University of Pennsylvania
UPenn Biostatistics Working Papers

Year 2009 *Paper* 31

A Hidden Markov Random Field Model for Genome-wide Association Studies

Hongzhe Li* Zhi Wei†
J M. Maris‡

*University of Pennsylvania, hongzhe@mail.med.upenn.edu
†zhiwei@mail.med.upenn.edu
‡maris@chop.edu

A Hidden Markov Random Field Model for Genome-wide Association Studies

Hongzhe Li, Zhi Wei, and J M. Maris

Abstract

Genome-wide association studies (GWAS) are increasingly utilized for identifying novel susceptible genetic variants for complex traits, but there is little consensus on analysis methods for such data. Most commonly used methods include single SNP analysis or haplotype analysis with Bonferroni correction for multiple comparisons. Since the SNPs in typical GWAS are often in linkage disequilibrium (LD), at least locally, Bonferonni correction of multiple comparisons often leads to conservative error control and therefore lower statistical power. In this paper, we propose a hidden Markov random field model (HMRF) for GWAS analysis based on a weighted LD graph built from the prior LD information among the SNPs and an efficient iterative conditional mode algorithm for estimating the model parameters. This model effectively utilizes the LD information in calculating the posterior probability that a SNP is associated with the disease. These posterior probabilities can then be used to define a false discovery controlling procedure in order to select the disease-associated SNPs. Simulation studies demonstrated the potential gain in power over single SNP analysis. The proposed method is especially effective in identifying SNPs with borderline significance at the single-marker level that nonetheless are in high LD with significant SNPs. In addition, by simultaneously considering the SNPs in LD, the proposed method can also help to reduce the number of false identifications of disease-associated SNPs. We demonstrate the application of the proposed HMRF model using data from a case-control genome-wide association study of neuroblastoma and identify one new SNP that is potentially associated with neuroblastoma.

A Hidden Markov Random Field Model for Genome-wide Association Studies

Hongzhe Li[1*], Zhi Wei[2] and John Maris[3]

[1*]Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, Philadelphia, PA 19104.

[2]Department of Computer Science, New Jersey Institute of Technology, Newark, NJ 07102.

[3]Division of Oncology and Center for Childhood Cancer Research, Childrens Hospital of Philadelphia; Department of Pediatrics, University of Pennsylvania School of Medicine, Philadelphia, PA 19104.

*Address correspondence to:

Hongzhe Li

Department of Biostatistics and Epidemiology

University of Pennsylvania School of Medicine

Philadelphia, PA 19104, USA.

Tel: (215) 573-5038

Email: hongzhe@upenn.edu

ABSTRACT

Genome-wide association studies (GWAS) are increasingly utilized for identifying novel susceptible genetic variants for complex traits, but there is little consensus on analysis methods for such data. Most commonly used methods include single SNP analysis or haplotype analysis with Bonferroni correction for multiple comparisons. Since the SNPs in typical GWAS are often in linkage disequilibrium (LD), at least locally, Bonferonni correction of multiple comparisons often leads to conservative error control and therefore lower statistical power. In this paper, we propose a hidden Markov random field model (HMRF) for GWAS analysis based on a weighted LD graph built from the prior LD information among the SNPs and an efficient iterative conditional mode algorithm for estimating the model parameters. This model effectively utilizes the LD information in calculating the posterior probability that a SNP is associated with the disease. These posterior probabilities can then be used to define a false discovery controlling procedure in order to select the disease-associated SNPs. Simulation studies demonstrated the potential gain in power over single SNP analysis. The proposed method is especially effective in identifying SNPs with borderline significance at the single-marker level that nonetheless are in high LD with significant SNPs. In addition, by simultaneously considering the SNPs in LD, the proposed method can also help to reduce the number of false identifications of disease-associated SNPs. We demonstrate the application of the proposed HMRF model using data from a case-control genome-wide association study of neuroblastoma and identify one new SNP that is potentially associated with neuroblastoma.

1 Introduction

Genome-wide association studies (GWAS) are increasingly utilized for identifying novel susceptible genetic variants for complex traits, but there is little consensus on optimal study design and analysis. GWAS are designed to scan the entire genome for the identification of genetic variations associated with phenotypic traits, such as a disease condition, blood

pressure, or body mass index. These studies usually examine several hundred thousand single nucleotide polymorphisms (SNPs) that explain most of the genetic variation across the genome as well as SNPs in a large array of candidate gene regions. Most GWAS involve some form of multistage design, which includes an initial scan of hundreds of thousands of SNPs on a sample of cases and controls, followed by testing a subset of the most promising markers on independent samples. Additional markers may also be included at later stages to better characterize the full spectrum of genetic variation in the targeted regions. GWAS have been demonstrated to be a powerful approach for the detection of genetic variants related to complex traits, such as age-related macular degenerative diseases (Klein *et al.*, 2005), prostate and breast cancers (Hunter *et al.*, 2007), and type 2 diabetes (Scott *et al.*, 2007). The Welcome Trust Case-Control Consortium has recently published a GWAS of seven diseases using 14,000 cases and 3000 shared controls (Nature, 2007). The success of these studies has provided solid evidence that GWAS represent a powerful approach to the identification of genes involved in common human diseases.

To date, the analytical methods of GWAS have largely been limited to the single SNP or SNP-SNP pair analysis, coupled with statistical techniques such as the Bonferroni procedure for controlling multiple comparisons. However, since SNPs in typical GWAS are in LD locally, simple Bonferonni correction can potentially be conservative and therefore lead to a loss of statistical power. In addition, if multiple SNPs are all in LD with the true disease variants, effectively utilizing the information from multiple SNPs in LD can potentially increase the power of detecting the SNPs associated with the disease. Local LD information has been mainly utilized in haplotype analysis based on sliding windows (Huang *et al.*, 2007). However, such haplotype analysis may lead to loss of power due to high degrees of freedom. In addition, it is often arbitrary to decide how many SNPs to consider in typical haplotype analysis. Browning and Browning (2007) developed an efficient multilocus association test for whole genome association studies using localized haplotype clustering, effectively using the local LD information, and demonstrated its application to the Wellcome Trust Case-Control Consortium data (Browning and Browning, 2008). Eskin (2008) proposed a data-

adaptive multiple comparison procedure to incorporate prior information of LD structure and molecular function in GWAS in order to increase power.

In this paper, we propose to formally incorporate LD information among the SNPs derived from the dataset itself or from the HapMap data (Altshuler *et al.*, 2005) into identifying the disease-associated SNPs. In particular, we will first build a weighted LD graph based on pair-wise LD measures among the SNPs. Such measures can either be derived from the HapMap project (Altshuler *et al.*, 2005) or from the dataset itself. We then propose a hidden Markov random field model on such an LD graph in order to compute the posterior probability that a SNP is associated with the disease. The key is that the posterior probability that a given SNP is associated with disease depends not only on the genotype data observed at this SNP, but also on the genotypes of the SNPs that are in strong LD with this particular SNP. In addition, we propose a simple empirical Bayes method to borrow information across all the markers in estimating the model parameters. We provide an efficient iterative conditional mode (ICM) algorithm (Besag, 1986) to estimate the parameters and Gibbs sampling approach for estimating the posterior probabilities. These posterior probabilities can be used to define a false discovery rate (FDR) controlling procedure in order to select the relevant SNPs. Different from the standard FDR control procedure of Benjamini and Hochberg (1995), this posterior probability-based FDR control accounts for SNP dependency explicitly and is expected to gain power in detecting the disease-associated SNPs. The proposed HMRF model for accounting for the LD dependency of the SNPs is somewhat similar to some recently developed HMRF models for the analysis of microarray gene expression data in order to account for known regulatory network information (Wei and Li, 2007; Wei and Li, 2008; Wei and Pan, 2007).

The rest of the paper is organized as follows: in Methods, we present our proposed HMRF model for SNPs data and the ICM algorithm for estimating the parameters. We also present a Gibbs sampling approach to estimate the posterior probabilities and use these probabilities to define the false discovery rate. In Results, we present results from the analysis of a case-control neuroblastoma dataset and simulation results to demonstrate the methods and to

compare the power with single-marker analysis. Finally, we present a brief summary and discussion of our methods and results.

2 A Hidden Markov Random Field Model for GWAS

2.1 Weighted LD graph and a Markov random field model

Suppose we have m cases and n controls that are genotyped over a set of p SNPs. Let $S = \{1, \cdots, p\}$ denote the SNP index. We want to determine which SNPs in S are associated with disease. Let $Y = (Y_1, ..., Y_s, \cdots, Y_p)$ be the observed genotype data for the p SNPs, where Y_s itself is a vector $Y_s = (y_{s1}, ..., y_{sm}; y_{s(m+1)}, ..., y_{s(m+n)})$, where y_{si} is the observed genotype for the ith individual at the sth SNP. The typical single SNP analysis often ignores the LD among these SNPs. Here we propose to develop a HMRF model to take into account the LD information in identifying the disease-associated SNPs. We first construct a weighted undirected LD graph G based on pair-wise LD information derived from the data or from the HapMap project. Specifically, an edge between SNPs s and s' is drawn with weight

$$w_{ss'} = I(r^2_{ss'} > \tau) r^2_{ss'},$$

if $w_{ss'} \neq 0$, where $I(.)$ is the indicator function, $r^2_{ss'}$ is the r^2 measurement of LD between SNPs s and s' and τ is a pre-determined cutoff value. We use $\tau = 0.4$ in our simulations and analysis. An example of such a LD graph is given in Figure 1 for the 998 SNPs we use for simulation in Section 3.2.

For a given SNP s, we then define a random indicator variable as

$$X_s = \begin{cases} 1 & \text{if SNP } s \text{ is associated with the disease} \\ 0 & \text{if SNP } s \text{ is not associated with the disease.} \end{cases}$$

For two SNPs s and s' that are linked on the LD graph, i.e., if the r^2 between these two SNPs are greater than τ, we expect that X_s and $X_{s'}$ are dependent. We propose to model such dependency using a simple discrete Markov random field model (Besag, 1974; Besag,

1986) with the following joint probability function for $X = (X_1, \cdots, X_p)$,

$$p(X; \Phi) \propto \exp(\gamma \sum_{s=1}^{p} X_s + \beta \sum_{s \sim s'} w_{s,s'} I(X_s = X_{s'})), \tag{1}$$

where γ and $\beta \geq 0$ are the two model parameters, and β measures dependencies of X_s for SNPs in LD. In this model, the parameter $\beta > 0$ encourages the SNPs that are in LD to have similar values of X_s. We assume that the true association state X is a realization of a locally dependent discrete MRF with a specified distribution $\{p(X)\}$. This is in contrast to the hidden Markov model where some time or spatial order of the SNPs has to be assumed. The conditional association state for SNP s, given the states of all neighboring SNPs is

$$p(X_s | X_{N_s}; \Phi) \propto \exp(\gamma X_s + \beta \sum_{s' \in N_s} w_{s,s'} I(X_s = X_{s'})),$$

where N_s represents the neighbors of the SNP s on the LD graph.

2.2 An empirical Bayes model for genotype data

In order to relate the latent vector X to the observed genotypes, we further assume that given any particular realization of X, the random variables $Y = (Y_1, ..., Y_p)$ are conditionally independent with the following conditional density,

$$l(Y|X) = \prod_{s=1}^{p} P(Y_s | X_s), \tag{2}$$

where $P(Y_s|X_s)$ is the joint probability of the observed genotypes over $m + n$ individuals at the SNP s given the latent state X_s.

In order to specify $f(Y_s|X_s)$, let $\theta_s = (\theta_{s1}, \theta_{s2}, \theta_{s3})$ be the genotype frequencies at the sth SNP in the case population, and $\rho_s = (\rho_{s1}, \rho_{s2}, \rho_{s3})$ be the genotype frequencies at the sth SNP in the control population, for genotype values of 0, 1 and 2, respectively. We assume that both of these frequencies across all the SNPs have a Dirichlet prior with parameter $\alpha = (\alpha_1, \alpha_2, \alpha_3)$,

$$f(\theta_s) = f(\theta_{s1}, \theta_{s2}, \theta_{s3}) = \frac{\Gamma(\sum_{j=1}^{3} \alpha_j)}{\prod_{j=1}^{3} \Gamma(\alpha_j)} \prod_{j=1}^{3} \theta_{sj}^{\alpha_j - 1}.$$

The same prior is also assumed for ρ_s. For SNP s, let $y_{s+} = (y_{s+,1}, y_{s+,2}, y_{s+,3})$ denote observed genotype counts data in the m cases and $y_{s-} = (y_{s-,1}, y_{s-,2}, y_{s-,3})$ denote the observed genotype counts data in n controls. So if SNP s is not associated with the disease, cases should have the same genotype frequencies as the controls. The combined genotype counts data $y_{s0} = y_{s+} + y_{s-}$ are generated from a trinomial distribution with the genotype frequencies of $\theta_s = (\theta_{s1}, \theta_{s2}, \theta_{s3})$. Thus, given $X_s = 0$ the probability of the combined genotype count data is

$$\begin{aligned} P(Y_s|X_s=0) &= P(y_{si}; i=1,...,m+n|X_s=0) = \int (y_{si}; i=1,...,m+n|X_s=0,\theta_s) f(\theta_s) d\theta_s \\ &= \int \prod_{i=1}^{3} \theta_i^{y_{s0}i} \times \frac{\Gamma(\sum_{j=1}^{3}\alpha_i)}{\prod_{i=1}^{3}\Gamma(\alpha_i)} \prod_{i=1}^{3} \theta_i^{\alpha_i - 1} d\theta_s \\ &= \frac{\Gamma(\sum_{j=1}^{3}\alpha_i)\prod_{i=1}^{3}\Gamma(\alpha_i + y_{s+,i} + y_{s-,i})}{\prod_{i=1}^{3}\Gamma(\alpha_i)\Gamma(\sum_{j=1}^{3}(\alpha_i + y_{s+,i} + y_{s-,i}))}. \end{aligned} \tag{3}$$

On the other hand, if SNP s is associated with the disease, i.e., when $X_s = 1$, cases and controls should have different genotype frequencies, in which case we have

$$\begin{aligned} P(Y_s|X_s=1) &= P(y_{si}; i=1,...m, m+1..., m+1|X_s=1) \\ &= \int (y_{si}; i=1,...,m|X_s=1,\theta_s) f(\theta_s) d\theta_s \\ &\quad \times \int (y_{si}; i=m+1,...,m+n|X_s=1,\rho_s) f(\rho_s) d\rho_s \\ &= \frac{\Gamma(\sum_{j=1}^{3}\alpha_i)\prod_{i=1}^{3}\Gamma(\alpha_i + y_{s+,i})}{\prod_{i=1}^{3}\Gamma(\alpha_i)\Gamma(\sum_{j=1}^{3}(\alpha_i + y_{s+,i}))} \times \frac{\Gamma(\sum_{j=1}^{3}\alpha_i)\prod_{i=1}^{3}\Gamma(\alpha_i + y_{s-,i})}{\prod_{i=1}^{3}\Gamma(\alpha_i)\Gamma(\sum_{j=1}^{3}(\alpha_i + y_{s-,i}))}. \end{aligned} \tag{4}$$

Together the probability models (1), (3) and (4) define a hidden Markov random field model, where X is the vector of the hidden states that follows a discrete Markov random field (1) and (3) and (4) define the emission probabilities.

2.3 Parameter estimation using the ICM algorithm and a FDR controlling procedure

In order to estimate the model parameters $\Phi = (\gamma, \beta)$ in the MRF model and $\alpha = (\alpha_1, \alpha_2, \alpha_3)$ in the emission probabilities, we propose to use the following ICM algorithm, which was

originally proposed by Besag (1986) for statistical image analysis. The algorithm involves iterative updates of model parameters and latent states and has the following steps:

1. Obtain an initial estimate $\hat{X}$ based on the single marker trend test using a p-value of 0.0001.

2. Estimate α with the value of $\hat{\alpha}$ by maximizing the probability of the observed data given by equation (2), $l(Y|\hat{X})$.

3. Estimate Φ with the value of $\hat{\phi}$ by maximizing the following pseudo-likelihood function,

$$\begin{aligned} l(\hat{X};\Phi) &= \prod_{s}^{p} p_s(\hat{X}_s|\hat{X}_{N_s};\Phi) \\ &= \prod_{s}^{p} \frac{\exp(\gamma\hat{X}_s + \beta\sum_{s'\in N_s} w_{s,s'} I(\hat{X}_s = \hat{X}_{s'}))}{\exp(\gamma + \beta\sum_{s'\in N_s} w_{s,s'})I(\hat{X}_{s'}=1) + \exp(\beta\sum_{s'\in N_s} w_{s,s'})I(\hat{X}_{s'}=0)}. \end{aligned}$$

The reason for using the pseudo-likelihood function here is that it is computationally difficult to compute the full likelihood function due to an unknown normalizing constant in the MRF model (1).

4. Carry out a single cycle of ICM based on the current $\hat{X}$, $\hat{\alpha}$ and $\hat{\phi}$, to obtain a new $\hat{X}$. Specifically, for $s = 1$ to p, update X_s based on

$$P(X_s|Y, \hat{X}_{S/s}) \propto f(Y_s|X_s;\hat{\alpha}) p_s(X_s|\hat{X}_{N_s};\hat{\Phi}). \tag{5}$$

5. Go to step **2** until $\max_{\theta\in(\alpha,\Phi)} \frac{|\theta^{(k+1)}-\theta^{(k)}|}{|\theta^{(k+1)}|} < 0.001$.

After the convergence of the algorithm, we propose to sample the latent vector X M times using Gibbs sampling based on the conditional probability (5). Based on these samples, we can estimate the posterior probability of $q_s = Pr(X_s = 0|Y)$. Let $q_{(s)}$ be the sorted values of q_s in descending order. For SNP s, the null hypothesis of interest is

$$\begin{aligned} H_{s0} &: \text{SNP } s \text{ is not associated with the disease,} \\ H_{s1} &: \text{SNP } s \text{ is associated with the disease.} \end{aligned}$$

Based on these posterior probabilities, let $k = max\{t : \frac{1}{t}\sum_{s=1}^{t} q_{(s)} \le \alpha\}$, then reject all $H_{(s)}, s = 1, \cdots, k$. This posterior probability-based definition of FDR has been widely used

in the analysis of microarray gene expression data (Newton *et al.*, 2001) and has been shown to control the FDR at the appropriate level α and to obtain optimal false non-discovery rates in the setting of hidden Markov models (Sun and Cai, 2008).

Note that if we fix $\beta = 0$ in the MRF model (1), i.e., if we do not consider the LD dependency of the SNPs, the procedure reduces to a simple empirical Bayes (EB) procedure for the analysis of genetic association data in order to borrow information across all the SNPs.

3 Results

3.1 Application to a Neuroblastoma Case-Control Data Set

Neuroblastoma (NB) is a common and lethal pediatric malignancy, but despite significant effort the genetic events that initiate tumorigenesis were until recently unknown (Mosse *et al.*, 2008). We had hypothesized that NB is a complex disease that results from the interaction of mutant alleles with relatively low to moderate effect on tumor initiation. To identify these genetic variants, Maris *et al.* (2008) reported a GWAS of NB where 1032 neuroblastoma cases and 2043 controls of European descent were genotyped using the Illumina 550K SNP chips and they observed a significant association between NB and the common minor alleles of three consecutive SNPs at chromosome band 6p22 and containing the predicted genes FLJ22536 and FLJ44180 (p-value $= 1.71\times10^{-9}$ to 7.01×10^{-10}; allelic odds ratio, 1.39 to 1.40) using single SNP trend tests. Homozygosity for the at-risk G allele of the most significantly associated SNP, rs6939340, resulted in an increased likelihood of NB development (odds ratio, 1.97; 95% confidence interval, 1.58 to 2.45).

To demonstrate our proposed HMRF modeling approach, we reanalyzed the 30,216 SNPs on chromosome 6 that passed the standard quality controls (see Maris *et al.* 2008 for details). We grouped these SNPs into groups of 1000 SNPs along the genome and fitted our proposed HMRF for each of these groups in order to facilitate the computation. We ran 10,000 Gibbs samples to obtain the posterior probability of a SNP being associated with the

Table 1: Results of analysis of NB data set for top nine SNPs with single SNP χ^2-test p-value less than 10^{-5}. Also shown are the the estimated posterior probability of being associated with NB based on the HMRF model (Post.prob) and the single SNP p-value for the replication data set (rep-p-value).

SNP	Location	χ^2 p-value	Post.Prob	rep-p-value
rs6939340	22247983	5.78e-09	1.00	0.0031
rs9295536	22239908	8.55e-09	1.00	0.012
rs4712653	22233943	9.23e-09	1.00	0.0048
rs12189640	27582636	2.57e-06	0.0046	0.28
rs4487594	133438173	2.73e-06	0.74	0.054
rs6929659	17773336	4.88e-06	0.014	0.43
rs2256175	31488428	5.45e-06	0.0025	0.73
rs10456051	27571415	8.56e-06	0.0041	0.29
rs858985	27286007	9.57e-06	0.0046	0.57

disease. Figure 2 shows the results from both single SNP analysis based on the chi-square tests and our proposed HMRF model. We observed that many SNPs have almost zero estimated probability of being associated with NB. We show in Table 1 the results for nine SNPs with chi-square test p-value less than 10^{-5}. In general, the results agree with each other well, where the three SNPs in gene FLJ22536 on chromosome 6, rs6939340, rs4712653 and rs9295536, showed both the most significant association and also the highest posterior probabilities (close to 1.00) of being associated with NB. However, two other SNPs, rs11759745 and rs9466269, with chi-square test p-values of 3.33×10^{-5} and 8.37×10^{-5}, have a posterior probability of being associated with NB of 0.9933 and 0.9881, clearly indicating that both are also associated with NB. These two SNPs are in high LD with the three SNPs that have a very strong association with NB, with r^2 values ranging from 0.42 to 0.69. This demonstrated that the proposed method is effective in identifying the SNPs with borderline

significance at the single marker level that nonetheless is in high LD with significant SNPs. If we chose these top five SNPs based on their estimated posterior probabilities, we expected a FDR of 0.0037.

Another SNP, rs4487594 on the 6q23.2 band of the chromosome 6, has a p-value of 2.73×10^{-6}, but with a relatively high posterior probability of 0.74 to be associated with NB. There are in fact no other SNPs that are in high LD with this SNP, and therefore the posterior probability was only determined by the observed genotypes at this SNP. This is in contrast to some other SNPs with similar single SNP p-values but low posterior probabilities; these SNPs all have neighboring SNPs that are in high LD with them but do not show any association with NB. If we also chose the SNP rs4487594 to be NB associated, the estimated FDR became 0.046. However, if we chose the top 10 SNPs, the FDR increased dramatically to 0.29. So at the FDR level of 0.046, we concluded that the SNP rs4487594 is also associated with NB risk. We tested the association between this SNP and NB in a replication set of 401 cases and 1178 controls and obtained a p-value of 0.054. This further indicated the possible association between the SNP rs4487594 and NB. In contrast, none of the five SNPs with similar single SNP p-values but with lower posterior posterior probabilities was significant in the replication set (see Table 1).

3.2 Simulation Studies

To demonstrate the proposed methods, we conducted a simulation study. In order to obtain more realistic LD patterns among the SNPs, our simulation was based on sampling from the 3075 individuals in the case-control study of neuroblastoma analyzed in the previous section. Particularly, we consider a region of 998 SNPs around the chromosome band 6p22 and select 20 SNPs as the true SNPs with relative risk of -1.5 for 10 such SNPs and 2.0 for another 10 SNPs. These disease-associated SNPs have 6 to 9 neighboring SNPs with $r^2 \geq 0.4$. Figure 1 shows the constructed LD graph for 857 SNPs with at least one neighbor based on the data from 3075 samples from the NB GWAS study. Based on the true genotypes of these 3075 individuals, we simulated the disease probability using the following logistic regression

model to give a disease rate of 0.10,

$$Pr(Y = 1|X) = \frac{\exp(\beta_0 + \sum_{k=1}^{20} \beta_k X_k)}{1 + \exp(\beta_0 + \sum_{k=1}^{20} \beta_k X_k)},$$

where $\beta_0 = -4.2$, $\beta_k = -\log(1.5)$ for $k = 1, \cdots, 10$ and $\beta_k = \log(2.0)$ for $k = 11, \cdots, 20$. Based on this model, we repeatedly generated the disease status for the 3075 individuals until we obtained 1000 cases and 1000 controls. The 20 true disease variants were then removed from our analysis. In order to assess the performance of selecting the relevant SNPs, the 117 SNPs that are in LD with the true variants with $r^2 \geq 0.4$ are defined as true positives and this definition of true positives is used in defining the sensitivities (SEN), specificities (SPE).

We repeated the simulation 100 times. For each simulation, we can calculate the SENs, SPEs and FDRs using different cutoff values of the posterior probabilities. Figure 3 shows the ROC curves for three different analysis methods, including the single SNP analysis, analysis based on an empirical Bayes method without utilizing the LD information and the proposed HMRF method using the LD information, averaged over 100 simulations. Due to the fact that the estimated posterior probability of being associated with disease for some SNPs can be 1, the ROC curve for the HMRF model does not start from the origin (0,0). It is clear that the methods without accounting for the LD resulted in lower sensitivity in identifying the disease-associated SNPs compared to our proposed HMRF method with LD information. However, we observed that the EB approach did not improve sensitivities over the single SNP analysis, resulting almost identical ROC curves. Similar results were also observed when SNPs that are in LD with the true variants with $r^2 \geq 0.6$ or $r^2 \geq 0.8$ are defined as true positives (results not shown).

4 Discussion

Since many SNPs in GWAS are in LD, it is important to efficiently utilize such LD information in order to increase the power of detecting disease-associated SNPs. In this paper, we introduce a hidden Markov random field model to account for LD in analysis of the

SNP data from GWAS, where the pair-wise LD measurements are used to define a weighted LD graph and a Markov random field model based on the constructed LD graph is used to model the dependency among the SNPs. In addition, a simple empirical Bayes model is proposed in order to borrow information across all the SNPs. Our simulation indicates that the proposed method can lead to an increase in sensitivity in identifying the SNPs that are associated with the disease. We used the GWAS of neuroblastoma to demonstrate the computational feasibility of the methods and the advantage that one can gain over the simple commonly used single SNP analysis. The proposed method is especially effective in identifying the SNPs with borderline significance at the single-marker level that nonetheless are in high LD with significant SNPs, as demonstrated by our analysis of the neuroblastoma dataset. In addition, by simultaneously considering the SNPs in LD, the proposed method can also help to reduce the number of false identifications of disease-associated SNPs.

Although Bonferonni correction has been widely applied in GWAS, analytical and simulation studies by Sabatti *et al.* (2003) have shown the the FDR procedure of Benjamini and Hochberg (1995) can effectively control the FDR for the dependent tests encountered in case-control association studies and increase power over more traditional methods. However, the direct application of the FDR procedure can lead to loss of power due to the dependency among tests, although the FDR can still be controlled (Benjamini and Yekutieli, 2001). The most effective way of correcting this problem relies on developing a precise model for the dependency among the SNPs and incorporating it in the definition of a FDR controlling procedure (Sabatti *et al.*, 2003). The HMRF model proposed provides one such model and we expect some gain in power over the standard FDR controlling procedure ignoring the dependency. For the hidden Markov model, Sun and Cai (2008) proved the optimal power of a posterior probability-based FDR procedure while controlling the FDR. It is easy to show that if the model we used is the true model and the true parameters are known, our proposed posterior probability-based FDR controlling procedure can indeed control the FDR with a minimal false non-discovery rate and therefore the optimal power. However, it is not clear that such a result holds when the parameters are estimated using the ICM algorithm. This

deserves further investigation.

The method in this paper was developed for case-control GWAS data; however, it can be easily extended to continuous quantitative traits where one only needs to redefine the emission probabilities. One such probability can be based on the empirical Bayes linear models as in the Limma method for microarray gene expression data analysis (Smyth, 2004; Li *et al.*, 2008). The method in this paper can also be extended to incorporate the prior genetic network such as the protein-protein interaction network information into the analysis of GWAS data, where one can create a weighted LD graph for the SNPs within a given gene. The LD-graphs are then linked based on the *a priori* gene or protein-protein interaction network to form a large SNP-network. A HMRF model can then be defined on this combined network that can help to identify the subnetworks of SNPs that might be related to disease risk. This network-based analysis of genomic data has shown some promise in the analysis of microarray gene expression data (Wei and Li, 2007; Wei and Li, 2008; Wei and Pan, 2007). We would also expect some potential gain in sensitivity in identifying the disease-associated genetic variants in the analysis of GWAS data. It would be interesting to investigate more on how much one can gain in analysis of GWAS data when the prior biological information is effectively utilized in the analysis.

Acknowledgments

This research was supported by NIH grants R01-ES009911, R01-CA127334 and R01-CA78454. We thank Tony Cai and Jichun Xie for discussing FDR controls and Mr. Edmund Weisberg, MS at Penn CCEB for editorial assistance.

References

Altshuler, D., Brooks L., Chakravarti, A., Collins, F., Daly, M., Donnelly, P., Consortium, I.H. 2005. A haplotype map of the human genome. Nature 437, 12991320.

Benjamini, Y. and Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society B 57, 289300.

Benjamini, Y. and Yekutieli, D. 2001. The control of the false discovery rate in multiple testing under independence. The Annals of Statistics 29:11651188.

Besag, J. 1974. Spatial interaction and the statistical analysis of lattice systems. Journal of the Royal Statistical Society, Series B 36, 192−225.

Besag, J. 1986. On the statistical analysis of dirty pictures. Journal of Royal Statistical Society B 48, 259−302.

Browning, B.L. and Browning, S.R. 2007. Efficient multilocus association testing for whole genome association studies using localized haplotype clustering. Genetic Epidemiology 31, 365−375.

Browning, B.L. and Browning, S.R. 2008. Haplotypic analysis of Wellcome Trust Case Control Consortium data. Human Genetics 123, 273−280.

Eskin, E. 2008. Increasing power in association studies by using linkage disequilibrium structure and molecular function as prior information. Genome Research 18, 653−660.

Huang, B.E., Amos, C.I. and Lin, D.Y. 2007. Detecting haplotype effects in genomewide association studies. Genetic Epidemiology 31, 803−812.

Hunter, D.J., Kraft, P., Jacobs, K.B., Cox, D.G., Yeager, M., Hankinson, S.E., Wacholder, S., Wang, Z., Welch, R., Hutchinson, A., Wang, J., Yu, K., Chatterjee, N., Orr, N., Willett, W.C., Colditz, G.A., Ziegler, R.G., Berg, C.D., Buys, S.S., McCarty, C.A., Feigelson, H.S., Calle, E.E., Thun, M.J., Hayes, R.B., Tucke,r M., Gerhard, D.S., Fraumeni, J.F. Jr, Hoover, R.N., Thomas, G., Chanock, S.J. 2007. A genome−wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nature Genetics 39, 870−874.

Klein, R.J., Zeiss, C., Chew, E.Y., Tsai, J.Y., Sackler, R.S., Haynes, C., Henning, A.K.,

SanGiovanni, J.P., Mane, S.M., Mayne, S.T., Bracken, M.B., Ferris, F.L., Ott, J., Barnstable, C., Hoh. J. 2005. Complement factor H polymorphism in age−related macular degeneration. Science 3085, 385−9.

Li, C., Wei, Z. and Li, H. 2008. A Network−constrained empirical Bayes method for Analysis of genomic data. UPenn Biostatistics Working paper.

Maris, J.M., Yael, P.M., Bradfield, J.P., Hou, C., Monni, S., Scott, R.H., Asgharzadeh, S., Attiveh, E.F., Diskin, S.J., Laudenslager, M., Winter, C., Cole, K., Glessner, J.T., Kim, C., Frackelton, E.C., Casalunovo, T., Eckert, A.W., Capasso, M., Rappaport, E.F., McConville, C., London, W.B., Seeger, R.C., Rahman, N., Devoto, M., Grant, S.F.A., Li, H. and Hakonarson, H. 2008. A genome-wide association study identifies a susceptibility locus to clinically aggressive neuroblastoma at 6p22. New England Journal of Medicine 358, 2585−2593.

Mosse, Y.P., Laudenslager, M., Longo, L., Cole, K.A., Wood, A., Attiyeh, E.F., Laquaglia, M.J., Sennett, R., Lynch, J.E., Perri, P., Laureys, G., Speleman, F., Kim, C., Hou, C., Hakonarson, H., Torkamani, A., Schork, N.J., Brodeur, G.M., Tonini, G.P., Rappaport, E., Devoto, M., Maris, J.M. 2008. Identification of ALK as a major familial neuroblastoma predisposition gene. Nature 455, 930-035.

Newton, M.A., Kendziorski, C.M., Richmond, C.S., Blattner, F.R. and Tsui, K.W. 2001. On differential variability of expression ratios: improving statistical inference about gene expression changes from micorarray data. Journal of Computational Biology 8, 37−52.

Sabatti, C., Service, S. and Freimer, N. 2003. False discovery rates in linkage and association linkage genome screens for complex disorders. Genetics 164, 829−833.

Scott, L.J., Mohlke, K.L., Bonnycastle, L.L., Willer, C.J., Li, Y., Duren, W.L., Erdos, M.R., Stringham, H.M., Chines, P.S., Jackson, A.U., Prokunina-Olsson, L., Ding, C.J., Swift, A.J., Narisu, N., Hu, T., Pruim, R., Xiao, R., Li, X.Y., Conneely, K.N., Riebow, N.L., Sprau, A.G., Tong, M., White, P.P., Hetrick, K.N., Barnhart, M.W., Bark, C.W., Goldstein, J.L., Watkins, L., Xiang, F., Saramies, J., Buchanan, T.A., Watanabe, R.M.,

Valle, T.T., Kinnunen, L., Abecasis, G.R., Pugh, E.W., Doheny, K.F., Bergman, R.N., Tuomilehto, J., Collins, F.S., Boehnke, M. 2007. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science 316, 1341−5.

Smyth, G.K. 2004. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Statistical Applications in Genetics and Molecular Biology 3(1), Article 3.

Sun, W, and Cai, T. 2008. Large-scale multiple testing under dependency. Journal of the Royal Statistical Society, Series B, in press.

Welcome Trust Case-control Consortium 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447, 661−678.

Wei, Z. and Li, H. 2007. A Markov random field model for network-based analysis of genomic data. Bioinformatics 23, 1537−1544.

Wei, Z. and Li, H. 2008. A hidden spatial-temporal Markov random field model for network-based analysis of time course gene expression data. Annals of Applied Statistics 2, 408−429.

Wei, P. and Pan, W. 2008. Incorporating gene networks into statistical tests for genomic data via a spatially correlated mixture model. Bioinformatics 24, 404−411.

Yeager, M., Orr, N., Hayes, R.B., Jacobs, K.B., Kraft, P., Wacholder, S., Minichiello, M.J., Fearnhead, P., Yu, K., Chatterjee, N., Wang, Z., Welch, R., Staats, B.J., Calle, E.E., Feigelson, H.S., Thun, M.J., Rodriguez, C., Albanes, D., Virtamo, J., Weinstein, S., Schumacher, F.R., Giovannucci, E., Willett, W.C., Cancel-Tassin, G., Cussenot, O., Valeri, A., Andriole, G.L., Gelmann, E.P., Tucker, M., Gerhard, D.S., Fraumeni, J.F. Jr, Hoover, R., Hunter, D.J., Chanock, S.J., Thomas, G. 2007. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. Nature Genetics 39, 645649.

Figure Legends

Figure 1: Example of a LD graph, which is used for the simulation study, where each node represents a SNP and a link between two SNPs indicates that r^2 is greater than 0.40 between them.

Figure 2: Results of analysis on 30,216 SNPs on chromosome 6 based on 1251 neuroblastoma cases and 2236 controls. Legend on the left-margin (-log of the p-values) and the circles are results from single SNP analysis and the legend on the right-margin (posterior probability of being associated with the neuroblastoma) and the triangles are results from the proposed HMRF model.

Figure 3: The ROC curves (sensitivity versus 1-specificity) for the proposed hidden MRF model (MRF), empirical Bayes method (EB) and the single SNP Cochran-Armitage trend test (Logistic). Simulation results based on 100 replications.

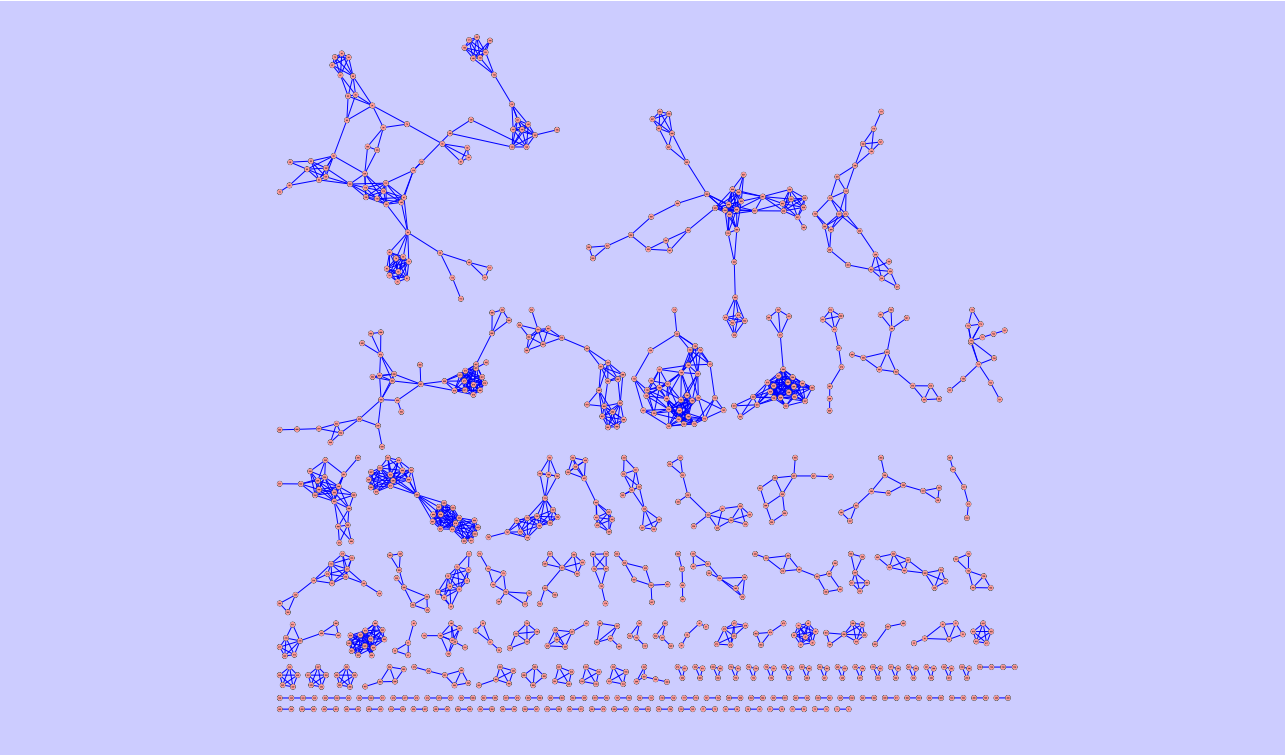

Figure 1: *Example of a LD graph, which is used for the simulation study, where each node represents a SNP and a link between two SNPs indicates that* r^2 *is greater than 0.40 between them.*

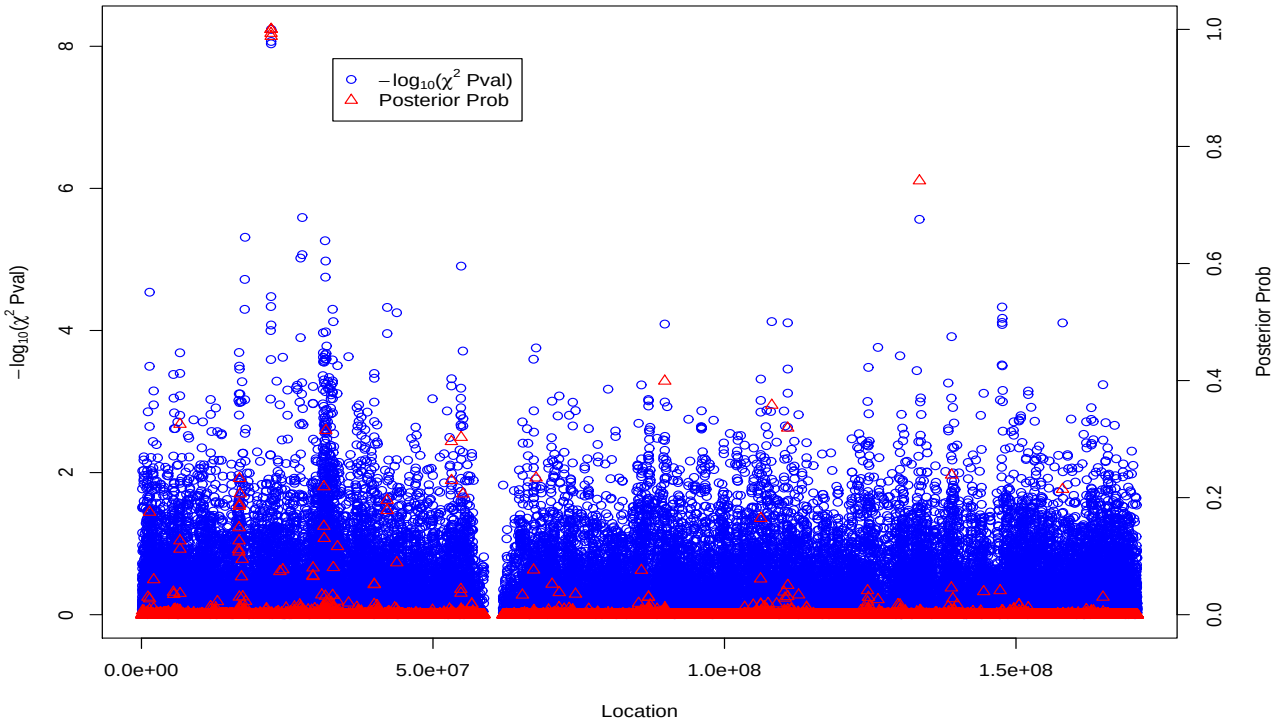

Figure 2: *Results of analysis on 30,216 SNPs on chromosome 6 based on 1251 neuroblastoma cases and 2236 controls. Legend on the left-margin (-log of the p-values) and the circles are results from single SNP analysis and the legend on the right-margin (posterior probability of being associated with the neuroblastoma) and the triangles are results from the proposed HMRF model.*

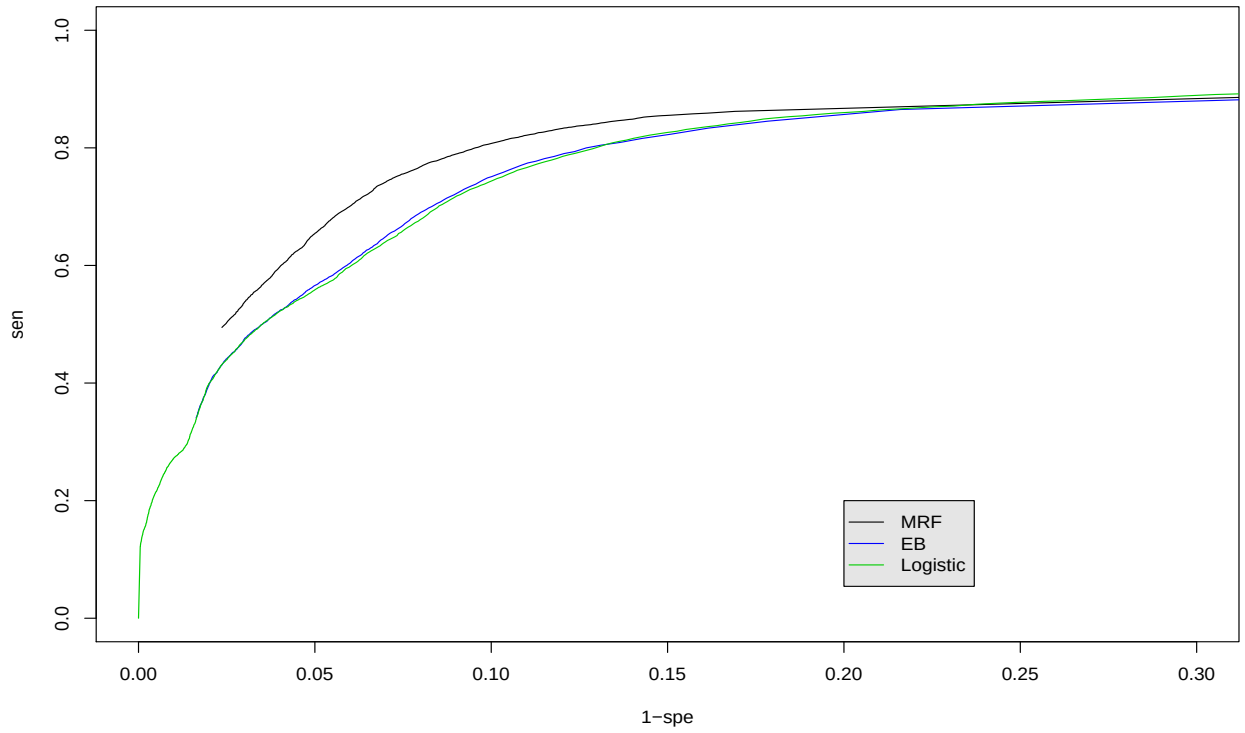

Figure 3: *The ROC curves (sensitivity versus 1-specificity) for the proposed hidden MRF model (MRF), empirical Bayes method (EB) and the single SNP Cochran-Armitage trend test (Logistic). Simulation results based on 100 replications.*