Covariate Information for Heterogeneity and its Comparison with HLOD," *BMC Genetics*, 6, S138.
- Biswas, S., Papachristou, C., Irwin, M. E., and Lin, S. (2003), "Linkage Analysis of the Simulated Data - Evaluations and Comparison of Methods," *BMC Genetics*, 4, S70.
- Chung, K. L. (1967), *Markov Chains with Stationary Transition Probabilities* (2nd ed.), New York: Springer-Verlag.
- Eerdewegh, P. V., Dowd, M., Dupuis, J., Falls, K., Hayward, B., and Santangelo, S. L. (2001), "On the Detection of Linkage in Multiple Data Sets: A Comparison of Various Statistical Approaches," *Genetic Epidemiology*, 21, S67-S72.
- Geman, S., and Geman, D. (1984), "Stochastic Relaxation, Gibbs Distribution, and the Bayesian Restoration of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721-741.
- Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (eds.) (1996), *Markov Chain Monte Carlo in Practice*, London, U.K.: Chapman & Hall
- Green, P. J. (1995), "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination," *Biometrika*, 82, 711-732.

- Hastings, W. K. (1970), "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," *Biometrika*, 57, 97-109.
- Janssen, B., Halley, D., and Sandkuijl, L. A. (1997), "Linkage Analysis Under Locus Heterogeneity: Behavior of the A-test in Complex Analyses," *Human Heredity*, 47, 223-233.
- Jones, G. L., Haran, M., Caffo, B. S., and Neath, R. (2006), "Fixed-width output analysis for Markov chain Monte Carlo," To appear in *Journal of the American Statistical Association*.
- Kleijnen, J. P. C. (1987), *Statistical Tools for Simulation Practitioners*, New York: Marcel Dekker.
- Kruglyak, L., Daly, M. J., Reeve-Daly, M. P., and Lander, E. S. (1996), "Parametric and Nonparametric Linkage Analysis: A Unified Multipoint Approach," *American Journal of Human Genetics*, 58, 1347-1363.
- Leal, S.M., and Ott, J. (2000), "Effects of Stratification in the Analysis of Affected Sib-Pair Data: Benefits and Costs," *American Journal of Human Genetics*, 66, 567-575.
- Lehmann, E. L. (1999), *Elements of Large Sample Theory*, New York: Springer.
- Lin, S., and Biswas, S. (2004), "On Modeling Locus Heterogeneity Using Mixture Distributions," *BMC Genetics*, 5, 29.
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., Teller, E. (1953), "Equation of State Calculations by Fast Computing Machines," *Journal of Chemical Physics*, 21, 1087-1092.
- Ott, J. (1999), *Analysis of Human Genetic Linkage*, Baltimore, MD: The John Hopkins University Press.
- Raftery, A. E. (1996), "Hypothesis Testing and Model Selection," in *Markov Chain Monte Carlo in Practice*, eds. W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, London, U.K.: Chapman & Hall, pp 163-187.
- Richardson, S., and Green, P. J. (1997), "On Bayesian Analysis of Mixtures With an Unknown Number of Components," *Journal of the Royal Statistical Society B*, 59, 731-792.

- Sobel, E., and Lange, K. (1996) "Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker sharing statistics," *American Journal of Human Genetics*, 58, 1323-1337.
- Schmeiser, B. (1982) , "Batch Size Effects in the Analysis of Simulation Output," *Operations Research*, 30, 556-568.
- Vieland, V. J., and Logue, M. (2002), "HLODs, Trait Models, and Ascertainment: Implications of Admixture for Parameter Estimation and Linkage Detection," *Human Heredity*, 53, 23-35.
- Vieland, V. J., Wang, K., and Huang, J. (2001), "Power to Detect Linkage Based on Multiple Sets of Data in the Presence of Locus Heterogeneity: Comparative Evaluation of Model-based Linkage Methods for Affected Sib Pair Data," *Human Heredity*, 51, 199-208.
- Wang, K., Huang, J., Logue, M., and Vieland, V. (2001), "Combined Multipoint Analysis of Multiple Asthma Data Sets Based on the Posterior Probability of Linkage," *Genetic Epidemiology*, 21, S73-S78.
- Whittemore, A. S., and Halpern, J. (2001), "Problems in the Definition, Interpretation, and Evaluation of Genetic Heterogeneity," *American Journal of Human Genetics*, 68, 457-465.
- Wijsman, E. M., Almasy, L., Amos, C. I., Borecki, I., Falk, C. T., King, T. M., Martinez, M. M., Meyers, D., Neuman, R., Olson, J. M., Rich, S., Spence, M. A., Thomas, D. C., Vieland, V. J., Witte, J. S., and MacCluer, J. W. (2001), "Genetic Analysis Workshop 12: Analysis of Complex Genetic Traits: Applications to Asthma and Simulated Data," *Genetic Epidemiology*, 21, S1-S853.

Table 1: Test statistics Z_y and Z_p for independence of batch means y_r 's and p_r 's, respectively. Hypothesis of independence is rejected at 5% level if $|Z| > 1.96$, where 1.96 is the 2.5% quantile of the $N(0, 1)$ distribution.

d^a	Z_y	Z_p	\hat{p}	\widehat{BF}	\hat{m}	SE(\hat{m}) ^b	95% CI	95% CS
5.00	0.585	1.257	0.946	367.9	6.46	0.091	(6.28,6.64)	(-8,23)
13.00	-0.758	-0.788	0.989	1888.1	10.32	0.042	(10.24,10.40)	(1, 18)
25.64	-0.316	0.198	0.876	148.4	16.38	0.046	(16.29,16.47)	(6,27)
38.00	0.644	1.441	0.589	30.1	39.02	0.077	(38.87,39.17)	(12,52)
49.77	-0.886	-1.668	0.724	55.1	44.84	0.055	(44.73,44.95)	(33,61)

a: True disease gene location (in cM).

b: Batch means SE based on 30 batches.

Table 2: Bayesian and admixture approach results at various positions of the disease gene, d (in cM), and various models. Columns 5-8 are about the estimated disease gene location under either approach and are computed over samples showing linkage signals.

Model	d	Method	Power ^a	Mean	SD	Mean 95% CS ^b	% CS containing d
Int	5.00	Bayesian	67.0	5.90	5.49	(-6.78, 20.12)	98.81
		Admix.	44.4	4.87	5.28		
Int	13.00	Bayesian	72.0	12.30	5.32	(-0.68, 26.06)	98.89
		Admix.	50.4	11.68	4.69		
Int	25.64	Bayesian	73.6	25.05	5.87	(12.83, 38.74)	97.28
		Admix.	46.8	24.28	5.55		
Int	38.00	Bayesian	75.8	38.22	5.32	(25.13, 51.51)	100.00
		Admix.	54.8	38.06	5.97		
Int	49.77	Bayesian	58.4	47.16	8.13	(30.60, 62.87)	93.84
		Admix.	36.2	47.97	7.79		
Dom	25.64	Bayesian	80.4	24.76	5.27	(14.01, 36.71)	97.26
		Admix.	60.2	24.32	5.41		
Rec	25.64	Bayesian	41.0	26.43	6.10	(11.80, 42.62)	98.54
		Admix.	18.6	25.82	6.18		
Hom	25.64	Bayesian	84.6	25.64	4.51	(13.85, 38.94)	99.29
		Admix.	71.8	25.43	4.60		
Int	25.64 ^c	Bayesian	85.4	25.67	2.99	(18.53, 32.67)	96.25
		Admix.	65.2	25.66	4.08		

a: percentage of samples showing linkage signals.

b: CS with negative lower limit contains positions beyond the first marker located at 0 cM.

c: a finer map of markers is used as described in the text.

Figure 1: Diagnostic plots for a sample generated with the disease gene located at 25.64 cM. Seven chains are run, each from a different starting point. Top plot shows \hat{p} and bottom plot shows \hat{m} .

Figure 2: Analysis of 100 GAW13 simulated replicates using both disease models I and II. Top: scatter plot of pairs of $\log(\widehat{BF})$ values for the two models. The dashed lines represent $\log(25)$. The discordant pairs (replicates for which the two models lead to different conclusions) are marked with a “+” symbol. Bottom: plot of 95% CSs (lines) and \hat{m} values (circles). The solid lines and circles are for Model I while the dashed lines and open circles are for Model II. These are plotted for a replicate only if $\widehat{BF} > 25$ for both models. The three horizontal gray lines represent the locations of the three disease genes.

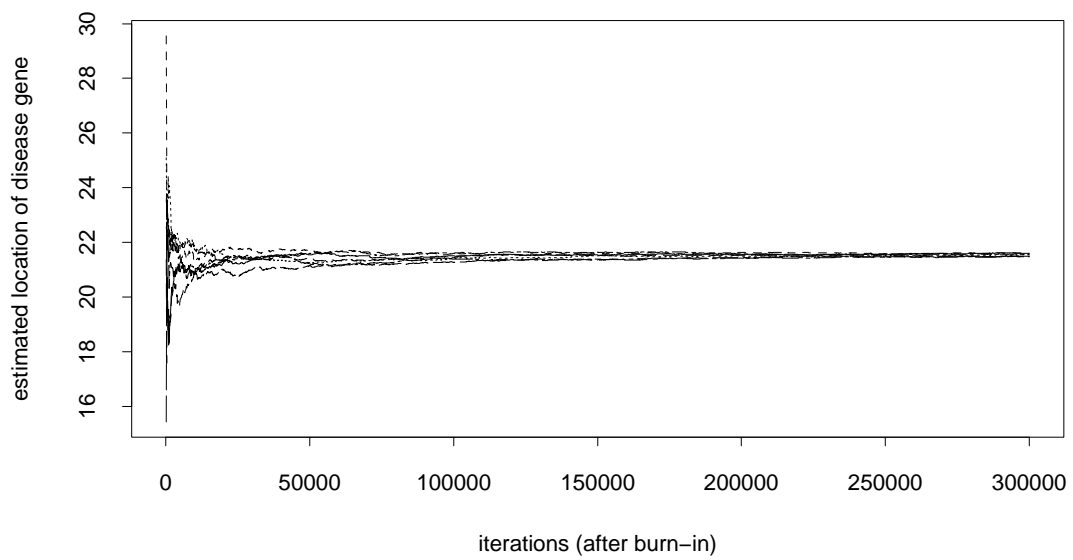
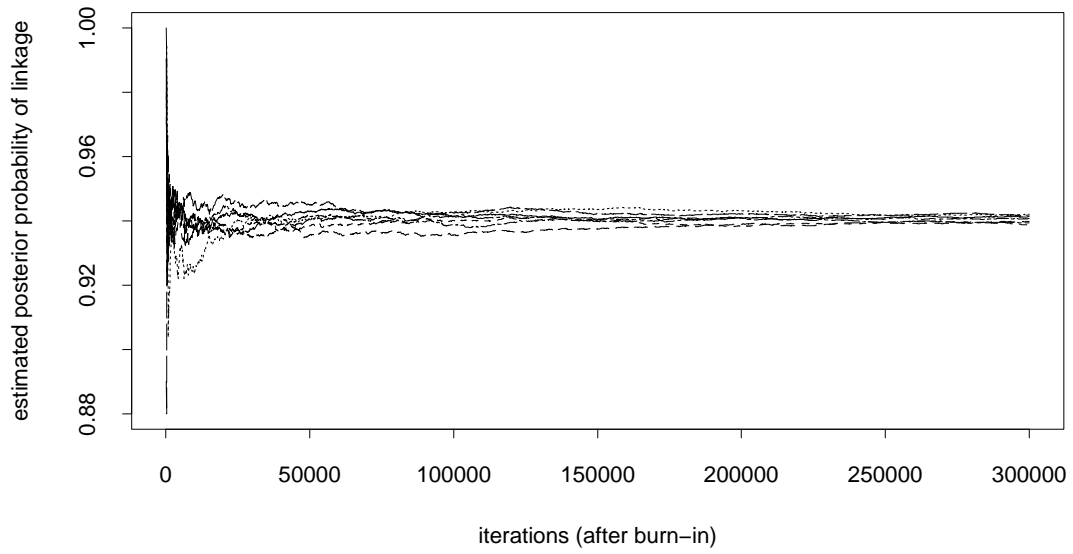


Figure 1:

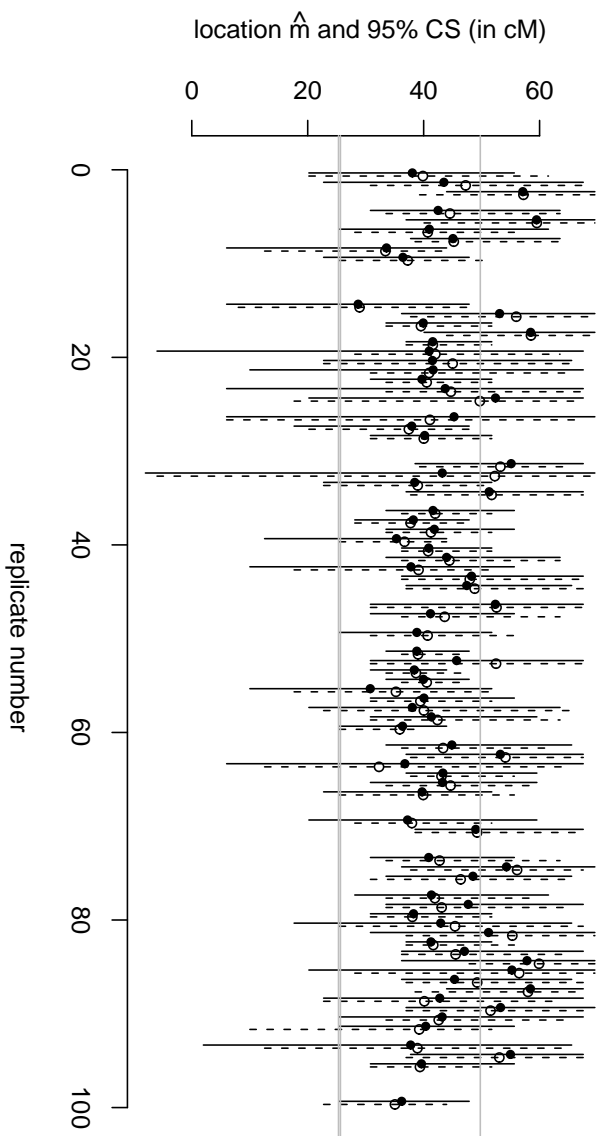
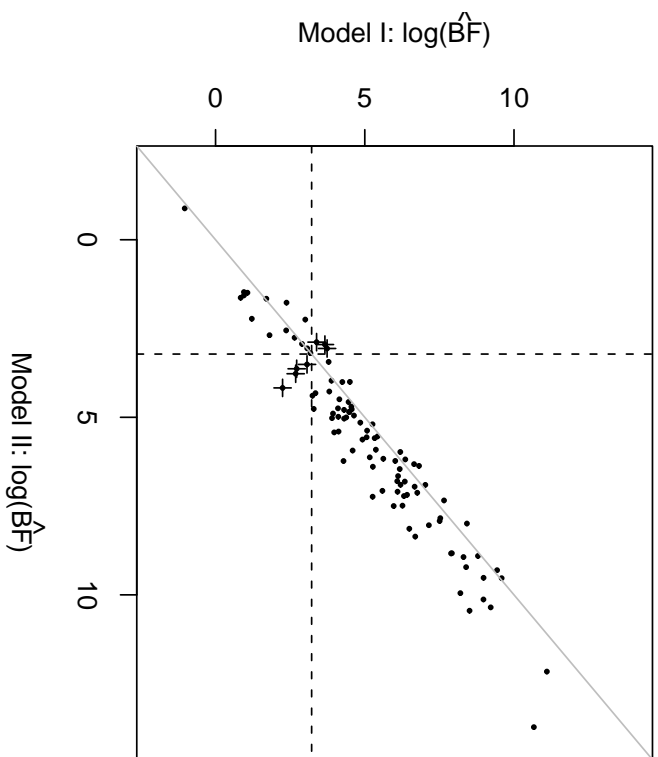


Figure 2: