Inference From Genome-Wide Association Studies Using a Novel Markov Model

Fay J. Hosking,[1]* Jonathan A. C. Sterne,[2] George Davey Smith,[2] and Peter J. Green[1]

[1]*Department of Mathematics, University of Bristol, Bristol, UK*
[2]*Department of Social Medicine, University of Bristol, Bristol, UK*

In this paper we propose a Bayesian modeling approach to the analysis of genome-wide association studies based on single nucleotide polymorphism (SNP) data. Our latent seed model combines various aspects of *k*-means clustering, hidden Markov models (HMMs) and logistic regression into a fully Bayesian model. It is fitted using the Markov chain Monte Carlo stochastic simulation method, with Metropolis-Hastings update steps. The approach is flexible, both in allowing different types of genetic models, and because it can be easily extended while remaining computationally feasible due to the use of fast algorithms for HMMs. It allows for inference primarily on the location of the causal locus and also on other parameters of interest. The latent seed model is used here to analyze three data sets, using both synthetic and real disease phenotypes with real SNP data, and shows promising results. Our method is able to correctly identify the causal locus in examples where single SNP analysis is both successful and unsuccessful at identifying the causal SNP. *Genet. Epidemiol.* 2008.

Key words: Markov chain Monte Carlo; hidden Markov model; logistic regression; SNP-base population association study

*Correspondence to: Fay J. Hosking, Department of Mathematics, University of Bristol, University Walk, Bristol BS8 4PB, UK.
E-mail: fay.hosking@bristol.ac.uk
Received 14 June 2007; Accepted 7 February 2008
Published online in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/gepi.20322

INTRODUCTION

Genome-wide association studies (GWAS) are now technologically and financially feasible. Such studies involve genotyping several thousand cases and controls across several thousand single nucleotide polymorphisms (SNPs), in order to identify loci with a causal effect on the risk, severity, age of onset or prognosis of human diseases. These studies, such as The Wellcome Trust Case Control Consortium [2007], will provide large amounts of data and the exciting possibility of identifying causal genes and variants.

GWAS pose difficult statistical challenges. The large number of associations tested can lead to spurious findings, unless steps are taken to control for multiple comparisons [Wacholder et al., 2004; Dudbridge and Koeleman, 2004]. Accounting for linkage disequilibrium (LD) between SNPs may provide means of reducing this problem and hence increasing power to locate causal loci [The International HapMap Consortium, 2005]. In this paper we propose a flexible Bayesian approach to the analysis of GWAS, which allows for inference to be drawn about the causal locus.

The paper is set out as follows: The second section introduces the idea of "seeds" as generators of the observed data. The notation and model are introduced in the third section; together with details of inference, convergence checking and computational aspects are also given. Results from three applications are given in the fifth section. The paper concludes with a discussion of possible extensions and future work.

MODEL BACKGROUND

Motivation for our approach came from examining the local patterns of densely genotyped SNP data. Figure 1 shows a section of phased data from HapMap [The International HapMap Consortium, 2003] for 50 SNPs starting at index 1,751 of Chromosome One. The data are from 60 unrelated members of the CEU population in this study, and therefore consists of 120 strands of genetic data. The left-hand plot shows the original data, with SNPs on vertical axis and the strands on the horizontal axis. In the right-hand plot the strands have been reordered, by using the first component obtained from principal components analysis, so that similar strands tend to be plotted consecutively.

This reordering reveals that a total of 73 of the chromosome sections are identical copies of one of three patterns. The remaining SNP patterns can be generated from these patterns by switching a few times between them and by changing a few isolated

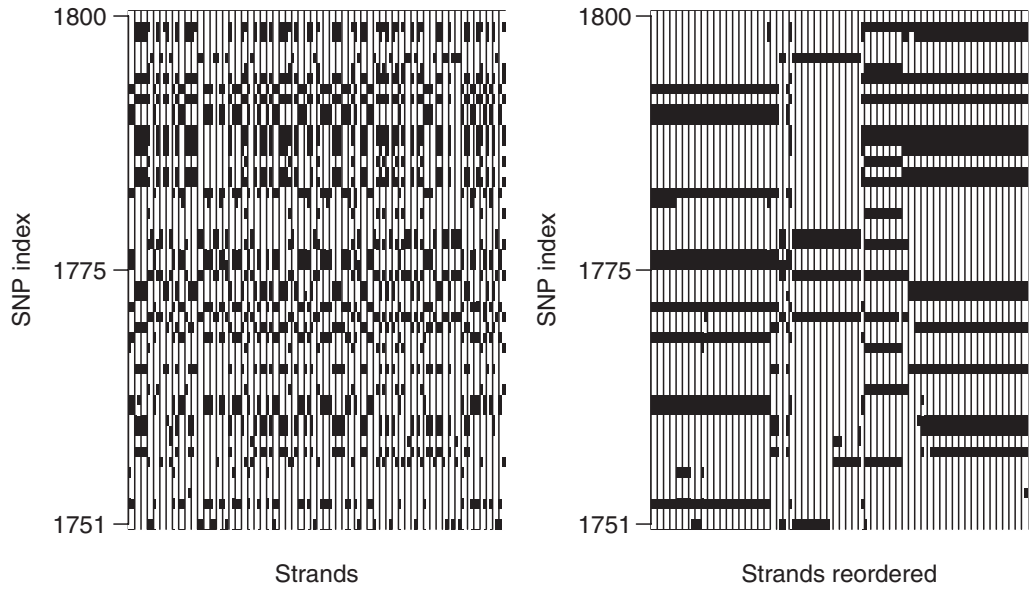

Fig. 1. The original genotype data (black as rare allele and white as common) for 50 SNPs starting at index 1,751 of Chromosome One of CEU population (a) in the original order and (b) with the strands reordered. SNP, single nucleotide polymorphism.

values. Therefore, in the formal statistical model described in Model, we will assume observed SNP patterns in a particular short section of the chromosome to be generated stochastically from combining and ''mutating'' the ''seeds,'' which are considered as resulting from ancestral haplotypes. Similar ideas have been used in various works [Scheet and Stephens, 2006; Rastas et al., 2005; Kimmel and Shamir, 2005] and are discussed in the sixth section.

METHODS

NOTATION

We assume that the data consist of binary-phased bi-allelic SNP data, X, for n people and on m SNPs on a single chromosome. The jth strand for the ith individual will be denoted as $\boldsymbol{x}_{i,j}$, $j = 1,2$. Each person also has a binary phenotype variable y_i. We consider possible causal locations, u, lying between each of the adjacent SNPs. All of our modeling is conditional on u and therefore this is considered fixed for Model. We assume that there is a single causal SNP, at u, which is not measured and that there are unobserved alleles $\boldsymbol{z}$ at u. The tth element of the jth strand for the ith individual with $z_{i,j}$ imputed at location u is given by $\widetilde{x}_{i,j}^{(t)}$:

$$\widetilde{x}_{i,j}^{(t)} = \begin{cases} x_{i,j}^{(t)} & \text{if } t < u, \\ z_{i,j} & \text{if } t = u, \\ x_{i,j}^{(t+1)} & \text{if } t > u. \end{cases}$$

In this model we have used a discrete parameter scale to label the SNPs, thereby using their indices as distance measure. An alternative approach would have been to use a continuous parameter model, with SNP locations measured in basepairs; however, our modeling is primarily driven by LD, which does not correlate more strongly with either distance measure. A continuous parameter model would also have been more computationally burdensome. Strictly speaking, this model is not consistent as u varies since the discrete parameter scale is perturbed by the insertion of the SNP as u. However, numerical experiments show that this is negligible (see Inference).

MODEL

We build a statistical model that reflects the observed data and unknown parameters, including the location of the causal locus. The model consists of two interconnecting parts. The first part, the ''strand model'' accounts for the dependence between the SNPs along the chromosome, with unknown parameters $\boldsymbol{\alpha}$. The second part the ''disease model'' models the relationship between the causal alleles, $\boldsymbol{z}$, and the phenotype, $\boldsymbol{y}$, using unknown parameters $\boldsymbol{\beta}$. These two sections together give us a framework upon which to build a Bayesian model.

Strand model. Each strand (observed SNPs; augmented by each possible SNP at u) is modeled as being generated by a discrete parameter Markov chain switching between $\mathcal{S}$ seeds (see the second section). For each u, $\mathcal{S}$ seeds of length m and centered at u are chosen by using a heuristic k-means clustering approach, details of which are given in Window length and choice of seeds.

The value of the SNP at u for each seed is calculated by examining its two neighbors. If they

are both the same then that value is imputed otherwise the next two adjacent SNPs on either side are considered. This process is repeated until the two SNPs being considered are the same when the most often occurring allele is imputed. The $\mathcal{S}$ augmented patterns of length $m+1$ formed in this way are used as seeds.

A possible series of switches to construct a particular observed strand, $\widetilde{x}_{i,j}$ from the seeds is represented by the underlying Markov chain $h_{i,j}^{(t)}$, $t = 1,\ldots,m+1$ and $h_{i,j}^{(t)} \in \mathcal{S} = \{0,\ldots,s-1\}$. The Markov chain is assumed to have a stationary transition matrix Γ with the (r,s)th element given by $\gamma(r,s)$ and initial distribution π_0, which is often assumed to be the stationary distribution of Γ. We assume that Γ has diagonal entries $1-\alpha_0$ and off-diagonal entries of $\alpha_0/(\mathcal{S}-1)$. Thus, $1-\alpha_0$ is the probability of the chain remaining in the current state and α_0 is a parameter of Γ and π_0.

Without noise the SNP value at locus t would necessarily be the same as that of the current seed and we need more flexibility; therefore, we introduce a ''mutation'' parameter α_1, defined by

$$\widetilde{x}_{i,j}^{(t)} = h_{i,j}^{(t)} \quad \text{with probability } 1-\alpha_1.$$

This means that the SNP value for a particular strand at locus t depends both on the current state of the hidden Markov chain, h_t, and on α_1.

The likelihood for a particular strand consisting of $\boldsymbol{x}_{i,j}$, with $z_{i,j}$ imputed at u is

$$\begin{aligned} p(\widetilde{\boldsymbol{x}}_{i,j}|\alpha_0,\alpha_1) = &\sum_{h\in\mathcal{S}^{m+1}} \pi_0(h_{i,j}^{(1)},\alpha_0)p(\widetilde{x}_{i,j}^{(1)}|h_{i,j}^{(1)},\alpha_1) \\ &\times \prod_{t=2}^{m+1} \gamma(h_{i,j}^{(t-1)},h_{i,j}^{(t)})p(\widetilde{x}_{i,j}^{(t)}|h_{i,j}^{(t)},\alpha_1). \end{aligned} \tag{1}$$

The sum is over $\mathcal{S}^{m+1}$ elements, so direct computation quickly becomes infeasible as the number of SNPs increases, even for a small number of seeds. However, a recursive procedure allows for the calculation of (1) in $\mathcal{O}(\mathcal{S}^2(m+1))$ time for each α_0, α_1, i, j. Denoting $\boldsymbol{x}^{1:t}$ as $\{x^{(1)},x^{(2)},\ldots,x^{(t)}\}$ for any vector $\boldsymbol{x}$, the forward variable [Scott, 2002] is $\ell_t(r) \equiv p(\widetilde{\boldsymbol{x}}_{i,j}^{1:t}, h_t = r|\alpha_1)$. That is, $\ell_t(r)$ is the joint likelihood contribution of $\widetilde{\boldsymbol{x}}_{i,j}^{1:t}$ as well as the event $h_t = r$ averaging over the previous $h_1,\ldots,h_{t-1}$.

The recursive procedure calculates

$$\ell_t(r) = p(\widetilde{x}_{i,j}^{(t)}|r,\alpha_1)\sum_{w=0}^{\mathcal{S}-1}\gamma(w,r)\ell_{t-1}(w)$$

and thus (1) is obtained by $\sum_{r=0}^{\mathcal{S}-1}\ell_{m+1}(r)$.

Disease model. The disease model likelihood involves three binomial probabilities p_0, p_1 and p_1, which are the probabilities of being a case for genotypes 00, {01, 10} and 11, respectively:

$$p(y_i|z_{i,1},z_{i,2},\boldsymbol{\beta}) = p_{z_{i,1}+z_{i,2}}^{y_i}(1-p_{z_{i,1}+z_{i,2}})^{1-y_i}.$$

We follow Minelli et al. [2005] and parameterize p_0, p_1 and p_2:

$$\begin{aligned} \text{logit}(p_0) &= \beta_0 - \beta_1/2, \\ \text{logit}(p_1) &= \beta_0 + \beta_1\beta_2 - \beta_1/2, \\ \text{logit}(p_2) &= \beta_0 + \beta_1/2. \end{aligned}$$

It is possible to work with the binomial probabilities, p_0, p_1, p_2 directly and this allows for computational saving when using conjugate priors. However, it is more convenient to place constraints on the possible genetic models using $\boldsymbol{\beta}$.

Overall model. We place this model in a Bayesian framework, with prior distributions $p(u)$, $p(\boldsymbol{\alpha})$ and $p(\boldsymbol{\beta})$. The full joint probability is then given by

$$\begin{aligned} &p(u,\boldsymbol{\alpha},\boldsymbol{\beta},X,\boldsymbol{y},\boldsymbol{z}) \\ &= p(u)p(\boldsymbol{\alpha})p(\boldsymbol{\beta})\prod_{i=1}^{n}\left\{\prod_{j=1}^{2}p(\boldsymbol{x}_{i,j},z_{i,j}|\boldsymbol{\alpha},u)\right\} \\ &\quad\times\prod_{i=1}^{n}p(y_i|\boldsymbol{z}_i,\boldsymbol{\beta}), \end{aligned}$$

where $\boldsymbol{z}_i = \{z_{i,1},z_{i,2}\}$.

The three parameters for the strand model, α_0, α_1 and β_2 are constrained to lie between zero and one. This constraint is placed on β_2 so that we consider only common genetic models. These include recessive, dominant and co-dominant, which are characterized by β_2 values of 0, 1 and 0.5, respectively. This choice of constraint allows for the rare allele to be protective ($\beta_1 < 0$) but since $\beta_2 \in [0,1]$ it does not allow for rare models of inheritance such as overdominance. For the three constrained parameters two possible noninformative priors are $\beta(1,1)$ and $\beta(0.5,0.5)$. The former is uniform over the interval while the latter corresponds to the Jeffreys prior for a binomial distribution and both have been used for modeling vague prior beliefs about proportions. For the other two parameters, β_0 and β_1, dispersed normal distributions centered at zero may be used as non-informative priors.

INFERENCE

We wish to infer the values of the unknowns u, $\boldsymbol{z}$, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ from the observed data with particular emphasis on u and perhaps $\boldsymbol{\beta}$. Our Bayesian approach means that this should be based on the posterior distribution, the conditional distribution of the unknowns given the data, X and $\boldsymbol{y}$:

$$\begin{aligned} &p(u,\boldsymbol{z},\boldsymbol{\alpha},\boldsymbol{\beta}|X,\boldsymbol{y}) \\ &\quad= p(u|X,\boldsymbol{y})p(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{z}|u,X,\boldsymbol{y}) \\ &\quad= p(u|X,\boldsymbol{y})p(\boldsymbol{z},\boldsymbol{\alpha}|X,u)p(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{z}). \end{aligned} \tag{2}$$

The second factorization in the above equation uses the fact that $\boldsymbol{\beta}$ and $\boldsymbol{y}$ given $\boldsymbol{z}$ are conditionally

independent of $X,u,\boldsymbol{\alpha}$. Calculation of (2) is intractable except by simulation. We therefore calculate $p(\boldsymbol{\alpha}, \boldsymbol{\beta}, z|u, X, \boldsymbol{y})$ by using Markov chain Monte Carlo (MCMC) techniques. One method for sampling from complex probability distributions is the Metropolis-Hastings algorithm [Metropolis et al., 1953; Hastings, 1970]. We use a deterministic-sweep Metropolis-Hastings method with random walk proposals for the five real parameters and sum out over z. Computational savings are made by using the calculations defined in strand model for the strand model.

Since $p(z, \boldsymbol{\alpha}|X, u)$ requires calculating and storing the seeds and since finding a suitable method for moving about the space of u is complex, u will be treated as a model indicator. Independent MCMC runs will be performed, to both approximate for each u, the marginal likelihood for each u and the posterior distribution of the parameters and causal SNP values, given u. These together determine the full posterior distributions of all the unknowns. These independent runs for each u mean that incompatible models are used for different u but this approximation is not serious.

Inference about the location of the causal locus is of primary interest although inference on the other parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, is possible. The posterior distribution of the logistic disease model parameters, $\boldsymbol{\beta}$, with u held fixed, for example at the posterior mode of u may give insight into the strength and type of association. Inference on $\boldsymbol{\alpha}$ averaged over u may give insight into the strength of LD in a region.

Inference about the location of the causal SNP is achieved by, $p(u|X, \boldsymbol{y}) \propto p(u)p(\boldsymbol{y}|X, u)p(X|u)$. The last factor, $p(X|u)$ is not constant due to the insertion of an unknown z for each u position, which has the effect of changing the distances between the first and second halves of the data in the current window of interest. However, numerical experiments have shown that the variability of this factor with respect to u is negligible.

IMPLEMENTATION OF MCMC

Although the strand model is formally defined along the whole chromosome, attention is limited to a moving window centered at the current u position. Limiting attention to a shorter moving window improves computational speed and is reasonable since LD is believed to be strong only over relatively short distances. We have fixed the number of seeds $\mathcal{S}$ and length of window and have run independent MCMC samplers for each u moving along the strands.

Inference about u requires the calculation of the marginal likelihood $p(\boldsymbol{y}|u, X)$. We have chosen to use two direct methods for this calculation both of which can be implemented by computationally quick algorithms. The first is motivated by importance sampling and is the harmonic mean of the posterior conditional likelihood,

$$\widehat{p}_1(\boldsymbol{y}|u, X) = \left[\frac{1}{N}\sum_{t=1}^{N} p(\boldsymbol{y}|\boldsymbol{\alpha}^{(t)}, \boldsymbol{\beta}^{(t)}, z^{(t)}, u, X)^{-1}\right]^{-1},$$

where $\{\boldsymbol{\alpha}^{(t)}, \boldsymbol{\beta}^{(t)}, z^{(t)}\}_{t=1}^{N}$ is a sample from the posterior distribution $p(\boldsymbol{\alpha}, \boldsymbol{\beta}, z|X, \boldsymbol{y}, u)$.

The second method is importance weighted marginal density estimation [IWMDE; Chen, 1994] and leads to the estimator,

$$\widehat{p}_2(\boldsymbol{y}|u, X) = \left[\frac{1}{N}\sum_{t=1}^{N} \frac{w(\boldsymbol{\alpha}^{(t)}, \boldsymbol{\beta}^{(t)}, z^{(t)}|u, X)}{p(\boldsymbol{y}|\boldsymbol{\alpha}^{(t)}, \boldsymbol{\beta}^{(t)}, z^{(t)}, u, X)p(\boldsymbol{\alpha}^{(t)}, \boldsymbol{\beta}^{(t)}, z^{(t)}|u, X)}\right]^{-1},$$

where $\{\boldsymbol{\alpha}^{(t)}, \boldsymbol{\beta}^{(t)}, z^{(t)}\}_{t=1}^{N}$ is as before and $w(\cdot)$ is a completely known conditional density. Taking $w(\cdot)$ as $p(\boldsymbol{\alpha}, \boldsymbol{\beta}, z|u, X)$ yields the harmonic mean estimator given before. These estimators are both derived from different exact expressions for $p(\boldsymbol{y}|u, X)$, with integrals replaced by averages across the simulation.

CONVERGENCE CHECKING

As with all MCMC-based approaches the convergence of the runs needs to be checked. However, due to the impractical nature of graphically assessing the convergence of each parameter chain at each u location, the usual graphical methods cannot be used. Instead a statistic is required that can alert us to possible convergence problems. In order to minimize the number of such statistics we would like to focus on the terms forming the empirical averages $\widehat{p}_1(\boldsymbol{y}|u, X)$ and $\widehat{p}_2(\boldsymbol{y}|u, X)$.

This statistic [Gelman and Rubin, 1992; Brooks and Gelman, 1998] requires K independent runs of length $2T$ and can be easily and quickly calculated. Denoting the terms in the summations of $\widehat{p}_1(\boldsymbol{y}|u, X)$ or $\widehat{p}_2(\boldsymbol{y}|u, X)$ for chain k at the tth iteration as ψ_{kt}, then the variance ratio R can be estimated by

$$\widehat{R} = \frac{\widehat{V}}{W},$$

where

$$\widehat{V} = \frac{T-1}{T}W + \frac{K+1}{K}\frac{1}{K-1}\sum_{k=1}^{K}(\bar{\psi}_{k.} - \bar{\psi}_{..})^2,$$

$$W = \frac{1}{K(T-1)}\sum_{k=1}^{K}\sum_{t=T+1}^{2T}(\psi_{kt} - \bar{\psi}_{k.})^2.$$

For overdispersed starting values $\widehat{R}$ overestimates R and should converge to one from above if the chain converges.

It is possible to use a correction factor to account for variability in the variance [Brooks and Gelman, 1998]. This correction factor leads to

$$\widehat{R}_c = \frac{d+3}{d+1}\widehat{R},$$

where d is approximated by $2\widehat{V}/\widehat{\mathrm{Var}}(\widehat{V})$. However, this correction is usually minor as at convergence d tends to be large [Brooks and Gelman, 1998]. Brooks and Gelman [1998] recommended plotting $\widehat{R}$ or $\widehat{R}_c$ across all iterations; in order to minimize output we have calculated it only at several iteration indices spread throughout the run, each time using the second half of all the preceding iterations. This means that examination of the output to flag anomalous values is quick and easy.

We have used a long burn-in of 20,000 iterations followed by a further 30,000 iterations, to ensure that the majority of the chains have no convergence problems. While once two or more independent samplers have been run, $\widehat{R}$ or $\widehat{R}_c$ can be calculated to assist with identifying those chains that show a lack of convergence. If necessary the model has been re-run and all results given are from fully converged chains.

WINDOW LENGTH AND CHOICE OF SEEDS

Choosing the length of window and the number of seeds are both effectively model-selection problems, but they are computationally burdensome. For the applications described in the fifth section our model appears robust to different window lengths and number of seeds.

Several different ways of choosing the seeds have been explored. We are currently using k-means clustering to obtain $\mathcal{S}$ clusters for each moving window. The k-means algorithm is run from several different starting positions to ensure convergence to a global minimum. Then, since k-means gives the centroids of the clusters and these are not guaranteed to be binary, the final cluster centroids are rounded to either zero or one and used as the seeds.

APPLICATIONS

HAPMAP-BASED STUDY

We conducted two simulation studies on two different sections of HapMap Phase 2 data from Chromosome One of the CEU population. The 60 unrelated individuals from this data set were used and two subsections of 101 SNPs were chosen. One nonmonomorphic SNP from each subsection was chosen to be "causal" and binary phenotypes were randomly generated. The causal SNP was then removed from each data set. Four different window widths of 20, 30, 40 and 50 SNPs were used in the analyses and for each three seeds were used. The results obtained are shown graphically in Figures 2 and 3.

These figures show the harmonic mean estimator $\widehat{p}_1$ and the IWMDE estimator $\widehat{p}_2$ for each of the data sets for each u location studied. The true causal SNP location is indicated by vertical dashed lines. The plotted values are approximately proportional to log posterior probabilities; they are approximate due to the omission of the term $P(X|u)$ as described in Inference. These estimates show the strength of association between each loci and the phenotype of interest. The lower panel shows the $\log_{10}$ P-values from Fisher's exact test for a 2×3 table. With both analyses, the graphs indicate the relative degree of support from the data for the hypotheses that the

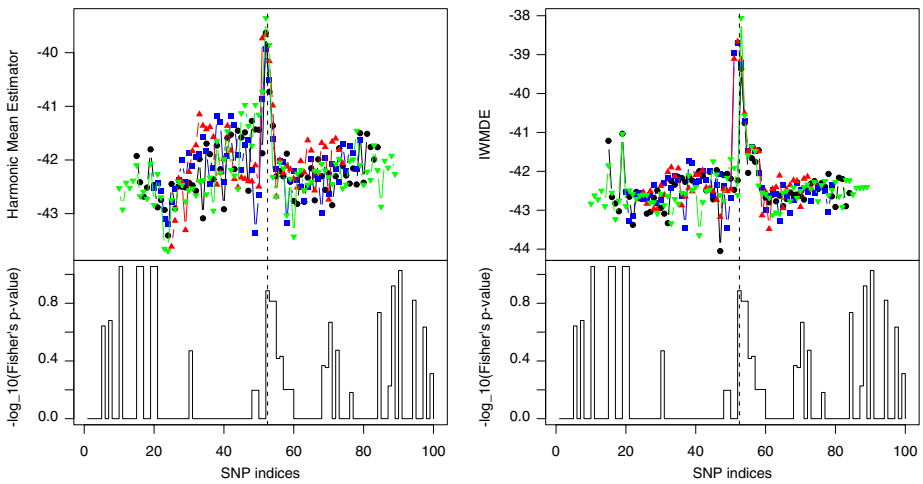

Fig. 2. Marginal log likelihood output using the harmonic mean estimator and IWMDE for data set 1 (50 -▲-, 40 - ■ -, 30 -●-, 20 -▼-). The lower panel shows the $\log_{10}$ values of the P-values from the Fisher's exact test for a 2×3 table. The location of the causal SNP is denoted by the vertical dashed line. SNP, single nucleotide polymorphism; IWMDE, importance weighted marginal density estimation. [Color figure can be viewed in the online issue which is available at www.interscience.wiley.com.]

corresponding loci are causal. However, *P*-values and posterior probabilities are not directly comparable, and the graphs cannot be plotted on the same axes. It might be noted that our Bayesian analysis tends to separate loci with high support from those with moderate support rather more clearly.

In the first data set (Fig. 2) both estimators reach their maximum values near the same location as the true causal locus. This time the maximum is unique, probably due to the weak LD in the region covered by this data set. In the second data set the Fisher *P*-values maximum are not maximized at the true causal locus. A signal is seen at the causal locus although this is not as strong as some at either end of the data set.

In the second data set (Fig. 3) both estimators again reach their maximum values at the same location as the true causal SNP as indicated by a dashed vertical line. Due to the strong LD in this second data set and the small number of individuals other local maxima are also seen, indicating other loci that are in LD with the causal locus. The appearance of these local maxima is exacerbated by the small number of individuals. The Fisher *P*-values also have their maximum at the causal locus and show a strong signal corresponding to most of the peaks in our estimators. In both data sets the results obtained from both estimators are consistent.

Sensitivity analysis, not shown, showed that the burn-in period was sufficient for convergence of R. From the above plots it is clear that the results are similar across the four different window widths used and very similar results were also obtained when increasing the number of seeds used to five.

An important feature of our model-based analysis is that it provides simultaneous joint inference about all unknown quantities. As an example, the parameters of the disease regression model can be estimated. For the second data set the posterior means (and standard deviations) are $\bar{\beta}_0 = 0.3947$ (0.451), $\bar{\beta}_1 = 1.910$ (1.098) and $\bar{\beta}_2 = 0.6154$ (0.216). The joint distribution can be

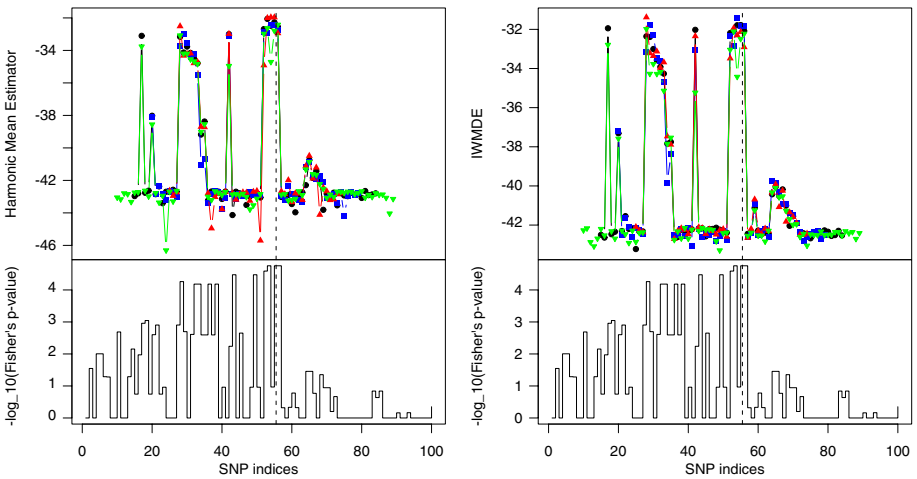

Fig. 3. Marginal log likelihood output using the harmonic mean estimator and IWMDE for data set 2 (50 -▲-, 40 - ■ -, 30 -●-, 20 -▼-). The lower panel shows the $\log_{10}$ values of the *P*-values from the Fisher's exact test for a 2×3 table. The location of the causal SNP is denoted by the vertical dashed line. SNP, single nucleotide polymorphism; IWMDE, importance weighted marginal density estimation. [Color figure can be viewed in the online issue which is available at www.interscience.wiley.com.]

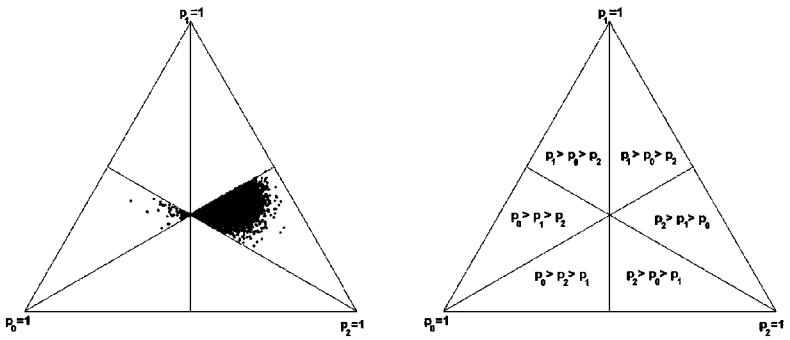

Fig. 4. Plot showing the posterior distributions of the probabilities of being a case given genotypes $\{00\}$, $\{01, 10\}$ and $\{11\}$ as denoted by p_0, p_1 and p_2 for the most likely causal locus as determined by the model. The right-hand plot divides the plotting area into regions annotated by the size order of the three relative probabilities.

most clearly plotted by using p_0, p_1 and p_2 as defined in strand model to be the probabilities of being a case given genotypes {00}, {01, 10} and {11}, respectively. In Figure 4 the joint distribution of these binomial probabilities is plotted for the most likely causal locus as determined by the model. Figure 4 is best examined together with looking at the marginal distributions of β_1 and β_2, which are shown in Figure 5.

The diagonal lines in Figure 4 correspond to $\beta_2 = 0$ and 1 as imposed by our model. The corresponding histogram in Figure 5 shows that the distribution of β_2 is skewed towards one suggesting a near-dominance disease model. Both the histogram for β_0 and Figure 5 show that there is a high probability (greater than 97%) that the SNP is not protective.

CYP2D6

The CYP2D6 gene is known to play an important role in drug metabolism. Hosking et al. [2002] genotyped 27 SNPs in a 880 kb region flanking the CYP2D6 gene and identified a 403 kb region of high LD spanning the CYP2D6 locus (Fig. 6). This data set has since been used several times to test new methodology [Morris et al., 2003; Waldron et al., 2006]. The data set consists of information from 1,018 individuals, 41 of whom were classified as cases due to their poor drug metabolism.

We used PHASE [Stephens et al., 2001; Stephens and Donnelly, 2003] to phase the data, and the resulting data were treated as if this phasing was fixed. To investigate the effect of uncertainty in the phasing we ran our model on five different output files from PHASE. For this data set, due to the small number of SNPs, one static window of length 27 was used for all the MCMC runs.

This plot has several interesting features. It shows a high peak very close to the CYP2D6 locus as indicated by two vertical dashed lines. The location of this peak slightly to the left of the CYP2D6 locus is consistent with other results [Morris et al., 2003; Waldron et al., 2006]. Association is also found in a

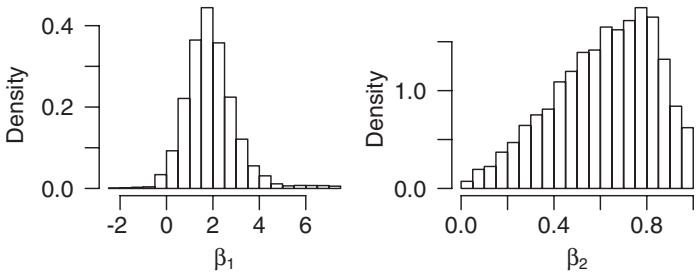

Fig. 5. Plots showing the marginal distributions of β_1 and β_2.

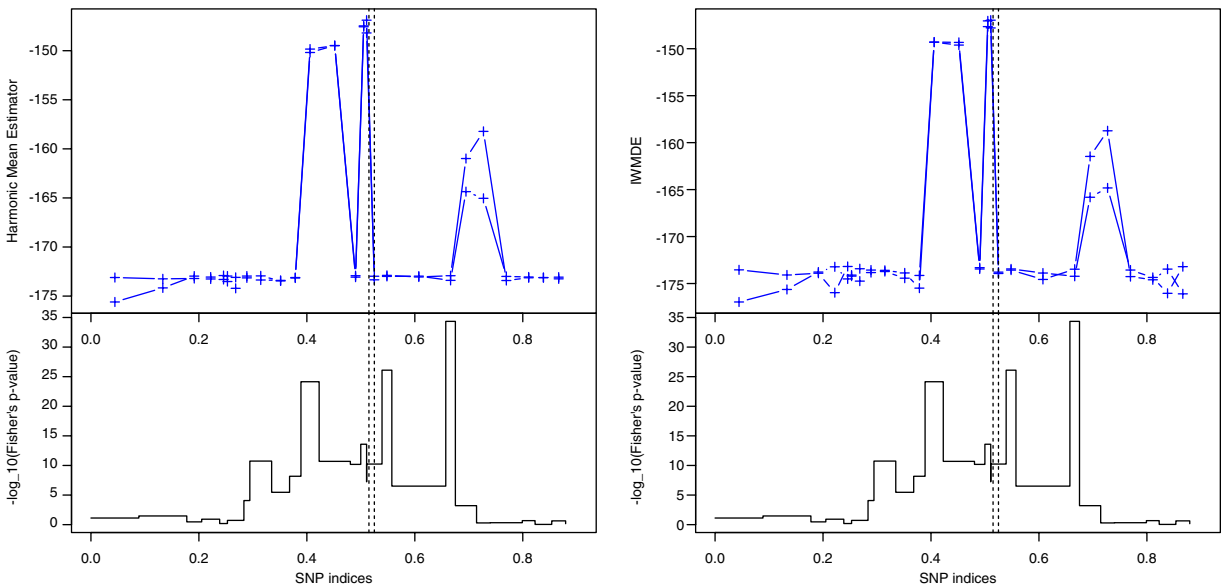

Fig. 6. Marginal log likelihood output using the harmonic mean and IWMDE for the CYP2D6 data set for five independent MCMC samplers. The lower panel shows the $\log_{10}$ values of the *P*-values from Hosking et al. [2002]. The vertical dashed lines show the location of the CYP2D6 gene. IWMDE, importance weighted marginal density estimation; MCMC, Markov chain Monte Carlo. [Color figure can be viewed in the online issue which is available at www.interscience.wiley.com.]

region 0.1 Mb to the left of the CYP2D6 gene but in strong LD with it, this association has also been noted in other work. Three Fisher's *P*-values are particularly strong, one of which is very close to the causal locus; however, this is not the strongest signal seen from the *P*-values.

DISCUSSION

In this paper we have described a general method to search for causal loci in GWAS, allowing for and exploiting LD between nearby SNPs. This latent seed model uses hidden Markov models (HMMs) to describe the observed patterns seen in SNP data and is set in a Bayesian framework allowing simultaneous coherent inference about the position of the causal locus, and other parameters including those predicting disease status. Results from several data sets, both with real and simulated phenotypes are very encouraging.

Although this method will always be computationally intensive due to requiring a large number of MCMC simulations this is justified by the richness of the resulting conclusions. In any case, several factors mitigate the computational load. This is maintained to be linear in the number of loci examined, through use of a finite window centered on each locus. Further, the ever-increasing speed of computers and the fact that the independent samplers are suited to being run on parallel processors, its computational intensity is not seen as being prohibitive.

Recently several methods for analyzing data from GWAS have been published, including work by Morris [2006], Waldron et al. [2006] and Verzilli et al. [2006]. Morris [2006] used a Bayesian partition model for clustering haplotypes and then used a Bayes factor to reflect the strength of evidence that the disease is associated with polymorphisms in the candidate region. Waldron et al. [2006] also defined haplotype clusters and then used these clusters to predict the genotype at a particular locus that can then be tested for association with the phenotype. Verzilli et al. [2006] used a Bayesian graphical model to locate causal SNPs in a very computationally efficient way.

Various other authors have recently published independent work specifically using HMMs in the analysis of different aspects of data from GWAS. Scheet and Stephens [2006] and Rastas et al. [2005] have focused on using HMMs primarily to phase the data and to impute missing genotype values. They used the idea of modeling clustering membership along a chromosome based on an HMM and used an Expectation-Maximization algorithm for maximum likelihood estimation of the causal locus. In these papers, the model-selection problem of choosing the number of clusters/seeds is treated differently: Scheet and Stephens [2006] used a cross-validation approach while Rastas et al. [2005] regarded the number as fixed. Our decision to use a discrete parameter model for the SNP positions is analogous with Rastas et al. [2005]. Scheet and Stephens [2006] do allow for the possibility of a continuous parameter scale; however, they report no improvement in performance over the discrete parameter scale.

Kimmel and Shamir [2005] examined the phasing of the data and the imputation of missing values, as well as predicting phenotypes. For the latter, a particular locus is chosen and this SNP is regarded as the phenotype, instead of an SNP predictor. A cross-validation method is then used to predict each individual's phenotype individually when all other individual's phenotypes are known.

We believe that our model is a promising avenue of approach for the analysis of GWAS and despite its computational intensity it shows good performance. It is also easily elaborated to incorporate missing and unphased data, environmental variables and continuous phenotypes as well as controlling for population substructure that will make it applicable to more general GWAS data.

ACKNOWLEDGMENTS

We thank Louise Hosking from GlaxoSmithKline for providing the CYP2D6 data.

REFERENCES

Brooks SP, Gelman A. 1998. General methods for monitoring convergence of iterative simulations. J Comput Graph Stat 7:434–455.

Chen MH. 1994. Importance-weighted marginal Bayesian posterior density estimation. J Am Stat Assoc 89:818–824.

Dudbridge F, Koeleman BPC. 2004. Efficient computation of significance levels for multiple associations in large studies of correlated data, including genomewide association studies. Am J Hum Genet 75:424–435.

Gelman A, Rubin DB. 1992. Inference from iterative simulation using multiple sequences. Stat Sci 7:457–511.

Hastings WK. 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 53:97–109.

Hosking LK, Boyd PR, Xu CF, Nissum M, Cantone K, Purvis IJ, Khakhar R, Barnes MR, Liberwirth U, Hagen-Mann K, Ehm MG, Riley JH. 2002. Linkage disequilibrium mapping identifies a 390 kb region associated with CYP2D6 poor drug metabolising activity. Pharmacogenomics 2:165–175.

Kimmel G, Shamir R. 2005. A block-free hidden Markov model for genotypes and its application to disease association. J Comput Biol 12:1243–1260.

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. 1953. Equations of state calculations by fast computing machines. J Chem Phys 21:1087–1091.

Minelli C, Tobin JR, Abrams KR. 2005. An integrated approach to the meta-analysis of genetic association studies using Mendelian randomization. Am J Epidemiol 160:445–452.

Morris AP. 2006. A flexible Bayesian framework for modeling haplotype association with disease, allowing for dominance

effects of the underlying causative variants. Am J Hum Genet 79:679–694.

Morris AP, Whittaker JC, Xu CH, Hosking LK, Balding DJ. 2003. Multipoint linkage-disequilibrium mapping narrows location interval and identifies mutation heterogeneity. Proc Natl Acad Sci USA 100:13442–13446.

Rastas P, Koivisto M, Mannila H, Ukkonen E. 2005. A hidden Markov technique for haplotype reconstruction, volume 3692. Lecture Notes in Computer Science. Berlin: Springer. p 140–151.

Scheet P, Stephens M. 2006. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet 78:629–644.

Scott SL. 2002. Bayesian methods for hidden Markov models: recursive computing in the 21st century. J Am Stat Assoc 97:337–351.

Stephens M, Donnelly P. 2003. A comparison of Bayesian methods for haplotype reconstruction. Am J Hum Genet 73:1162–1169.

Stephens M, Smith NJ, Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68:978–989.

The International HapMap Consortium. 2003. The International HapMap Project. Nature 426:789–796.

The International HapMap Consortium. 2005. A haplotype map of the human genome. Nature 437:1299–1320.

The Wellcome Trust Case Control Consortium. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661–684.

Verzilli CJ, Stallard N, Whittaker JC. 2006. Bayesian graphical models for genomewide association studies. Am J Hum Genet 79:100–112.

Wacholder S, Chanock S, Garcia-Closas M, El Ghormli L, Rothman N. 2004. Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. J Natl Cancer Inst 96:434–442.

Waldron ERB, Whittaker JC, Balding DJ. 2006. Fine mapping of disease genes via haplotype clustering. Genet Epidemiol 30:170–179.