Genetic Epidemiology 29: 91–107 (2005)

Direct Analysis of Unphased SNP Genotype Data in Population-Based Association Studies Via Bayesian Partition Modelling of Haplotypes

Andrew P. Morris*

Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom

We describe a novel method for assessing the strength of disease association with single nucleotide polymorphisms (SNPs) in a candidate gene or small candidate region, and for estimating the corresponding haplotype relative risks of disease, using unphased genotype data directly. We begin by estimating the relative frequencies of haplotypes consistent with observed SNP genotypes. Under the Bayesian partition model, we specify cluster centres from this set of consistent SNP haplotypes. The remaining haplotypes are then assigned to the cluster with the ''nearest'' centre, where distance is defined in terms of SNP allele matches. Within a logistic regression modelling framework, each haplotype within a cluster is assigned the same disease risk, reducing the number of parameters required. Uncertainty in phase assignment is addressed by considering all possible haplotype configurations consistent with each unphased genotype, weighted in the logistic regression likelihood by their probabilities, calculated according to the estimated relative haplotype frequencies. We develop a Markov chain Monte Carlo algorithm to sample over the space of haplotype clusters and corresponding disease risks, allowing for covariates that might include environmental risk factors or polygenic effects. Application of the algorithm to SNP genotype data in an 890-kb region flanking the CYP2D6 gene illustrates that we can identify clusters of haplotypes with similar risk of poor drug metaboliser (PDM) phenotype, and can distinguish PDM cases carrying different high-risk variants. Further, the results of a detailed simulation study suggest that we can identify positive evidence of association for moderate relative disease risks with a sample of 1,000 cases and 1,000 controls. *Genet. Epidemiol.* 29:91–107, 2005. © 2005 Wiley-Liss, Inc.

Key words: population-based association study; unphased SNP genotype data; haplotype disease-risk estimation; Bayesian partition model

Contract grant sponsor: Wellcome Trust.
*Correspondence to: Dr. Andrew Morris, Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK.
E-mail: amorris@well.ox.ac.uk
Received 6 January 2005; Accepted 16 March 2005
Published online 6 June 2005 in Wiley InterScience (www.interscience.wiley.com)
DOI: 10.1002/gepi.20080

INTRODUCTION

The most promising approach for mapping genes contributing to complex traits is generally accepted to be disease-marker association studies of samples of unrelated affected cases and unaffected controls, provided that the high-risk variant is not too rare [Risch and Merikangas, 1996; Zondervan and Cardon, 2004]. The power of this approach depends, in part, on the extent of linkage disequilibrium (LD) of the high-risk variant with alleles at flanking markers within a population of individuals, generated as a result of the shared ancestry of their chromosomes at the disease gene. Chromosomes carrying the same variant tend to share a more recent common ancestor at the disease gene than random chromosomes in the population, and hence are expected to carry similar marker haplotypes in the flanking region. The extent of haplotype sharing will depend on the time to the most recent common ancestor and the rates of recombination and marker mutation. As a result, we then expect to see correlation between disease phenotype and marker haplotypes, the strength of the correlation dependent on the penetrances of variants in the disease gene.

In this report, we focus on association studies using single nucleotide polymorphism (SNP) markers in candidate genes or small candidate regions. One of the attractive features of these markers for mapping is their abundance through-

out the genome, although each individual polymorphism provides relatively little information about LD. Single-locus analyses, testing for association of each SNP in turn with the disease, are thus extremely inefficient, even before addressing the issue of multiple testing with many markers. For high-density panels of SNPs in candidate genes or small candidate regions, strong correlations are expected between alleles at different loci on the same chromosome as a result of LD. Thus, appropriate multi-locus analyses of SNP haplotypes can jointly provide evidence of association for relatively modest gene effects with realistic sample sizes, even when the individual markers do not.

A convenient foundation for the development of statistical methods that take account of the joint information across multiple linked SNPs is the logistic regression modelling framework. Assuming multiplicative disease risks, the model is parameterised in terms of the odds of disease for each haplotype. Within this framework, it is straightforward to accommodate covariates, which may include environmental risk factors, polygenic effects, or genotypes at unlinked SNPs to allow for population structure [Pritchard and Rosenberg, 1999]. The logistic regression model can also be extended to allow for epistasis and gene-environment interactions.

A major drawback of haplotype-based analyses is the requirement of phase information, which cannot generally be recovered from the genotypes generated by current SNP typing technology. An obvious approach to dealing with this problem is to first reconstruct haplotypes using a statistical algorithm such as PHASE [Stephens et al., 2001; Stephens and Donnelly, 2003], and then proceed to analyse these haplotypes as if they were known to be correct. However, this approach does not allow for the uncertainty in the haplotype reconstruction process, and may lead to inflated estimates of the level of LD across the region [Morris et al., 2003] and over-confidence in any results obtained from the subsequent haplotype-based association analysis [Morris et al., 2004]. The appropriate approach to deal with unknown phase is to consider all possible haplotype configurations consistent with the observed SNP genotype data, weighted in the logistic regression likelihood by the corresponding phase assignment probabilities [Schaid et al., 2002; Zaykin et al., 2002; Stram et al., 2003].

However, a further problem with haplotype-based analyses is lack of parsimony, since one odds parameter is required for each haplotype. To reduce the dimensionality of the problem, we can take advantage of the expectation that ''similar'' marker haplotypes in the region flanking the disease gene have comparable disease risks. A number of methods have been proposed that cluster SNP haplotypes according to some similarity metric, and then assign the same disease odds to all haplotypes within the same cluster, reducing the number of parameters required [Templeton et al., 1987, 1988, 1992; Templeton and Sing, 1993; Molitor et al., 2003a,b; Durrant et al., 2004].

In this report, we develop a novel method for the analysis of population-based association studies using unphased SNP genotype data directly. We begin by obtaining maximum likelihood estimates of the relative frequencies of haplotypes consistent with the observed SNP genotypes via implementation of the expectation-maximisation (E-M) algorithm [Excoffier and Slatkin, 1995]. Under the Bayesian partition model [Knorr-Held and Rasser, 2000; Denison and Holmes, 2001], we specify ''cluster centres'' from the set of consistent SNP haplotypes, with each cluster allocated a disease odds parameter. The remaining SNP haplotypes are then assigned to the ''nearest'' centre, where similarity is defined in terms of marker allele matches. A similar approach has been utilised by Molitor et al. [2003b] in the context of fine-scale mapping with phased haplotype data. Uncertainty in phase assignment is addressed by considering all possible haplotype configurations consistent with each unphased genotype, weighted in the logistic regression likelihood by their probabilities, calculated according to the estimated relative haplotype frequencies [Schaid et al., 2002; Zaykin et al., 2002]. In this way, we naturally allow for missing genotype data by considering all haplotypes consistent with each possible genotype at an untyped locus. We develop a reversible jump Markov chain Monte Carlo (MCMC) algorithm to sample over the space of haplotype clusters and corresponding odds, allowing for additional covariates. Output from the algorithm can be used to: (1) estimate relative-risks of disease for each haplotype consistent with the observed unphased genotype data, treating the most common haplotype as baseline; (2) identify clusters of haplotypes with similar disease risks; (3) identify groups of cases carrying the same high-risk variants; and (4) estimate the posterior probability of haplotype association with the disease.

We illustrate the method by application to high-density unphased genotype data collected across an 890-kb region flanking the CYP2D6 for association with a recessive poor drug metaboliser (PDM) phenotype [Hosking et al., 2002]. Our analysis provides overwhelming evidence of association of the PDM phenotype with SNP haplotypes across the candidate region. Further, by constructing a dendogram of common SNP haplotypes consistent with the observed unphased marker genotype data, we identify two high-risk clusters, each associated with a different mutation in CYP2D6. We are also able to distinguish PDM cases carrying two copies of the most common high-risk mutation at the CYP2D6 locus, from those carrying other, rarer mutations. Finally, we present a detailed simulation study to evaluate the performance of the method to detect haplotype associations across a candidate gene (or small candidate region <100-kb) for a putative disease gene. The results are encouraging, indicating that this approach can be used to identify associations for moderate relative disease risks with a sample of 1,000 cases and 1,000 controls.

MODEL AND METHODS

Consider a case-control sample of N unrelated individuals, typed at M marker SNPs in a candidate gene or region, yielding genotypes $\mathbf{G}$, with alleles coded 1 and 2 at each locus, and 0 denoting missing data. The disease status of individual i is denoted y_i=1 if affected and y_i=0 if unaffected, with additional covariates denoted $\mathbf{x}_i$. The set of n distinct marker SNP haplotypes consistent with the observed genotypes, $\mathbf{G}$, is denoted $\mathcal{H} = \{\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_n\}$ where $\mathcal{H}_1$ is the most common haplotype. Relative haplotype frequencies, $\mathbf{h}$, are estimated by means of maximum likelihood from the available genotype data via implementation of the E-M algorithm.

Within a logistic regression framework, the key parameters of interest are haplotype log-odds of disease, ψ. Under the Bayesian partition model, ψ is determined precisely by the assignment of haplotypes to clusters, referred to as a tessellation, $\mathbf{T}$, and corresponding cluster log-odds, β. Thus, we can obtain estimates of the haplotype log-odds by considering the joint posterior density function $f(\mathbf{T}, \beta, \theta \,|\, \mathbf{y}, \mathbf{G}, \mathbf{h}, \mathbf{x})$, where θ denotes a set of additional model parameters, including covariate regression coefficients, γ. In particular, the marginal posterior distribution of haplotype log-odds can be obtained by integration,

$$f(\psi \,|\, \mathcal{D}) = \int_{\theta} f(\mathbf{T}, \beta, \theta \,|\, \mathcal{D}) \partial\theta,$$

where $\mathcal{D} = \{\mathbf{y}, \mathbf{G}, \mathbf{x}, \mathbf{h}\}$ denotes observed data and relative haplotype frequencies. By Bayes' theorem,

$$f(\mathbf{T}, \beta, \theta \,|\, \mathcal{D}) \propto f(\mathbf{y} \,|\, \mathbf{G}, \mathbf{h}, \mathbf{x}, \mathbf{T}, \beta, \theta) f(\mathbf{T}, \beta, \theta), \quad (1)$$

where $f(\mathbf{y} \,|\, \mathbf{G}, \mathbf{h}, \mathbf{x}, \mathbf{T}, \beta, \theta)$ denotes the likelihood of disease phenotypes given the assignment of marker SNP haplotypes to clusters in tessellation $\mathbf{T}$, the corresponding cluster log-odds β and additional model parameters, θ, and $f(\mathbf{T}, \beta, \theta)$ denotes their joint prior density.

HAPLOTYPE TESSELLATION STRUCTURE

A tessellation, $\mathbf{T}$, is defined by specifying K cluster centres, $\mathbf{C}$=$\{C_1, C_2, \ldots, C_K\}$, ordered and without replacement from the set of haplotypes $\mathcal{H}$. The haplotype $\mathcal{H}_j$ is then assigned to the cluster with maximum similarity metric, defined as

$$S_{jk} = \frac{1}{M} \sum_{m=1}^{M} s_{jk[m]}$$

for cluster center C_k, denoted $\mathcal{T}[\mathcal{H}_j]$. The SNP similarity metric, $s_{jk[m]}$, is given by

$$s_{jk[m]} = \begin{cases} 1 & \text{if } \mathcal{H}_{j[m]} = C_{k[m]} \\ 0 & \text{if } \mathcal{H}_{j[m]} \neq C_{k[m]}, \end{cases}$$

where $\mathcal{H}_{j[m]}$ and $C_{k[m]}$ denote the allele present at SNP m on haplotype $\mathcal{H}_j$ and cluster centre C_k, respectively. If haplotype $\mathcal{H}_j$ is equidistant from more than one cluster centre, it is assigned to that with minimum k.

LIKELIHOOD CALCULATION

For unphased genotype data, there is a set of P_i ordered haplotype pairs, $H_i = \{H_i^1, H_i^2, \ldots, H_i^{P_i}\}$, consistent with the observed genotype G_i for individual i. The likelihood term can then be expressed as a summation over H_i, weighted by the corresponding phase probabilities,

$$f(\mathbf{y} \,|\, \mathbf{G}, \mathbf{h}, \mathbf{x}, \mathbf{T}, \beta, \theta) \propto \prod_{i=1}^{N} \sum_{p=1}^{P_i} f(Y_i \,|\, H_i^p, \mathbf{x}_i, \mathbf{T}, \beta, \gamma) f(H_i^p \,|\, G_i, \mathbf{h}). \quad (2)$$

Assuming Hardy-Weinberg equilibrium,

$$f(H_i^p \,|\, G_i, \mathbf{h}) = \frac{h_{i1}^p h_{i2}^p}{f(G_i \,|\, \mathbf{h})},$$

where h_{i1}^p and h_{i2}^p are the relative frequencies of the pair of haplotypes H_{i1}^p and H_{i2}^p in configuration H_i^p,

and $f(G_i \mid \mathbf{h}) = \sum_{p=1}^{p_i} h_{i1}^p h_{i2}^p$. Within a logistic regression framework,

$$f\left(Y_i \mid H_i^p, \mathbf{x}_i, \mathbf{T}, \beta, \gamma\right) = \frac{\exp\left(\eta_i^p\right)^{y_i}}{1 + \exp\left(\eta_i^p\right)}.$$

Assuming a multiplicative disease model, the linear component is given by

$$\eta_i^p = \beta_{T\left[H_{i1}^p\right]} + \beta_{T\left[H_{i2}^p\right]} + \sum_{l=1}^{L} \gamma_l x_{il},$$

where $T\left[H_{i1}^p\right]$ and $T\left[H_{i2}^p\right]$ denote the assignments of the pair of haplotypes H_{i1}^p and H_{i2}^p, respectively, to clusters in tessellation $\mathbf{T}$.

PRIOR DENSITY FUNCTION

Covariate regression coefficients, γ, are assumed to be distributed MVN(0, $\sigma_C^2(\mathbf{x}'\mathbf{x})^{-1}$), independent of the tessellation and cluster log-odds, a priori [George and McCullogh, 1993]. Conditional on the number of clusters, K, in the tessellation, cluster log-odds, β, are assumed to be distributed MVN($\mu 1$, $\sigma_B^2 \mathbf{I}_K$), and independent of the choice of cluster centres, $\mathbf{C}$, a priori. In defining the tessellation, each of the n distinct haplotypes in $\mathcal{H}$ has equal prior probability of selection as one of the K cluster centres, $\mathbf{C}$. Hence, the joint prior density function can be expressed as

$$\begin{aligned} f(\mathbf{T}, \beta, \theta) =& f(\mathbf{C} \mid K) f(\beta \mid K, \mu, \sigma_B) \\ & \times f(K) f(\mu) f(\sigma_B) f(\gamma \mid \sigma_C) f(\sigma_C), \end{aligned}$$

where $f(\mathbf{C} \mid K) \propto (n - K)!$ and

$$f(\gamma \mid \sigma_C) \propto \frac{\det(\mathbf{x}'\mathbf{x})}{\sigma_C} \exp\left[-\frac{\gamma'(\mathbf{x}'\mathbf{x})\gamma}{2\sigma_C^2}\right],$$

$$f(\beta \mid K, \mu, \sigma_B) \propto \frac{1}{\sigma_B} \prod_{k=1}^{K} \exp\left[-\frac{(\beta_k - \mu)^2}{2\sigma_B^2}\right].$$

The unconditional prior density of the number of clusters is given by

$$f(K) = \begin{cases} 0.5 & \text{if } K = 1 \\ \left[0.5^{-K}\left(1 - 0.5^{(n-1)}\right)\right]^{-1} & \text{if } K > 1 \end{cases}$$

Under this model, the prior probability of exactly one cluster is 0.5, whilst the prior probability of more than one cluster has a truncate geometric distribution. The prior mean cluster log-odds has a prior uniform distribution so that $f(\mu) \propto 1$. The prior standard deviations of β and γ, have exponential distributions with expectation 1, a priori, given by $f(\sigma_B) \propto \exp[-\sigma_B]$ and $f(\sigma_C) \propto \exp[-\sigma_C]$, respectively.

MCMC ALGORITHM

We have developed a Metropolis-Hastings MCMC algorithm [Metropolis et al., 1953; Hastings, 1970] to approximate the joint posterior density of model parameters $\mathcal{Z} = \{\mathbf{C}, K, \beta, \gamma, \mu, \sigma_B, \sigma_C\}$, expressed as $f(\mathcal{Z} \mid \mathcal{D})$ in equation (1). The dimensionality of the parameter space depends on the number of clusters, K, of haplotypes. To account for this, we incorporate a birth-death process for the number of clusters via implementation of a reversible-jump step in the MCMC algorithm [Green, 1995]. At each step of the algorithm, a candidate set of new parameter values, $\mathcal{Z}'$, is proposed by making a ''small'' change to the current parameter set. The candidate values are accepted in place of $\mathcal{Z}$ with probability $f(\mathcal{Z}' \mid \mathcal{D})/f(\mathcal{Z} \mid \mathcal{D})$. Otherwise, the current values of $\mathcal{Z}$ are retained. Full details of the algorithm are presented in Appendix A2.

The algorithm is run for an initial ''burn-in'' period to allow convergence from a randomly selected set of starting values of $\mathcal{Z}$. In the subsequent sampling period, each parameter set accepted, or retained, by the algorithm represents a random draw from the posterior density (1). To reduce autocorrelation between consecutive draws, only every tth set of parameter values, $\mathcal{Z}$, is recorded for some suitably large t.

Output of the MCMC algorithm can be used directly to approximate the posterior distribution of the log-odds, ψ_j, of haplotype $\mathcal{H}_j$. Over R recorded MCMC outputs, the posterior mean of the log-odds of haplotype $\mathcal{H}_j$ is given by

$$\hat{\psi}_j = \frac{1}{R} \sum_{r=1}^{R} \beta_{T\left[\mathcal{H}_j\right]^{(r)}}^{(r)},$$

where $\beta_{T\left[\mathcal{H}_j\right]^{(r)}}^{(r)}$ denotes the log-odds of the cluster to which haplotype $\mathcal{H}_j$ is assigned in the rth output.

Assuming the disease to be rare, we can approximate the posterior mean log-relative risk, ϕ_j, of haplotype $\mathcal{H}_j$, treating the most common haplotype, $\mathcal{H}_j$ as baseline, given by

$$\hat{\phi}_j = \frac{1}{R} \sum_{r=1}^{R} \left[\beta_{T\left[\mathcal{H}_j\right]^{(r)}}^{(r)} - \beta_{T[\mathcal{H}_1]^{(r)}}^{(r)}\right],$$

with posterior variance

$$\mathbf{V}\left(\phi_j\right) = \frac{1}{(R-1)} \sum_{r=1}^{R} \left[\hat{\phi}_j - \beta_{T\left[\mathcal{H}_j\right]^{(r)}}^{(r)} + \beta_{T[\mathcal{H}_1]^{(r)}}^{(r)}\right]^2.$$

POSTERIOR PROBABILITY OF ASSOCIATION

In the absence of disease-marker association in the candidate gene, we expect all haplotypes to have the same risk, and hence to fall into a single cluster. We can thus approximate the posterior probability of haplotype association with disease, $\rho = f(K>1 \mid \mathcal{D})$, given by the proportion of MCMC outputs for which the number of clusters exceeds 1. The prior probability of more than one cluster of haplotypes, $f(K>1)$, is 0.5, so that $\hat{\rho}>0.5$ is ''suggestive'' of association. By convention, $\hat{\rho}>0.75$ is taken as ''positive'' evidence of association, $\hat{\rho}>0.95$ as ''strong'' evidence, whilst ''overwhelming'' evidence corresponds to $\hat{\rho}>0.99$ [Kass and Raftery, 1995].

POSTERIOR SUMMARY OF HAPLOTYPE DIVERSITY

Our prior model of the haplotype tesselation structure takes account only of pairwise diversity due to their allelic makeup. We propose a posterior measure of similarity between a pair of haplotypes that also takes account of similarity due to disease risk, given by the proportion of MCMC outputs for which they are assigned to the same cluster of the tesselation. This similarity metric can be used to construct a dendogram to summarise the posterior tesselation of haplotypes using standard average-linkage hierarchical clustering techniques [Hartigan, 1975].

SOFTWARE AVAILABILITY

The GENEBPM software has been developed to: (1) obtain maximum likelihood estimates of the relative frequencies of haplotypes consistent with a sample of observed SNP genotypes via application of the E-M algorithm; and (2) implement the reversible jump algorithm to sample over the space of haplotype clusters and corresponding odds under the Bayesian partition model, allowing for additional covariates in the logistic regression framework. GENEBPM is available as a linux executable on request from the author, together with additional software to summarise the output of the algorithm.

EXAMPLE APPLICATION

The gene CYP2D6 on human chromosome 22q13 has an established role in drug metabolisation [Evans and Relling, 2000]. Hosking et al. [2002] genotyped 1,018 individuals at 32 SNP markers across an 890-kb region flanking CYP2D6 to evaluate the efficacy of LD mapping methods to identify the gene. The sample was also typed for four known functional polymorphisms in CYP2D6. A total of 41 individuals were found to carry two mutant alleles across any of the four functional polymorphisms, and hence were predicted to be recessive poor drug metaboliser (PDM) cases. No additional covariates were obtained.

We present the results of analysis of the candidate region using the GENEBPM algorithm developed here to identify high-risk haplotypes across the 32 marker SNPs, but excluding the functional polymorphisms. Implementation of the E-M algorithm identified 878 marker SNP haplotypes consistent with the observed genotype data. Each run of the MCMC algorithm consisted of an initial 100,000-iteration burn-in period to allow convergence from a random starting parameter set. In the subsequent 1-million-iteration sampling period, output of the algorithm was recorded every t=1,000th iteration. The total run time of the algorithm, including relative haplotype frequency estimation, was less than 12 h on a dedicated Pentium IV processor.

Figure 1 presents a summary of the output from a single run of the MCMC algorithm (1,000 recorded sampling outputs). Figure 1a illustrates a trace of the scaled log-likelihood to check convergence. Figure 1b illustrates the corresponding autocorrelation function, providing no evidence of correlation between outputs. Figure 1c presents a trace of the number of clusters, K, of haplotypes, with the corresponding approximation to the posterior distribution presented in Figure 1d. The number of clusters ranges from 3 to 22, with a mode of 5. There is overwhelming evidence of haplotype association with PDM, with posterior probability $\hat{\rho}>0.999$. Figures 1e and 1f present approximations to the posterior distributions of model hyperparameters (μ and σ_B). The posterior mean of μ is -5.026, with posterior standard deviation 2.231, compared with the prior mean of 0. Similarly, the posterior distribution of σ_B has mean 3.596, with standard deviation 1.221, compared with the prior mean of 1.

Figure 2 presents a dendogram of the 41 marker SNP haplotypes with estimated relative frequency $h_j \geq 0.5\%$, to illustrate the posterior similarities between them in terms of disease risk and allelic makeup, generated from the output of a single run of the MCMC algorithm. Haplotypes are coded according to their relative frequency, where 1 denotes the most common. Broadly, we can

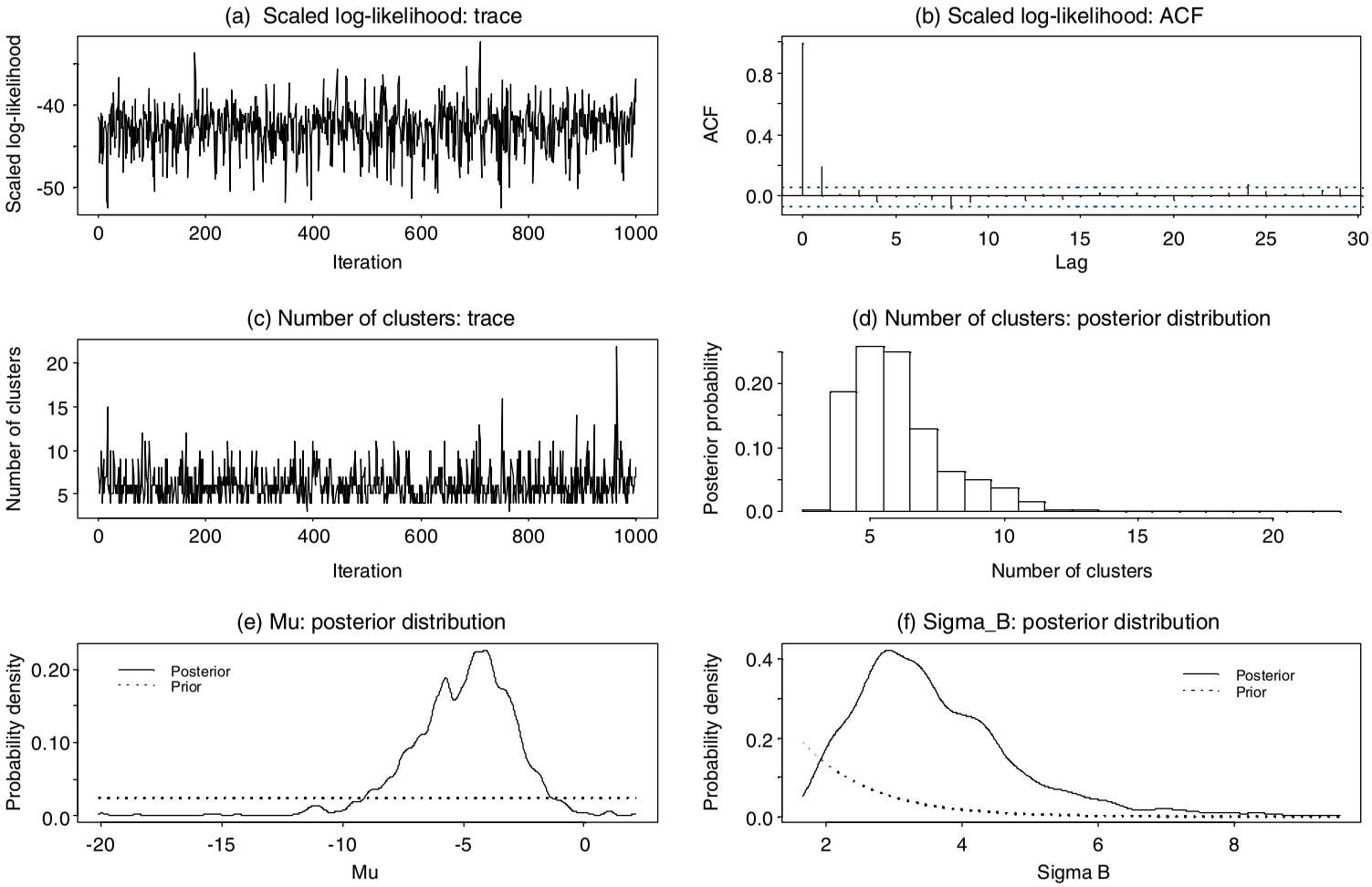

Fig. 1. Summary of the output from a single run of the MCMC algorithm: (a) trace of the scaled log-likelihood; (b) autocorrelation function of the scaled log-likelihood; (c) trace of the number of clusters, *K*, of haplotypes; (d) posterior distribution of the number of clusters, *K*, of haplotypes; (e) posterior distribution of μ; (f) posterior distribution of σ_B.

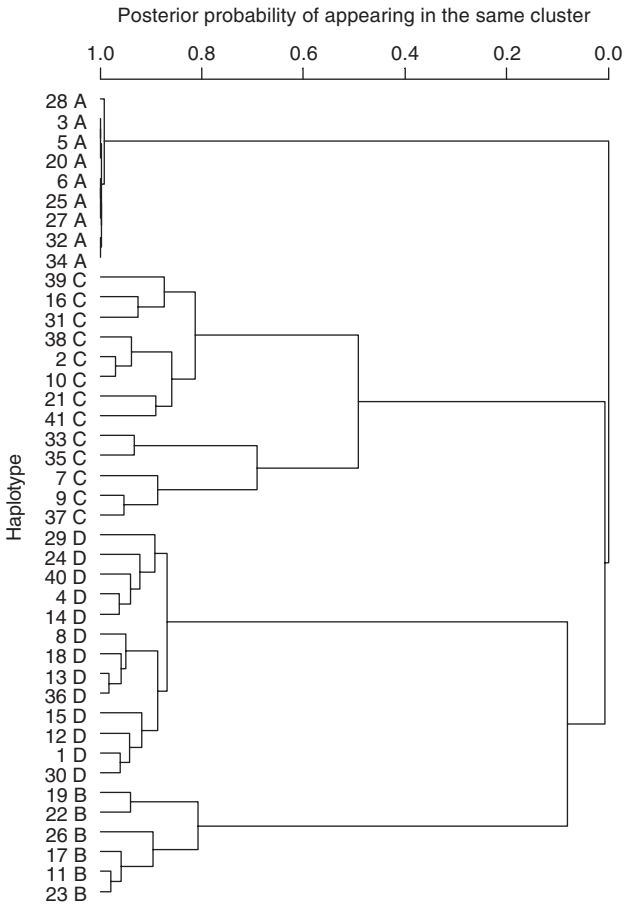

Fig. 2. Dendogram of the 41 marker SNP haplotypes with estimated relative frequency $h_j \geq 0.5\%$, to illustrate the similarities between them in terms of disease risk and allele matching, generated from the output of a single run of the MCMC algorithm. Haplotypes are coded according to their relative frequency, where 1 denotes the most common.

identify four clusters of haplotypes, summarised in Table I, and labelled A–D in Figure 2. Cluster A contains high-risk haplotypes, with posterior mean log-relative risks in the range 7.25–7.28, taking the most common haplotype as baseline. This cluster of haplotypes carries the most common high-risk mutation in CYP2D6. Cluster B also contains high-risk haplotypes, this time carrying a rarer mutation in CYP2D6, with lower posterior mean log-relative risks, ranging from 2.12 to 3.05. Clusters C and D contain lower-risk haplotypes, with posterior mean log-relative risks of less than 0.24.

As a final stage in the analysis, we consider the relatedness of the 41 PDM cases, illustrated by the dendogram presented in Figure 3, constructed using hierarchical clustering techniques, based on the output from a single run of the MCMC algorithm. Here, the similarity between a pair of individuals is defined by the posterior mean number of haplotypes they share from the same cluster over all MCMC outputs. For the rth output, the mean sharing is calculated over all combinations of possible phase assignments for the two individuals, weighted by the product of their relative probabilities. For each combination of phase assignments, sharing is scored as 2 if the individuals share both pairs of haplotypes from the same cluster(s), 1 if the individuals share one pair of haplotypes from the same cluster, and 0 otherwise. Figure 3 indicates the genotype of each PDM case at the CYP2D6 locus, where 1 and 2 are mutations associated with clusters A and B, respectively, and 3 is a much rarer mutation. The dendogram distinguishes individuals with different genotypes at CYP2D6 with remarkable accuracy. The 32 individuals carrying the 1/1 genotype at CYP2D6 form a tight cluster, with posterior mean haplotype cluster sharing of near 2, as expected. The same is true for the 7 individuals with CYP2D6 genotype 1/2. The 40 individuals carrying at least one copy of the common mutation at CYP2D6 (genotypes 1/1, 1/2, and 1/3) also form a cluster, with posterior mean haplotype sharing of approximately 1, again as expected.

To assess the effect of sporadic, non-genetic cases of PDM on the haplotype association, we repeated our analysis of the Hosking et al. [2002] sample, but with 41 randomly selected controls mislabeled as cases. Controls carry at most one copy of any mutation in CYP2D6: among the subset of mislabeled individuals, genotypes 0/0 (no mutations), 0/1 (one copy of the mutation carried by cluster A), and 0/2 (one copy of the mutation carried by cluster B) were observed. Despite the increased heterogeneity among the cases, our analysis still provided overwhelming evidence of haplotype association with PDM across the candidate region ($\hat{\rho} > 0.999$). Figure 4 presents a dendogram to illustrate the relatedness of the 41 PDM cases, and 41 mislabeled controls, generated from the output from a single run of the MCMC algorithm. This time, the two main clusters of the dendogram distinguish individuals carrying at least one copy of the most common mutation (genotypes 0/1, 1/1, 1/2, and 1/3) from those not carrying the most common mutation (genotypes 0/0, 0/2, and 2/3). Within the common mutation clade, the dendogram successfully identifies a single cluster of individuals carrying

TABLE I. Posterior mean (Standard Deviation) log-relative haplotype risks, $\hat{\phi}_j$, for PDM phenotype across 890kb candidate region flanking gene CYP2D6, treating the most common haplotype, $\mathcal{H}_1$, as baseline

Cluster	j	h_j (%)	ψ_j	ϕ_j(SD)	Marker SNP haplotype $\mathcal{H}_j$
A	3	3.42	1.322	7.282 (1.567)	1111111112112111112111121222111
	5	2.42	1.322	7.282 (1.567)	1111111112112111112111121221111
	6	2.35	1.324	7.284 (1.567)	1111111112112111112111121111111
	20	0.97	1.322	7.282 (1.567)	1111111112112111112111121221112
	25	0.86	1.324	7.284 (1.567)	1111111112112111112111121111112
	27	0.82	1.323	7.283 (1.568)	1111111112112111112112121111111
	28	0.78	1.288	7.248 (1.635)	1112111111112111112111111221111
	32	0.73	1.324	7.284 (1.568)	1111111111112111112111121111111
	34	0.67	1.322	7.282 (1.568)	1111111111112111112111121221111
B	11	1.66	−2.983	2.977 (1.537)	2212211111121221111211111111111
	17	1.10	−2.907	3.054 (1.502)	2212211111121221111211111111221
	19	0.98	−3.748	2.212 (2.405)	2212211111121221112211111111111
	22	0.93	−3.841	2.119 (2.533)	2212211111121221112211111112111
	23	0.92	−3.023	2.937 (1.567)	2212211111121221111211111111121
	26	0.86	−3.298	2.662 (1.892)	2212211111121221111211111221111
C	2	4.66	−8.591	−2.631 (2.943)	1111111111111122211111112221111
	7	2.34	−8.460	−2.500 (3.250)	2212211121212112121111111112111
	9	1.71	−8.524	−2.564 (3.178)	2212211121212112121111111111121
	10	1.71	−8.574	−2.614 (2.951)	1111111112111112221111112221111
	16	1.20	−8.531	−2.571 (2.914)	1111122211211112221111112221111
	21	0.96	−8.567	−2.607 (3.075)	1111111111111122211111112111221
	31	0.76	−8.539	−2.579 (3.076)	1111122211211112221111112221112
	33	0.68	−8.124	−2.164 (3.220)	2212211121211112221111112221111
	35	0.65	−8.159	−2.198 (3.151)	2212211122211112221111112221111
	37	0.64	−8.388	−2.427 (3.242)	2212211121212112121111111111221
	38	0.64	−8.598	−2.638 (3.075)	1111111111111112221111112221112
	39	0.55	−8.495	−2.535 (3.040)	1111122211211112221111112221221
	41	0.51	−8.634	−2.674 (3.118)	1111111111211112221111112111111
D	1	6.90	−5.960	BASELINE	1111111111121221111211111221111
	4	3.02	−5.725	0.235 (1.384)	1111111111121221111211111111121
	8	1.91	−6.069	−0.108 (1.342)	1111111111121221111211111112111
	12	1.65	−5.988	−0.027 (0.710)	1111111111121221111211111221112
	13	1.52	−5.980	−0.020 (0.821)	1111111111121221111211111111111
	14	1.40	−5.778	0.182 (1.572)	1111111111121221111211111111221
	15	1.33	−6.009	−0.049 (0.902)	1111111111121221112211111221111
	18	1.00	−6.020	−0.060 (1.099)	1111111111121221111111111111111
	24	0.90	−5.853	0.107 (1.773)	1111112221121221111211111111121
	29	0.76	−5.834	0.126 (1.488)	1111111111121221111211111221221
	30	0.76	−5.980	−0.019 (0.784)	1111111111121221111111111221111
	36	0.65	−6.009	−0.048 (0.983)	1111111112121221111211111111111
	40	0.53	−5.781	0.179 (1.364)	1111111111121221111211111121121

Rank j refers to ordered relative frequency among all haplotypes, where 1 is the most common (baseline), and corresponds to labels in Figure 2.

two copies of the most common mutation (genotype 1/1).

SIMULATION STUDY

We present details of a simulation study to investigate the utility of the proposed method for detecting haplotype associations across a candidate gene or small region (<100-kb) for a complex disease with 1% population prevalence. We investigate the properties of the posterior probability of haplotype association, $\hat{\rho}$, for a range of complex disease models, encompassing one or two high-risk variants. For each model, we generate 500 replicates of unphased case-control

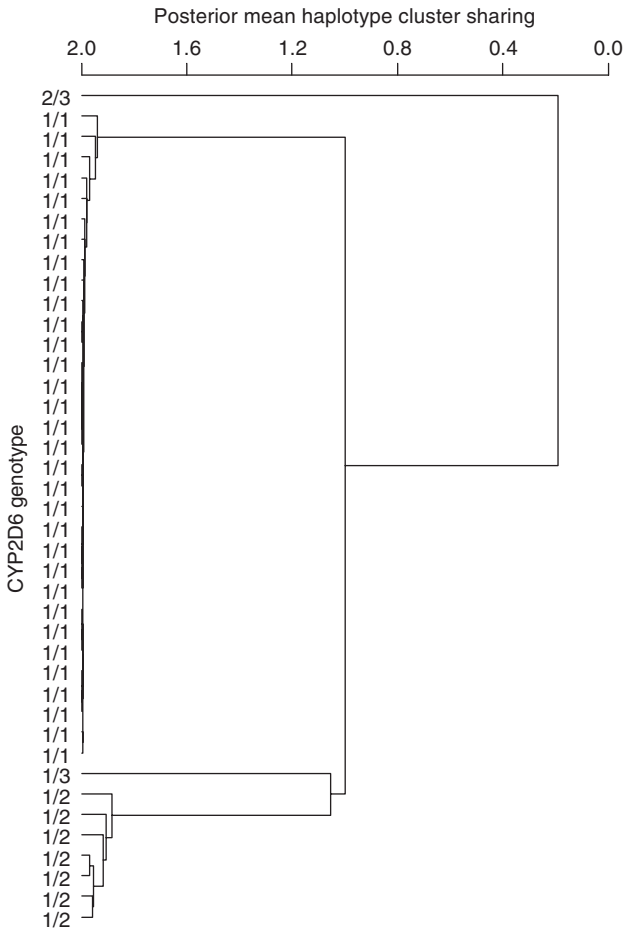

Fig. 3. Dendogram of the 41 PDM cases to illustrate their relatedness in terms of the mean number of haplotypes shared from the same cluster. PDM cases are coded according to their genotype at the CYP2D6 locus, where 1 is the most common high-risk mutation, and 2 and 3 are rarer high-risk variants.

marker genotype data, allowing for the bias towards common SNPs in the public databases. Each replicate is obtained as follows:

1. Generate an ancestral recombination graph [Griffiths and Marjoram, 1997] for a population of 20,000 SNP haplotypes from a realisation of the coalescent process with recombination, obtained using the MS software [Hudson, 2002]. We assume a mutation rate of 10^{-8} per base (in each generation) and a recombination rate of 1 cM per Mb, for an effective population size of 10,000 individuals, corresponding to scaled recombination and mutation rates of 400 per Mb [Nordborg, 2001].
2. Select the high-risk disease variant(s) at random from all SNPs ''close'' to a pre-specified minor allele frequency (MAF), q.
3. Test each remaining SNP for ascertainment in the public databases: the probability that a SNP is ascertained is given by $4 \times \text{MAF} \times (1\text{-MAF})$.
4. Select the required number of marker SNPs, at random, from all ascertained SNPs.
5. Generate a diploid individual by sampling a pair of haplotypes at random, and with replacement, from the population of 20,000 chromosomes. Generate the phenotype of this individual according to their genotype(s) at the disease variant(s) and the disease model. Repeat this step until the required numbers of affected cases and unaffected controls are obtained.
6. Retain the unphased genotypes of each individual only at the marker SNPs, ordered by physical location.

For each replicate of data, we perform a single run of the GENEBPM algorithm. After haplotype frequency estimation, each run of the MCMC algorithm consists of an initial 100,000-iteration burn-in period to allow convergence from a random starting parameter set, with output recorded every 1,000th iteration of the subsequent 1-million-iteration sampling period. From the output of the algorithm, we estimate the posterior probability, $\hat{\rho}$, of haplotype association with the disease.

NULL MODEL: NO HIGH-RISK DISEASE VARIANTS

We begin by considering the null model of no disease-marker association across the candidate region. Table II presents the mean posterior probability of association, $\hat{\rho}$, together with the proportion of replicates for which there is positive evidence ($\hat{\rho} > 0.75$) and strong evidence ($\hat{\rho} > 0.95$) of haplotype association, for a range of sample sizes, across two different candidate regions. There is some evidence to suggest that the mean posterior probability of association increases as the sample size is reduced. However, the mean posterior probability of association remains below the prior probability of 0.5, and the proportion of replicates with false-positive evidence of association is less than 1%, regardless of sample size or the candidate region under consideration.

ONE HIGH-RISK DISEASE VARIANT

We consider next the case of disease-marker association generated as a result of a single high-risk variant in the candidate gene for a sample of 1,000 cases and 1,000 controls. This model of

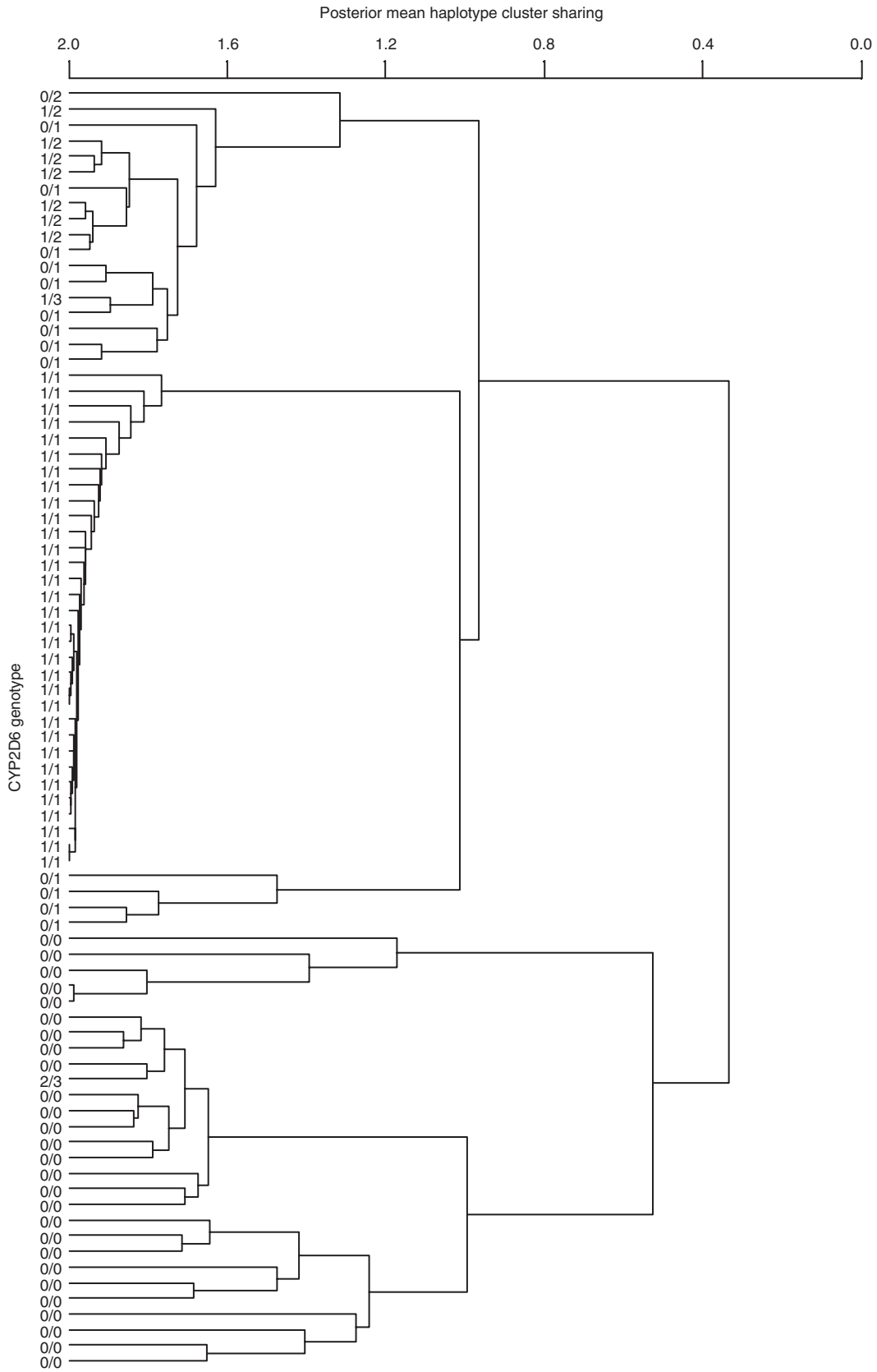

Fig. 4. Dendogram of the 41 PDM cases and 41 mislabeled controls to illustrate their relatedness in terms of the mean number of haplotypes shared from the same cluster. Individuals are coded according to their genotype at the CYP2D6 locus, where 0 corresponds to a low-risk variant, and 1, 2 and 3 are high-risk mutations.

TABLE II. Mean posterior probability of association, $\hat{\rho}$ together with the proportion of replicates for which there is positive evidence ($\hat{\rho} > 0.75$) and strong evidence ($\hat{\rho} > 0.95$) of haplotype association, for a range of sample sizes, across two different candidate regions, in the absence of a disease gene

				Proportion of replicates	
Candidate region	Sample size cases/controls	Mean number of haplotypes	Mean $\hat{\rho}$	$\hat{\rho} > 0.75$	$\hat{\rho} > 0.95$
5 SNPs in 50kb	200/200	11.73	0.415	0.000	0.000
	500/500	12.68	0.398	0.006	0.000
	1,000/1,000	13.11	0.385	0.002	0.002
10 SNPs in 100kb	200/200	41.05	0.417	0.004	0.002
	500/500	47.27	0.400	0.002	0.000
	1,000/1,000	50.51	0.387	0.002	0.002

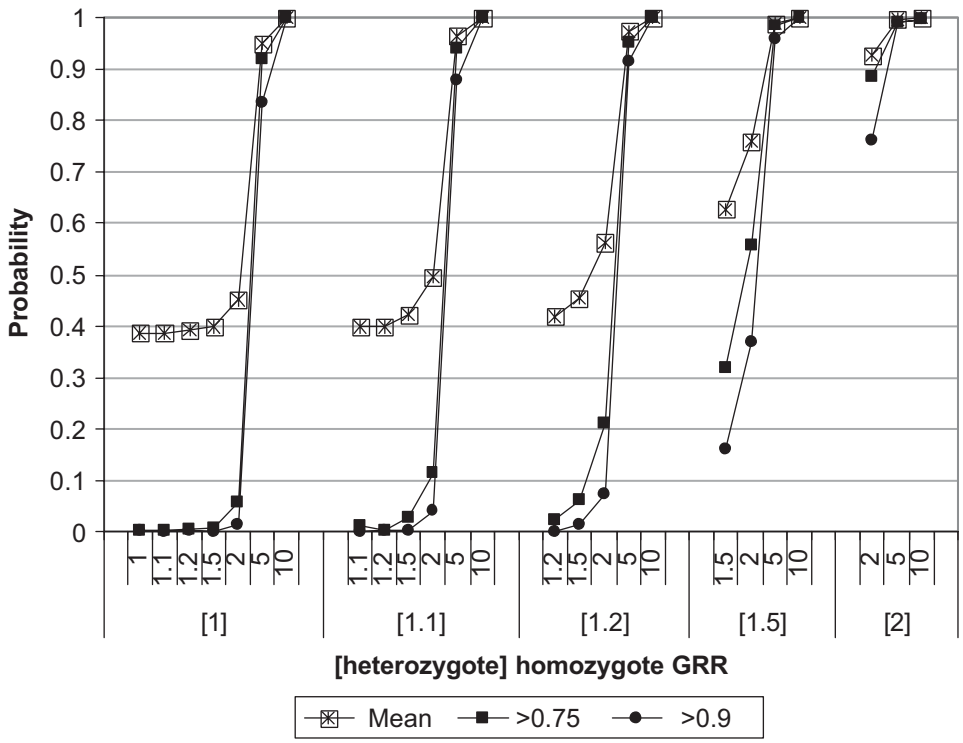

Fig. 5. Mean posterior probability of association, $\hat{\rho}$, and proportion of replicates for which there is positive evidence ($\hat{\rho} > 0.75$) and strong evidence ($\hat{\rho} > 0.95$) of haplotype association, as a function of homozygote and heterozygote genotype relative risks (GRR) for a high-risk variant frequency of 0.2, in a 100-kb candidate region spanned by 10 SNPs, for a sample of 1,000 cases and 1,000 controls.

association is parameterised in terms of the population relative frequency of the high-risk variant and the genotype relative risks (GRRs) of individuals homozygous and heterozygous for the high-risk variant, with the homozygous low-risk variant genotype taken as baseline.

Figure 5 presents the mean posterior probability of association, $\hat{\rho}$, together with the proportion of replicates for which there is positive evidence ($\hat{\rho} > 0.75$) and strong evidence ($\hat{\rho} > 0.95$) of haplotype association, as a function of GRRs for a high-risk variant frequency of 0.2, in a 100-kb candidate region spanned by 10 SNPs. For moderate relative risks (GRRs of 1.5), the mean posterior probability of association exceeds the prior of 0.5. In addition, the proportion of replicates with positive evidence of association exceeds 30%, even with a sample of just 1,000 cases and 1,000 controls.

Figures 6 and 7 present the proportion of replicates for which there is positive evidence

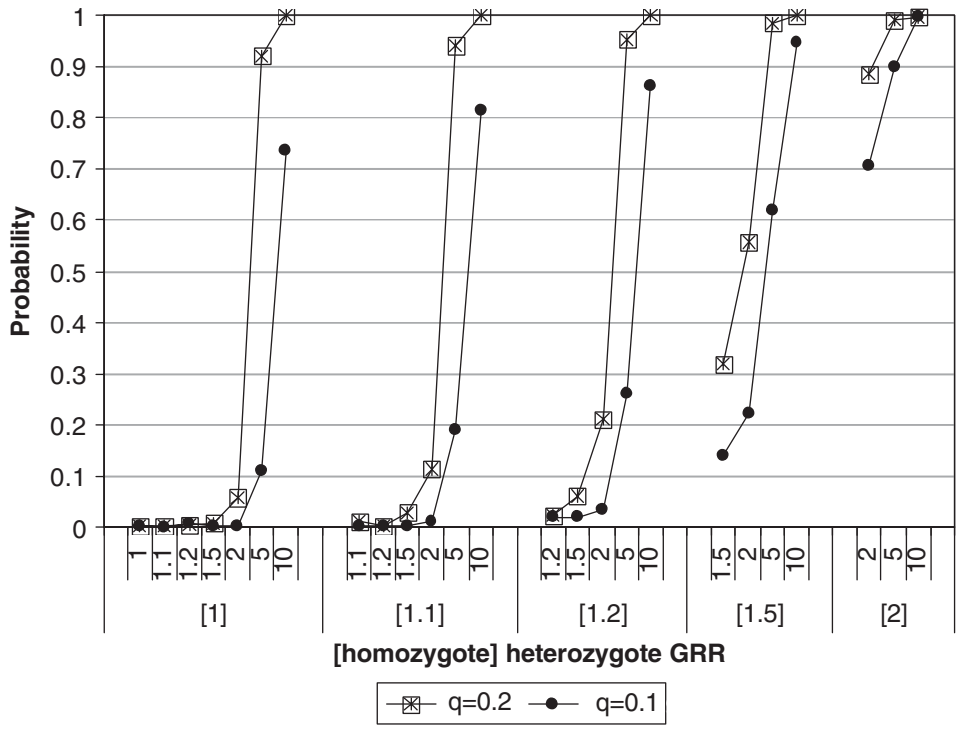

Fig. 6. Proportion of replicates for which there is positive evidence ($\hat{\rho} > 0.75$) of association as a function of homozygote and heterozygote genotype relative risks (GRR), for high-risk variant frequencies, q, of 0.1 and 0.2, in a 100-kb candidate region spanned by 10 SNPs, for a sample of 1,000 cases and 1,000 controls.

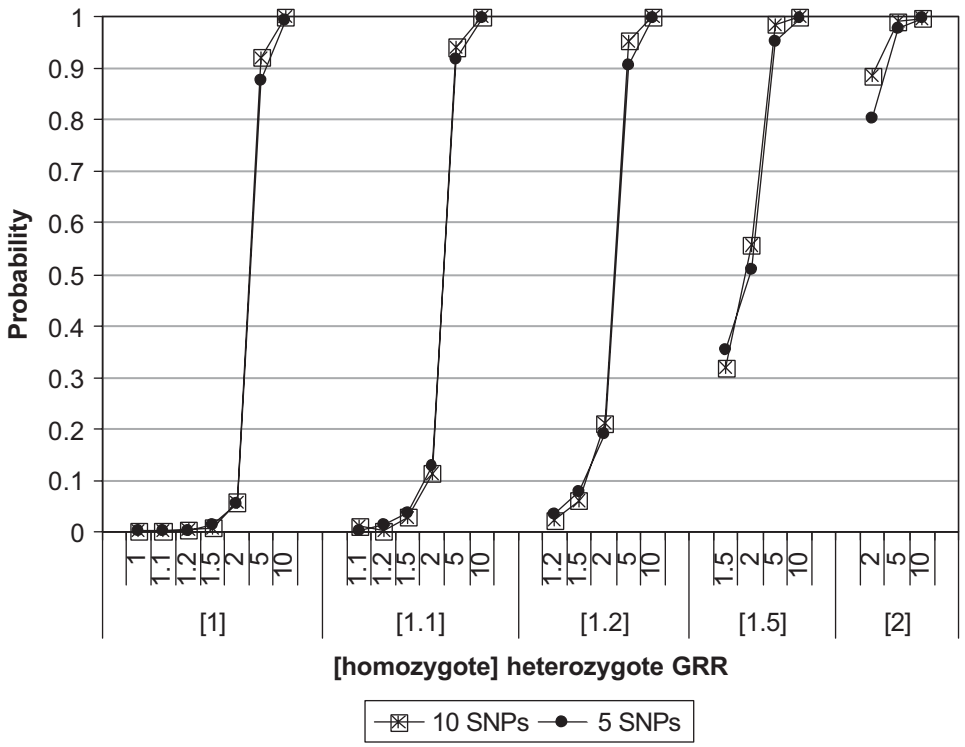

Fig. 7. Proportion of replicates for which there is positive evidence ($\hat{\rho} > 0.75$) of association as a function of homozygote and heterozygote genotype relative risks (GRR), for a high-risk variant with frequency 0.2, in two different candidate regions: 10 SNPs spanning 100-kb and 5 SNPs spanning 50-kb, each for a sample of 1,000 cases and 1,000 controls.

TABLE III. Mean posterior probability of association, $\hat{\rho}$, together with the proportion of replicates for which there is positive evidence ($\hat{\rho} > 0.75$) and strong evidence ($\hat{\rho} > 0.95$) of haplotype association, as a function of genotype relative risk (GRR) for two interacting high-risk variants, each with frequency of 0.1 or 0.2, in a 100-kb candidate region spanned by 10 SNPs, for a sample of 1,000 cases and 1,000 controls

Frequency of high-risk variants	GRR	Mean $\hat{\rho}$	Proportion of replicates	
			$\hat{\rho} > 0.75$	$\hat{\rho} > 0.95$
0.2	1.1	0.391	0.006	0.002
	1.2	0.404	0.014	0.006
	1.5	0.506	0.146	0.068
	2	0.719	0.516	0.408
	5	0.993	0.998	0.966
0.1	1.1	0.388	0.006	0.002
	1.2	0.390	0.006	0.000
	1.5	0.422	0.040	0.014
	2	0.562	0.270	0.208
	5	0.759	0.570	0.456

($\hat{\rho} > 0.75$) of association as a function of GRRs, for high-risk variant frequencies of 0.1 and 0.2, in two different candidate regions. The probability that we obtain positive evidence of association is less for the rarer high-risk variant. This is not unexpected since, for rarer variants, individuals carrying two high-risk alleles contribute less to the overall prevalence of the disease. A similar pattern of results is observed for both candidate regions.

TWO HIGH-RISK DISEASE VARIANTS

Finally, we consider a range of models for disease-marker association generated by two interacting high-risk variants in the same candidate gene. We assume that the two high-risk variants occur with the same population relative frequency, and interact with positive epistasis. Under this model, any individual carrying at least one high-risk variant at *both* loci has the same GRR of disease, with all other genotypes taken as baseline.

Table III presents the mean posterior probability of association, $\hat{\rho}$, together with the proportion of replicates for which there is positive evidence ($\hat{\rho} > 0.75$) and strong evidence ($\hat{\rho} > 0.95$) of haplotype association, as a function of GRR for two interacting high-risk variants, each with frequency of 0.1 or 0.2, in a 100-kb candidate region spanned by 10 SNPs. For the more common high-risk variants (frequency 0.2), the mean posterior probability exceeds the prior of 0.5 for moderate relative GRRs of 1.5, although the proportion of replicates with positive evidence of association is only 15%. A similar pattern of results is observed for the rarer high-risk variants (frequency 0.1), although the evidence of association is not as strong.

DISCUSSION

It is widely accepted that appropriate analyses of SNP haplotypes may provide evidence of association for the modest gene effects expected for complex traits with realistic sample sizes, even when the individual SNPs themselves do not. However, there are two major drawbacks of haplotype-based analyses with many SNPs: lack of parsimony and unknown phase.

Reducing the dimensionality of haplotype space to obtain a more parsimonious model of disease-marker association is not a new idea [Templeton et al., 1987]. Here, we take the same approach as Molitor et al. [2003b], clustering haplotypes according to a Bayesian partition model. We measure the similarity between pairs of haplotypes by the proportion of SNPs at which they carry the same allele. Such a metric is consistent with haplotype diversity driven by marker mutation, with minimal ancestral recombination, a reasonable assumption for candidate genes or small candidate regions. There are, of course, many other metrics. For example, we could weight SNP matches according to allele frequency [Durrant et al., 2004] or measure haplotype sharing around a putative disease locus [Molitor et al., 2003a, b] to extend the method for fine mapping. An alternative metric would treat all haplotypes as equally similar, with the result that clustering occurs with respect to disease risk, without regard to allelic makeup. However, this metric does not take account of the expected patterns of haplotype diversity generated as a result of their shared ancestry, and may have a less stable tesselation structure when there are rare haplotypes. In general, we would expect the equal similarity metric to perform less well than those taking account of allelic makeup, unless our model of haplotype evolution were inapporopriate. One such example would be a candidate region with high rates of recombination although, in this scenario, the phase assignment process would be inaccurate, and haplotype-based analyses would not be recommended.

We deal with unknown phase in the same way as Schaid et al. [2002] and Zaykin et al. [2002] by first obtaining maximum likelihood estimates of the relative frequencies of haplotypes consistent with the observed unphased SNP genotype data via implementation of the E-M algorithm. The resulting phase assignment probabilities calculated from these estimated haplotype frequencies are then treated as weights for each unphased genotype in the logistic regression model. However, we could easily incorporate the posterior probabilities of phase assignment generated by PHASE [Stephens et al., 2001; Stephens and Donnelly, 2003], or other Bayesian haplotype reconstruction algorithms. Alternatively, we could treat unknown phase as a latent variable to be updated in the MCMC algorithm, as implemented by Morris et al. [2004] in the context of fine mapping with unphased SNP genotype data, using the estimated phase assignment probabilities, a priori. However, this approach would add considerably to the computational burden of the algorithm.

We have illustrated here the utility of the GENEBPM algorithm in the analysis of retrospective case-control studies (as in the Simulation Study) and prospective cohorts (as in the Example Application). However, the prospective likelihood (2) does not take account of ascertainment. The over-representation of affected individuals in case-control samples will lead to inflated estimates of high-risk haplotype frequencies in the E-M algorithm, and will introduce bias in the corresponding haplotype relative-disease risk estimates. One solution to the problem would be to restrict haplotype frequency estimation to the control sample, but this may exclude rare high-risk haplotypes in the case sample. The correct approach would be to include an ascertainment correction in the propsective likelihood (2) as developed by Stram et al. [2003], with joint estimation of population haplotype frequencies and haplotype relative-disease risks. However, this is not possible here as the phase assignment probabilities are fixed in the prospective likelihood (2), given by relative haplotype frequencies estimated by the E-M algorithm, without regard to ascertainment. Nevertheless, if we are prepared to accept that the haplotype frequencies are nuisance parameters, Stram et al. [2003] have demonstrated that haplotype relative disease risk estimates are generally only slightly biased.

Within the Bayesian paradigm, we cannot formally test the null hypothesis of no haplotype association with disease. However, the GENEBPM algorithm can be used to approximate the posterior probability of association, given by the proportion of MCMC outputs for which there is more than one cluster in the haplotype tessellation. Assuming the prior probability of one cluster to be 0.5, a posterior probability of association of 0.75 corresponds to odds of 3:1 against the null hypothesis. Our simulation study suggests that this cut-off corresponds to a false-positive error rate of less than 1% for a range of sample sizes and candidate regions. However, to test more formally for association, we recommend generating the empirical null distribution of the posterior probability of association over many permutations obtained by randomly exchanging the case and control labels of pairs of individuals. Such an approach is computationally intensive, but not unrealistic.

The method presented here is designed for use in candidate genes or small candidate regions. The GENEBPM algorithm is currently limited to the analysis of haplotypes of up to 100 marker SNPs. To improve the efficiency of the E-M algorithm to allow for large numbers of SNPs in the initial haplotype frequency estimation procedure, haplotypes are built up locus-by-locus, ''culling'' phase assignments with low probability at each stage, in the same way as SNPHAP (http://www-gene.cimr.cam.ac.uk/clayton/software/snphap.txt). The algorithm could be used to analyse all SNPs in a candidate region, or a subset of tag SNPs selected to take advantage of the underlying patterns of LD between markers. Analysis of all SNPs may provide more ''refined'' clustering of haplotypes, but will be considerably less cost efficient than the tag SNP subset in terms of genotyping. However, the joint analysis of many SNPs across *large* candidate regions, or complete chromosomes in a genome scan, would not be realistic because of the effects of recombination on the clustering process and the expected inaccuracies in the phase assignment process. We could consider breaking the candidate region into ''blocks'' of strong LD, and to treat each block as independent. Alternatively, we could treat the candidate region as a sliding window of adjacent SNPs, with independent analyses performed within each window, and appropriate correction for multiple testing. We would expect peaks in the posterior probability of association to indicate the most likely regions to harbour genes contributing to disease risk, and this may help to prioritise further genotyping in an attempt to refine location.

ACKNOWLEDGMENTS

A.P.M. is grateful to Prof. David Balding, Dr. John Whittaker, and Dr. Nicky Best, from Imperial College Faculty of Medicine, London, for useful discussions. A.P.M. also thanks Prof. Duncan Thomas and an anonymous reviewer for helpful comments on the original version of the manuscript.

REFERENCES

Denison DGT, Holmes CC. 2001. Bayesian partitioning for estimating disease risk. Biometrics 57:143–149.

Durrant C, Zondervan KT, Cardon LR, Hunt S, Deloukas P, Morris AP. 2004. Linkage disequilibrium mapping via cladistic analysis of SNP haplotypes. Am J Hum Genet 75:35–43.

Excoffier L, Slatkin M. 1995. Maximum likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 12:921–927.

Evans W, Relling M. 2000. Pharmacogenomics: translating functional genomics into rational therapeutics. Science 286:487–491.

George EI, McCullogh RE. 1993. Variable selection via Gibbs sampling. J Am Stat Ass 88:881–889.

Green PJ. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82:711–732.

Griffiths RD, Marjoram P. 1997. An ancestral recombination graph. In: Donnelly P, Tavare S, editors. Progress in population genetics and human evolution. New York: Springer-Verlag. p. 257–270.

Hartigan JA. 1975. Clustering algorithms. New York: Wiley.

Hastings WK. 1970. Monte-Carlo sampling methods using Markov chains and their applications. Biometrika 57:97–109.

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

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model. Bioinformatics 18:337–338.

Kass RE, Raftery AE. 1995. Bayes factors. J Am Stat Assoc 90: 773–795.

Knorr-Held L, Rasser G. 2000. Bayesian detection of clusters and discontinuities in disease maps. Biometrics 46:13–21.

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. 1953. Equation of state calculations by fast computing machines. J Chem Phys 21:1087–1092.

Molitor J, Marjoram P, Thomas D. 2003a. Application of Bayesian spatial statistical methods to the analysis of haplotype effects and gene mapping. Genet Epidemiol 25:95–105.

Molitor J, Marjoram P, Thomas D. 2003b. Fine scale mapping of disease genes with multiple mutations via spatial clustering techniques. Am J Hum Genet 73:1368–1384.

Morris AP, Pedder A, Ayres K. 2003. Linkage disequilibrium assessment via log-linear modelling of SNP haplotype frequencies. Genet Epidemiol 25:106–114.

Morris AP, Whittaker JC, Balding DJ. 2004. Little loss of information due to unknown phase for fine-scale LD mapping with single-nucleotide-polymorphism genotype data. Am J Hum Genet 74:945–953.

Nordborg M. 2001. Coalescent theory. In: Balding DJ, Bishop M, Cannings C, editors. Handbook of statistical genetics. Chichester: Wiley. p. 179–212.

Pritchard JK, Rosenberg NA. 1999. Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet 65:220–228.

Risch N, Merikangas K. 1996. The future of genetic studies of complex human diseases. Science 273:1516–1517.

Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA. 2002. Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am J Hum Genet 70:425–434.

Stephens M, Donnelly P. 2003. A comparison of Bayesian methods for haplotype reconstruction from population genetic data. 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.

Stram DO, Pearce CL, Bretsky P, Freedman M, Hirschhorn JN, Altshuler D, Kolonel LN, Henderson BE, Thomas DC. 2003. Modelling and E-M estimation of haplotype-specific relative risks from genotype data for a case-control study of unrelated individuals. Hum Hered 55:179–190.

Templeton AR, Sing CF. 1993. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. IV. Nested analyses with cladogram uncertainty and recombination. Genetics 134:659–669.

Templeton AR, Boerwinkle E, Sing CF. 1987. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. I. Basic theory and an analysis of Alcohol Dehydrogenase activity in Drosophila. Genetics 117:343–351.

Templeton AR, Sing CF, Kessling A, Humphries S. 1988. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. II. The analysis of natural populations. Genetics 120:1145–1154.

Templeton AR, Crandall KA, Sing CF. 1992. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. Genetics 132:619–633.

Zaykin DV, Westfall PH, Young SS, Karnoub MA, Wagner MJ, Ehm MG. 2002. Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. Hum Hered 53:79–91.

Zondervan KT, Cardon LR. 2004. The complex interplay among factors that influence allelic association Nat Rev Genet 5:89–100.

APPENDIX A1

GLOSSARY OF NOTATION

y_i	phenotype of individual i: 0 unaffected and 1 affected
G_{im}	genotype of individual i at marker SNP m
x_{il}	response of individual i for lth covariate
$\mathcal{H}_j$	jth most frequent marker SNP haplotype consistent with genotype data
h_j	estimated relative frequency of haplotype $\mathcal{H}_j$

P_i	number of phase assignments consistent with marker SNP genotype of individual i
$\{H_{i1}^p, H_{i2}^p\}$	pair of marker SNP haplotypes constituting pth phase assignment consistent with genotype of individual i
ψ_j	log-odds of disease for marker SNP haplotype $\mathcal{H}_j$
ϕ_j	log-relative risk of disease for marker SNP haplotype $\mathcal{H}_j$, treating most common haplotype, $\mathcal{H}_1$, as baseline
K	number of clusters of marker SNP haplotypes
C_k	marker SNP haplotype centre of cluster k
$T[\mathcal{H}_j]$	cluster assignment of haplotype $\mathcal{H}_j$ in tessellation $\mathbf{T}$
β_k	log-odds of disease for cluster k
γ_l	regression parameter for lth covariate
μ	mean cluster log-odds of disease
σ_B	standard deviation of cluster log-odds of disease
σ_C	standard deviation of covariate regression parameters
ρ	posterior probability of haplotype association
θ	additional model parameters: $\{\gamma, \mu, \sigma_B, \sigma_C\}$

APPENDIX A2

DETAILS OF THE MCMC ALGORITHM

We have developed a reversible jump Metropolis-Hastings MCMC algorithm to approximate the posterior density function $f(\mathcal{Z} \mid \mathcal{D})$, given by equation (1) where $\mathcal{Z} = \{\mathbf{C}, K, \beta, \gamma, \mu, \sigma_B, \sigma_C\}$, and observed data $\mathcal{D} = \{\mathbf{y}, \mathbf{G}, \mathbf{x}, \mathbf{h}\}$. For each iteration of the algorithm, a new set of parameter values, $\mathcal{Z}'$, is proposed, according to predetermined weights, $\mathbf{w}$, summarised in Table IV. The proposed parameter values are substituted for the current set, provided that

$$\Delta \frac{f(\mathcal{Z}' \mid \mathcal{D})}{f(\mathcal{Z} \mid \mathcal{D})} > \varepsilon,$$

where ε is a standard uniform random variable, and Δ denotes the Hastings' ratio of proposal probabilities,

$$\Delta = \frac{\tau(\mathcal{Z}' \to \mathcal{Z})}{\tau(\mathcal{Z} \to \mathcal{Z}')}.$$

Otherwise, the current set of parameter values is retained. The possible changes to the parameter set are summarised below, where ε is a standard uniform random variable.

Change 1: Propose a cluster birth. The proposed number of clusters is given by K=K+1. Select a position, k^*, at random for the new cluster in the list of ordered cluster centres. Select a haplotype, $\mathcal{H}_j$, at random from $\mathcal{H}$, that is not already a cluster centre so that $C'_{k^*} = \mathcal{H}_j$. Generate a new cluster log-odds, β'_{k^*}, at random from a $N(\mu,\sigma_B^2)$ distribution. Then,

$$C'_k = C_k \text{ and } \beta'_k = \beta_k \text{ if } k < k^*$$
$$C'_k = C_{k+1} \text{ and } \beta'_k = \beta_{k+1} \text{ if } k > k^*.$$

To ensure reversibility, the Hastings ratio

$$\Delta = \frac{w_2(K')}{w_1(K) f(\beta'_{k^*} \mid \sigma_B)}$$

Change 2: Propose a Cluster Death. The proposed number of clusters is given by K=K−1. Select a cluster, k^*, at random for death. The

TABLE IV. Possible changes to the current parameter set in the reversible jump MCMC algorithm

			Relative weights $w_j(K)$		
Change j	Proposal	Parameters	K=1	1<K<n	K=n
1	Cluster birth	K, **C**, **T**, β	0.385	0.25	0
2	Cluster death	K, **C**, **T**, β	0	0.25	0.455
3	Cluster centre swap	**C**, **T**	0	0.1	0
4	Cluster centre change	**C**, **T**	0.154	0.1	0
5	Cluster log odds	β	0.092	0.06	0.109
6	Covariate regression coefficient	γ	0.092	0.06	0.109
7	Mean cluster log-odds	μ	0.092	0.06	0.109
8	Cluster log-odds SD	σ_B	0.092	0.06	0.109
9	Covariate regression coefficient SD	σ_C	0.092	0.06	0.109

proposed cluster centres and log-odds are then given by

$$C'_k = C_k \text{ and } \beta'_k = \beta_k \text{ if } k < k^*$$
$$C'_k = C_{k+1} \text{ and } \beta'_k = \beta_{k+1} \text{ if } k > k^*.$$

To ensure reversibility, the Hastings ratio

$$\Delta = \frac{w_1(K') f(\beta_{k^*} \mid \sigma_B)}{w_2(K)}$$

Change 3: Propose a Cluster Centre Swap. The following proposal procedure is carried out K times. Select a pair of clusters, k_1 and k_2, at random. The proposed cluster centre swap is given by

$$C'_{k1} = C_{k2} \text{ and } \beta'_{k1} = \beta_{k2}$$
$$C'_{k2} = C_{k1} \text{ and } \beta'_{k2} = \beta_{k1}.$$

The Hastings ratio $\Delta = 1$.

Change 4: Propose a cluster centre change. The following proposal procedure is carried out K times. Select a cluster k at random. Select a haplotype, $\mathcal{H}_j$, at random from $\mathcal{H}$, that is not already a cluster centre so that $C'_k = \mathcal{H}_j$. The Hastings ratio $\Delta = 1$.

Change 5: Propose a new cluster log-odds. The following proposal procedure is carried out K times. Select a cluster k at random. The proposed log-odds for the selected cluster is given by $\beta'_k = \beta_k + \nu_B(\varepsilon - 0.5)$, where ν_B denotes the maximum change in the parameter value. The Hastings ratio $\Delta = 1$.

Change 6: Propose a new covariate regression coefficient. The following proposal procedure is carried out L times. Select a covariate l at random. The proposed regression coefficient for the selected covariate is given by $\gamma'_l = \gamma_l + \nu_C(\varepsilon - 0.5)$, where ν_C denotes the maximum change in the parameter value. The Hastings ratio $\Delta = 1$.

Change 7: Propose a new prior mean cluster log-odds. The proposed mean is given by $\mu' = \mu + \nu_M(\varepsilon - 0.5)$, where ν_M denotes the maximum change in the parameter value.

Change 8: Propose a new prior cluster log-odds standard deviation. The proposed standard deviation is given by $\sigma'_B = \sigma_B + \nu_{SB}(\varepsilon - 0.5)$, where ν_{SB} denotes the maximum change in the parameter value. To ensure reversibility, $\sigma'_B = -\sigma'_B$ if $\sigma_B < 0$. The Hastings ratio $\Delta = 1$.

Change 9: Propose a new prior covariate regression coefficient standard deviation. The proposed standard deviation is given by $\sigma'_C = \sigma_C + \nu_{SC}(\varepsilon - 0.5)$, where ν_{SC} denotes the maximum change in the parameter value. To ensure reversibility, $\sigma'_C = -\sigma'_C$ if $\sigma_C < 0$. The Hastings ratio $\Delta = 1$.