Incorporating Biological Pathways via a Markov Random Field Model in Genome-Wide Association Studies

Min Chen[1], **Judy Cho**[2], **Hongyu Zhao**[3]*

1 Division of Biostatistics, Department of Clinical Sciences, University of Texas Southwestern Medical Center, Dallas, Texas, United States of America, **2** Internal Medicine, Yale University, New Haven, Connecticut, United States of America, **3** Center for Statistical Genomics and Proteomics, Department of Epidemiology and Public Health, Yale University, New Haven, Connecticut, United States of America

Abstract

Genome-wide association studies (GWAS) examine a large number of markers across the genome to identify associations between genetic variants and disease. Most published studies examine only single markers, which may be less informative than considering multiple markers and multiple genes jointly because genes may interact with each other to affect disease risk. Much knowledge has been accumulated in the literature on biological pathways and interactions. It is conceivable that appropriate incorporation of such prior knowledge may improve the likelihood of making genuine discoveries. Although a number of methods have been developed recently to prioritize genes using prior biological knowledge, such as pathways, most methods treat genes in a specific pathway as an exchangeable set without considering the topological structure of a pathway. However, how genes are related with each other in a pathway may be very informative to identify association signals. To make use of the connectivity information among genes in a pathway in GWAS analysis, we propose a Markov Random Field (MRF) model to incorporate pathway topology for association analysis. We show that the conditional distribution of our MRF model takes on a simple logistic regression form, and we propose an iterated conditional modes algorithm as well as a decision theoretic approach for statistical inference of each gene's association with disease. Simulation studies show that our proposed framework is more effective to identify genes associated with disease than a single gene–based method. We also illustrate the usefulness of our approach through its applications to a real data example.

Citation: Chen M, Cho J, Zhao H (2011) Incorporating Biological Pathways via a Markov Random Field Model in Genome-Wide Association Studies. PLoS Genet 7(4): e1001353. doi:10.1371/journal.pgen.1001353

Editor: David B. Allison, University of Alabama at Birmingham, United States of America

Received June 17, 2010; **Accepted** February 24, 2011; **Published** April 7, 2011

Funding: This work was supported in part by NIH grants R01 GM59507, U01 DK62422, R01 DK72373, and U01 RR24139 and NSF grant DMS-0714817. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: hongyu.zhao@yale.edu

Introduction

In genome-wide association studies (GWAS) researchers examine a large number of markers across the genome in many individuals to identify associations between genetic variants and disease, or to prioritize markers for follow up studies. However, most of the times the signals from individual markers are weak and the sample size is not large enough to have adequate power for true discoveries, especially when the minor allele frequency is low. Various approaches have been developed to increase statistical power, including aggregating multiple markers from the same gene or in the same haplotype block region and incorporating information from other sources into the GWAS analysis. It has been found that the gene level analysis has the ability to identify new associations in addition to those identified using individual Single Nucleotide Polymorphisms (SNPs) [1,2]. Gene-based analyses include those using the most significant SNP within and near a gene [1]; combination statistics (Fisher, Sidat, and Simes) from all individual markers [2]; Principal Component Analysis (PCA) regressions [3] and the sparse partial least squares regressions [4]. To incorporate prior biological knowledge, one information rich resource is biological pathways. It is believed that genes interact with each other in biological processes, and it is conceivable that they may jointly affect the risk of a complex disease. There exist an abundance of databases containing known gene pathways and protein-protein interactions, such as KEGG, BioCarta, GenMAPP, and HPRD. A number of gene prioritization methods incorporating prior biological knowledge have been developed for GWAS. Some examples include Prioritizer [5], Endeavour [6], CGI [7], CANDID [8], GeneWanderer [9], CIPHER [10], GIN [11], and the pathway based gene set enrichment approach [1]. These methods have shown that incorporating prior biological information in GWAS is useful. However, they do not consider functional relationships among genes. The general input of these approaches is a list of genes as a set, in which genes are treated as exchangeable without taking into account the regulatory relationships among them. As a result, information from the pathway topology and interactions among genes is usually ignored. However, how genes are functionally related to each other in a pathway may be very informative for GWAS analysis and such information can be utilized to increase the power of detecting real associations. When associations have been firmly established for some genes either through GWAS or prior candidate gene-based studies, we can take advantage of this knowledge to examine other genes related to these known genes through the same pathways they all participate in.

In this paper we propose a Markov Random Field (MRF) model to incorporate biological pathway information in GWAS. MRF

Author Summary

Statistical methods used in most GWAS are based on the analysis of single markers. Prior biological information about markers, genes, and pathways is not commonly incorporated in the detection of associated disease loci. Recently a number of methods have been developed to incorporate such information, and it has been shown that they may make use of prior biological knowledge in association analysis. However, most of these methods ignore the regulatory relationships and functional interactions among genes. In this article, we propose a statistical method that can explicitly model the interactions of genes in a neighborhood defined by the topology of a pathway. Simulation studies and a real data example show that the proposed method can improve the power of identifying associated genes when they are in the neighborhood of other genes whose association has been firmly established in previous studies.

has been considered by several authors to combine data from different sources in genomics studies, e.g., a spatial normal mixture model [12] for gene expression and CHIP-chip data, a Gamma-Gamma model and MRF for mRNA microarray data [13], and prioritizing genes by combining gene expression and protein interaction data [7]. However, little has been done in the context for GWAS, with the exception of Li et al. [14] who proposed a hidden MRF for GWAS. But their method is developed in the context of jointly analyzing markers in linkage disequilibrium.

We first present a motivating example from a GWAS of Crohn's disease [15] for the proposed method. As will be shown next, the result clearly suggests that genes in the same neighborhood within a pathway tend to show similar association status. This Crohn's disease cohort includes 401 cases and 433 controls, and the Illumina HumanHap300 BeadChip (Illumina, San Diego) were used for genotyping. We first mapped SNPs to genes and then applied PCA regressions to obtain gene-level p values of the association tests with Crohn's disease status [3]. More details about this data set are given in the Materials and Methods section. We then obtained pathway and interaction from BioCarta (http://www.biocarta.com/), GeneMAPP [16] and KEGG [17]. We consider a total of 3,735 genes in over 350 pathways. Genes on the same chromosome that are within 1 million base pairs are excluded to avoid effects caused by possible linkage disequilibrium. To see whether genes connected with each other in the same pathway tend to show similar evidence for association, we use a cut-off value 0.15 where genes whose p values are below this cut-off are considered interesting and labeled with 1. Note that we use a relatively loose threshold so that a sufficiently large number of genes are called "interesting" and this loose cut-off also reflects our belief that many genes have weak effects and only show moderate evidence of association. In a pathway k, we consider the number of edges connecting a pair of "interesting" genes, which depends on the labels of all genes. We denote this number by D_k. A large value of D_k would suggest that "interesting" genes are more likely to be neighboring genes. To assess the statistical evidence for the tendency to observe large D_k values, we employ a permutation procedure as follows. In each permutation, we randomly permute the "interesting" labels of all genes and derive a permuted statistic and these permuted statistics are used to arrive at an empirical distribution of D_k under the null hypothesis that there is no tendency for neighboring genes to have similar disease association status, i.e. "interesting" or not. We then compare the observed D_k statistic with the empirical distribution. Finally the p value of the observed D_k in this empirical distribution is calculated. A p value close to 0 indicates that "interesting" genes tend to be neighbors. This procedure is repeated for all pathways, and the histogram of p values of D_k for all pathways is plotted in Figure 1. It is evident that this distribution is highly skewed to the left, which suggests associated genes tend to be neighbors in a given pathway.

In the rest of this article, we first introduce our model and statistical inferential procedures. The performance of our methods is then assessed through both simulation studies and real data applications.

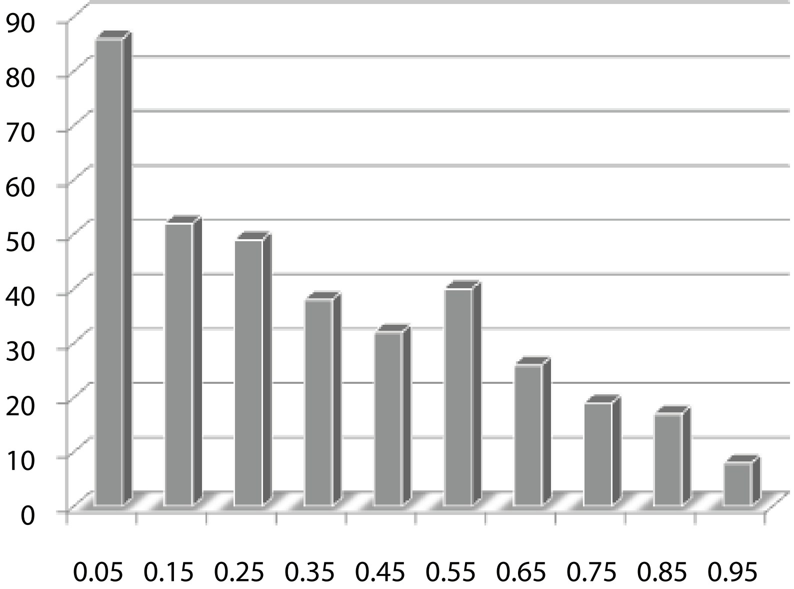

Figure 1. Histogram of p values of D_k, the number of edges connecting a pair of "interesting" genes in a pathway k, which depends on the labels of all genes. A large value of D_k would suggest that "interesting" genes are more likely to be neighbors. A permutation procedure is used to derive an empirical distribution of D_k under the null hypothesis. The p value of an observed D_k is calculated with respect to this empirical distribution. See the Introduction section for more details.
doi:10.1371/journal.pgen.1001353.g001

Results

A MRF model of gene pathways

We start by considering a simple model in which a pathway is represented by an undirected graph $G=(\mathcal{V},\mathcal{E})$ where $\mathcal{V}=\{1,\cdots,n\}$ is a set of n genes (nodes) and $\mathcal{E}=\{<i,j>: i \text{ and } j$ are directly connected$\}$ denotes the set of all edges. For the ith gene in $\mathcal{V}$, let $N_i=\{j: <i,j>\in\mathcal{E}\}$ denote the set of its neighbors, and $d_i=|N_i|$ denote the number of its neighbors. Let S_i denote the true association status where

$$S_i=+1 \text{ if gene } i \text{ is associated with the disease,}$$

$$S_i=-1 \text{ if gene } i \text{ is NOT associated with the disease.}$$

The values ± 1 are referred to as labels of a node hereafter. Let $\boldsymbol{S}=(S_1,\cdots,S_n)$ denote the labeling of $\mathcal{V}$. Thus $\boldsymbol{S}$ is a spatial random vector whose elements may be correlated with each other. Note that each node can be labeled either -1 or $+1$, and so there are a total of 2^n unique configurations of the pathway. The ultimate goal is to infer the value of S_i based on the pathway topology and the observed association data.

To formalize the idea that neighboring genes tend to have similar association status, we need a probability measure so that nodes connected with each other tend to have the same labels. Here we consider a nearest neighbor Gibbs measure [18] that has the following form:

$$Pr(\boldsymbol{S}|\boldsymbol{\theta}_0)=\frac{1}{z(\boldsymbol{\theta}_0)}\exp\left\{h\sum_{i\in\mathcal{V}}I_1(S_i)+\tau_0\sum_{<i,j>\in\mathcal{E}}(w_i+w_j)I_{-1}(S_i)I_{-1}(S_j)+\tau_1\sum_{<i,j>\in\mathcal{E}}(w_i+w_j)I_1(S_i)I_1(S_j)\right\}, \tag{1}$$

where $\boldsymbol{\theta}_0=(h,\tau_1,\tau_0)$ are the prior parameters or hyperparameters, $I_1(\cdot)$ and $I_{-1}(\cdot)$ are the indicator functions, $w_i=d_i^{1/2}$, and $z(\boldsymbol{\theta}_0)$ is a normalizing function that is the sum over all 2^n possible configurations:

$$z(\boldsymbol{\theta}_0)=\sum_{\boldsymbol{S}}\exp\left\{h\sum_{i\in\mathcal{V}}I_1(S_i)+\tau_0\sum_{<i,j>\in\mathcal{E}}(w_i+w_j)I_{-1}(S_i)I_{-1}(S_j)+\tau_1\sum_{<i,j>\in\mathcal{E}}(w_i+w_j)I_1(S_i)I_1(S_j)\right\}. \tag{2}$$

Note that it is prohibitive to evaluate $z(\boldsymbol{\theta}_0)$ when n is large. Here τ_0 and τ_1 assign prior weights to edges connecting two non-associated nodes and two associated nodes, respectively. The function w_i will be elaborated in more details in the context of the conditional probability later.

In (1), the second sum is taken over all edges connecting direct neighbors in which both end nodes are labeled -1, and the third sum is taken over all edges in which both end nodes are labeled 1. Positive τ_0 and τ_1 will put more weights on configurations in which directly linked nodes have the same labels, which is desirable in our context. The hyperparameter h determines the marginal probability of S_i when $\tau_0=\tau_1=0$, i.e., all nodes are treated as singletons that are independent:

$$Pr(S_i=1|h,\tau_0=\tau_1=0)=\frac{\exp(h)}{\exp(h)+1}.$$

The simple form Gibbs measure in (1) has the Markov property that makes it attractive to model a biological pathway, in which directly linked genes interact with each other. It defines a MRF, which by definition is a probability measure that satisfies $Pr(S_i|S_{\mathcal{V}-i})=Pr(S_i|S_{N_i})$, where $\mathcal{V}-i$ denote all nodes but i, and N_i is the set of all direct neighbors of node i. Please see Materials and Methods for details.

Posterior distribution

Now we discuss the posterior distribution of association status after combining the evidence from the observed association statistics at the gene level and the structure of the gene pathway. Before we proceed, it is necessary to present the likelihood function of the observed data. We consider the situation where the observed evidence of association is summarized by p values, which are assumed to be conditionally independent given the true association status $\boldsymbol{S}$. Under the null hypothesis of no association, each p-value has a uniform (0,1) distribution. In this article, we consider $\boldsymbol{y}=(y_1,\cdots,y_n)$, where $y_i=\Phi^{-1}(1-p_i/2)$ and $\Phi(\cdot)$ is the CDF (Cumulative Distribution Function) of $N(0,1)$. Therefore, under the null hypothesis of no association, i.e., $S_i=-1$, the density of y_i is $f_0(y_i)\sim N(0,1)$. However, if there is association between the gene and disease, i.e., $S_i=+1$, the distribution of y_i is usually unknown. For simplicity, we assume that it is from $N(\mu_i,\sigma_i^2)$, where μ_i is the location parameter and σ_i is the scale parameter that usually depends on the true effect size, allele frequencies, and the sample size. To account for the uncertainty about the parameters, we can put prior distributions on μ_i and σ_i^2, and marginalize over them to obtain the predictive density of y_i. Here we consider conjugate priors $\mu_i|\sigma_i^2\sim N(\bar{\mu},\sigma_i^2/a)$ and $\sigma_i^2\sim$ Inverse Gamma$(\nu/2,\nu d/2)$, or $\nu d/\sigma_i^2\sim\chi_\nu^2$. We denote $\boldsymbol{\theta}_1=(\bar{\mu},a,\nu,d)$ that are hyperparameters. The prior mean of μ_i is $\bar{\mu}$ and its variance is σ_i^2/a. The prior mean of σ_i^{-2} is d^{-1} and the prior variance is $Var[\sigma_i^{-2}]=2/(\nu d^2)$. This prior is of conjugate form so that the integration over μ_i and σ_i^2 is analytically tractable. We note that the hyperparameters can be estimated from the observed data via an empirical Bayes method (see Text S2, Figures S1 and S2). Under this prior setting, the marginal density of y_i is

$$f_1(y_i|S_i=1,\boldsymbol{\theta}_1)=\pi^{-1/2}(\nu d)^{\nu/2}\frac{\sqrt{a}}{\sqrt{a+1}}\frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)}\left(\frac{a}{a+1}(y_i-\bar{\mu})^2+\nu d\right)^{-(1+\nu)/2}.$$

This is equivalent to $(y_i-\bar{\mu})/\sqrt{(a+1)d/a}\sim t(\nu)$ when $\nu=1$, 2, and others.

The joint marginal density of $\boldsymbol{y}$ is

$$f(\boldsymbol{y}|\boldsymbol{S},\boldsymbol{\theta}_1)=\prod_{\{j:S_j=-1\}}f_0(y_j)\times\prod_{\{j:S_j=+1\}}f_1(y_i|\boldsymbol{\theta}_1).$$

Thus, the posterior distribution of $\boldsymbol{S}$ given the observed data $\mathbf{y}$ is

$$Pr(\boldsymbol{S}|\boldsymbol{y},\boldsymbol{\theta}_0,\boldsymbol{\theta}_1) \propto f(\boldsymbol{y}|\boldsymbol{S},\boldsymbol{\theta}_1)Pr(\boldsymbol{S}|\boldsymbol{\theta}_0). \tag{3}$$

Similar to the MRF interpretation of the prior distribution (1), the posterior also has a nice conditional distribution and is actually a MRF, as will be shown in the Materials and Methods section.

When n is large, since it is prohibitive to evaluate posterior probabilities on the entire space of configurations, we implement a Markov chain Monte Carlo (MCMC) method to sample from the posterior distribution. Naturally a Gibbs sampler is well suited for a MRF. As will be shown later, due to the MRF property, the posterior has a nice closed-form conditional distribution that can greatly facilitate the MCMC.

Making inference based on the posterior distribution

Most GWAS lead to a set of candidate genes/SNPs that will need to be validated in follow-up studies. Therefore, it is important to include as many truly associated genes as possible among the top ranked genes. Our proposed method allows us to rank order genes as detailed below.

There are several ways of inferring the labels according to the posterior distribution of $\boldsymbol{S}$. The first one is to use maximum *a posteriori* (MAP) estimate, which is the configuration with the largest posterior probability, a reasonable point estimate for $\boldsymbol{S}$. Let us denote it by $\hat{\boldsymbol{s}}^A = (\hat{s}_1^A, \cdots, \hat{s}_n^A)$. The MAP is the maximizer of the joint posterior distribution:

$$\hat{\boldsymbol{s}}^A = \arg\max_{\boldsymbol{s}} f(\boldsymbol{y}|\boldsymbol{s},\boldsymbol{\theta}_1)Pr(\boldsymbol{s}|\boldsymbol{\theta}_0).$$

A Gibbs sampler outlined above can be applied to stochastically search for the solution to the above optimization problem. Multiple restarts with different initial configurations are recommended. An alternative approach is to base the estimate on the posterior conditional probability of S_i given the observed data and all the other nodes $\boldsymbol{s}_{\mathcal{V}-i}$. We can estimate s_i by maximizing this conditional probability (MCP):

$$\hat{s}_i^C = \arg\max_{s_i} f(y_i|s_i,\boldsymbol{\theta}_1)Pr(s_i|s_{N_i},\boldsymbol{\theta}_0). \tag{4}$$

The advantage of this approach is that the above problem is trivial to solve. As will be explained in equation (8) of the Materials and Methods section, the second term in formula (4) can be evaluated in closed form. Besag [19] proposed an algorithm known as iterated conditional modes (ICM) that iteratively updates s_i. Note that the convergence of ICM is assured because the posterior is proportional to

$$Pr(\boldsymbol{y}|\boldsymbol{s},\boldsymbol{\theta}_0,\boldsymbol{\theta}_1) = Pr(s_i|\boldsymbol{y},s_{\mathcal{V}-i},\boldsymbol{\theta}_0,\boldsymbol{\theta}_1)Pr(s_{\mathcal{V}-i}|\boldsymbol{y},\boldsymbol{\theta}_0,\boldsymbol{\theta}_1),$$

which never decreases at any iteration because the first term is non-decreasing and the second one is a constant. So it is easy to see the ICM will converge to a local maximum in the posterior distribution. Since the ICM runs fast and usually converges in several iterations, multiple restarts with different initial configurations are recommended. Finally the resulting configurations can be compared by evaluating $f(\boldsymbol{y}|\hat{\boldsymbol{s}}^C,\boldsymbol{\theta}_1)Pr(\hat{\boldsymbol{s}}^C|\boldsymbol{\theta}_0)$ up to a normalizing constant to pick the largest one.

The inference can also be based on the marginal posterior probability. Let $m_i = Pr(S_1 = 1|\boldsymbol{y})$. We consider a decision rule in the form $\delta(m_i) = I(m_i \geq m^*)$, where $I(\cdot)$ is an indicator function and m^* is the sought decision threshold. If $\delta(m_i) = 1$, the decision is positive (also referred to as discovery) and gene i is called to be associated with the disease. Likewise if $\delta(m_i) = 0$ the decision is negative. To address the problem of multiple comparisons, we consider loss functions associated with making wrong decisions (false discoveries and false negatives), and solve the decision problem by minimizing the expectation of the loss functions under the posterior distribution. Here we consider two loss functions. First, if we are interested in the 0-1 loss function $L_1(\boldsymbol{S},\delta) = \sum_{i=1}^n |I_1(S_i) - \delta(m_i)|$, we may want to minimize the expected loss

$$\begin{aligned} m_1^* &= \arg\min_{m^*} E\{L_1(\boldsymbol{S},\delta)|\boldsymbol{y},\boldsymbol{\theta}_0,\boldsymbol{\theta}_1\} \\ &= \arg\min_{m^*} \sum_{\boldsymbol{S}} \left\{ \sum_{i=1}^n |I_1(S_i) - \delta(m_i)| \right\} \cdot Pr(\boldsymbol{S}|\boldsymbol{y},\boldsymbol{\theta}_0,\boldsymbol{\theta}_1), \end{aligned} \tag{5}$$

under the posterior distribution of $\boldsymbol{S}$. The solution is $m_1^* = 0.5$. Note that L_1 assigns equal loss to the false positive and false negative errors. This is to minimize the expected frequency of making wrong calls for the association status. Note that the performance of the decision rule δ is based on the frequentist operating characteristic in the Bayesian framework, which is common in medical decision makings [20]. The second loss function we consider is the false discovery rate (FDR):

$$FDR = L_2(\boldsymbol{S},\delta) = \frac{\sum_i \delta(m_i)I_{-1}(S_i)}{\sum_i \delta(m_i)}. \tag{6}$$

Suppose the goal is to control the expected FDR, under the posterior distribution, such that it is no more than α, i.e., $E\{L_2(\boldsymbol{S},\delta)|\boldsymbol{y},\boldsymbol{\theta}_0,\boldsymbol{\theta}_1\} \leq \alpha$. If we rank order all genes by their posterior probabilities from the largest to the smallest, and let $m_{(i)}$ denote the ith order statistics, then the solution is to choose a cut off value $m_2^* = m_{(j)}$ where j is the largest integer that makes $j^{-1}\sum_{i=1}^j m_{(i)} \geq (1-\alpha)$. We should mention that more complicated loss functions can be considered under the framework of our model. See Müller et al. [20] for other examples.

Simulation studies

First we use simulated data to study the performance of the proposed method. The simulation is based on a simple 6-node network shown in Figure 2. Genes G1 through G3 are assumed to be associated with the disease (labeled +1) while G4 through G6 are not associated with the disease (labeled -1). Data are simulated from a disease model as follows. We assume G1, G2 and G3 have independent effects on disease risk and each has a disease related SNP. The genotypes and minor allele frequencies of these three SNPs are denoted by (g_1,g_2,g_3) and (p_1,p_2,p_3), respectively, where $g_i \in \{0,1,2\}$ for $i = 1,2,3$. A multiplicative genetic model is assumed for the risk of having the disease. More specifically, for an individual with genotype (g_1,g_2,g_3), the risk is $r_0 r_1^{g_1} r_2^{g_2} r_3^{g_3}$, where r_0 is the baseline risk of those carrying two normal alleles in all three genes, and r_i is the relative risk, or effect size, of gene i, $i \in \{1,2,3\}$. For each SNP the Hardy-Weinberg equilibrium (HWE) is assumed to hold in the general population so that the genotype probabilities are $(1-p_i)^2$, $2p_i(1-p_i)$, and p_i^2 for $g_i = 0$, 1, and 2, respectively. In the simulation we use three minor allele frequencies $p = (0.05,\ 0.10,\ 0.15)$, three disease prevalence values $k = (0.05,\ 0.10,\ 0.15)$, and six effect sizes $r = (1.05,\ 1.10,\ 1.15,\ 1.20,\ 1.25,\ 1.30)$. As a result, there are a total of 54 settings of (p,k,r), for each of which we first let $p_1 = p_2 = p_3 = p$ and

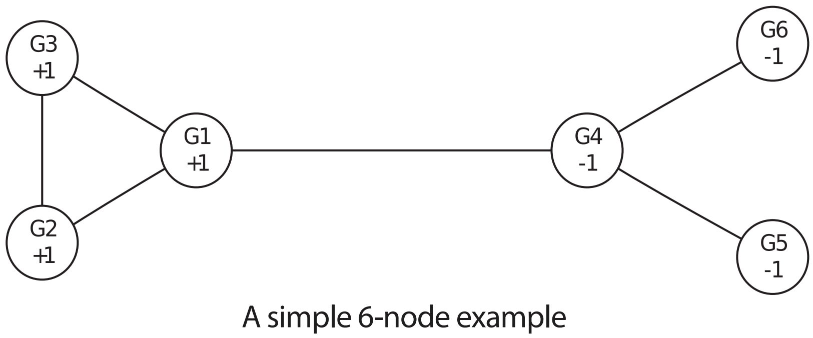

Figure 2. A simple 6-node network.
doi:10.1371/journal.pgen.1001353.g002

$r_1 = r_2 = r_3 = r$, and then calculate the baseline risk r_0, and finally obtain the conditional distribution of the genotypes (g_1, g_2, g_3) given the disease status. Then genotypes of G1, G2 and G3 of 600 cases and 600 controls are simulated according to the conditional genotype distribution. The p values of the three causal genes are calculated from a logistic regression of the data. For G4 through G6, the p values are simulated from Uniform(0,1). The power of detecting the true association depends on the disease model. In this case, larger values of relative risk, MAF and prevalence corresponds to association tests with higher power.

In the simulation we set the hyperparameters $(h, \tau_1, \tau_0) = (-1, 0.25, 0.01)$ where more weights are assigned to edges connecting two associated genes. This corresponds to a prior belief that the probability of association is roughly between 0.35 and 0.50. The hyperparameters $(\bar{\mu}, a, v, d)$ are set to (3, 1, 10, 1) where a large value of v puts a large prior variance on σ_i^2, which allows a wide range of values for both μ_i and σ_i^2. For each simulated data set, the posterior probabilities are enumerated since there are only 64 possible configurations in this simple example. The simulation is repeated 500 times. We compare the proposed method using the posterior mean with the one using the p value, and apply cut-off values of 0.7 and 0.05 for posterior probabilities and p values, respectively. For each simulated data set, we calculate the false positive rate (FPR), sensitivity (Sens.), and false discovery rate (FDR) by thresholding on p values and posterior probabilities. In addition, genes can be rank ordered by the two methods and the area under the Receiver Operating Characteristic curve (AUC) can be calculated. The average values of the three rates plus the AUC over the 500 simulated data sets are shown in Table 1. As can be seen, the proposed method of the posterior probability has higher sensitivity, smaller false discovery rate, and higher AUC than the p value thresholding in every setting of the prevalence, MAF and effect size, while the FPR of both methods are controlled at 0.05.

The second simulation study is based on the network shown in Figure 3. This network was adapted from BioCarta "Human Rho cell motility signaling pathway" and we deleted a few genes that are either absent from our Crohn's disease data or not connected to others. We assume three different sets of truly associated genes, plotted in triangles, rectangles and pentagons, each of which contains three, five, and seven nodes, respectively. To simulate different levels in the power of the association tests, for each gene with $S_i = +1$, the p value is computed from a two-sided z test where z scores are randomly drawn from $N(1,1)$, $N(1.5,1)$ and $N(2,1)$, respectively, corresponding to the power 0.16 (low), 0.32 (median) and 0.51 (high) in the association tests. The p values for $S_i = -1$ are generated randomly from Uniform(0, 1) as before.

To examine the effects of hyperparameters of the network, we consider eight priors, listed in Table 2, that roughly form four main groups indexed by numbers 1 through 4, and two subgroups indexed by letters a and b. For each set of hyperparameters a Gibbs sampler is run to draw samples from the corresponding prior distribution, and we can estimate $Pr(S_i = 1)$, the prior mean, and $Pr(S_i = S_j = 1)$ and $Pr(S_i = S_j = -1)$ where $<i,j> \in \mathcal{E}$, the probabilities of edge $<i,j>$ linking two nodes with identical labels. The averages of the estimated probabilities are listed in the last three columns of Table 2. The average prior means of all nodes are about 0.05, 0.15, 0.25, and 0.4, respectively for the four main groups. Roughly speaking, it means that group 1 is in favor of a small number, and group 4 a large number, while groups 2 and 3 in between, of nodes labeled with +1. Furthermore, values of (τ_0, τ_1) in subgroup b are larger than those in subgroup a, meaning that nodes with identical labels are more likely to be next to each other apriori in subgroup b than subgroup a, as can be seen from the last two columns in Table 2. On the other hand, because the posteriors are found to be insensitive to the hyperparameters $(\bar{\mu}, a, v, d)$ when v is large, they are set to (3, 1, 10, 1) as in the previous example.

We simulate 200 data sets for each combination of the three power settings (low, median and high) and three truly associated sets (3, 5, and 7 nodes). For each data set, we run eight Gibbs samplers using eight different hyperparameters described above. Each Gibbs sampler is run with 100 restarts and each start contains 100 steps. We compare the average AUC of 200 simulated data sets using p value and the posterior mean and plot the results in Figure 4. In general, the AUC of the proposed method is larger than that using p values alone. It achieves good AUC if the prior mean is close to the truth, especially when the power is low. For example, in the middle column panels where there are 5 truly associated genes, prior settings 2 and 3, favoring median number of truly associated nodes, outperform prior settings 1 and 4. Similarly, in the right panel where the true model contains 7 genes, prior settings 3 and 4, which are in favor of large models, perform better than the other prior settings. Furthermore, priors in subgroup b are better than subgroup a in general. It is not surprising because the priors in subgroup b encourages nodes labeled with +1 to group together, which agrees with the simulation setting.

To evaluate the control of the false positive rates and the false discovery rates of the proposed methods in relatively large pathways with only a few associated genes, we conduct a third simulation study based on a simulated network shown in Figure 5 that contains 60 nodes. We consider three truly associated gene sets, namely (2, 11, 19), (2, 11, 19, 41), and (2, 11, 19, 20, 41), and label them as models 1, 2 and 3 in Table 3. Similar to the previous study, we simulate p values from z scores randomly drawn from $N(1,1)$, $N(1.5,1)$ and $N(2,1)$, corresponding to weak, median and strong associations, respectively. Three prior settings are considered for (h, τ_1, τ_0), namely $(-1.5, 0.15, 0.02)$, $(-1.50, 0.10, 0.01)$ and $(-2,$

Table 1. Average FPR, sensitivities, FDR, and AUC of the 6-node network.

Prevelance	Effect Size		MAF = 0.05				MAF = 0.10				MAF = 0.15			
		Method	FPR	Sens.	FDR	AUC	FPR	Sens.	FDR	AUC	FPR	Sens.	FDR	AUC
0.05	1.05	*p* value	0.05	0.06	0.42	0.50	0.05	0.07	0.39	0.52	0.04	0.08	0.35	0.54
		Posterior	0.04	0.07	0.39	0.55	0.04	0.08	0.35	0.56	0.04	0.08	0.33	0.59
	1.10	*p* value	0.05	0.08	0.38	0.55	0.05	0.12	0.25	0.60	0.05	0.13	0.26	0.60
		Posterior	0.05	0.08	0.38	0.60	0.04	0.13	0.22	0.65	0.04	0.16	0.21	0.67
	1.15	*p* value	0.05	0.13	0.25	0.60	0.04	0.20	0.18	0.67	0.06	0.28	0.17	0.72
		Posterior	0.04	0.15	0.21	0.67	0.04	0.23	0.16	0.74	0.05	0.33	0.13	0.80
	1.20	*p* value	0.05	0.18	0.19	0.64	0.04	0.32	0.10	0.76	0.05	0.40	0.11	0.81
		Posterior	0.05	0.20	0.17	0.71	0.04	0.38	0.09	0.84	0.05	0.49	0.09	0.88
	1.25	*p* value	0.05	0.25	0.16	0.70	0.06	0.41	0.12	0.80	0.05	0.56	0.07	0.88
		Posterior	0.05	0.29	0.14	0.77	0.06	0.51	0.10	0.87	0.06	0.68	0.06	0.93
	1.30	*p* value	0.05	0.34	0.11	0.76	0.05	0.57	0.07	0.88	0.05	0.72	0.05	0.92
		Posterior	0.05	0.41	0.10	0.83	0.06	0.68	0.06	0.93	0.05	0.83	0.05	0.96
0.1	1.05	*p* value	0.05	0.06	0.43	0.54	0.05	0.08	0.39	0.52	0.05	0.07	0.41	0.53
		Posterior	0.05	0.07	0.42	0.58	0.05	0.08	0.39	0.58	0.05	0.07	0.38	0.59
	1.10	*p* value	0.05	0.09	0.34	0.57	0.04	0.13	0.24	0.60	0.05	0.16	0.25	0.62
		Posterior	0.04	0.09	0.31	0.62	0.04	0.15	0.20	0.66	0.05	0.18	0.21	0.70
	1.15	*p* value	0.04	0.14	0.22	0.62	0.05	0.21	0.18	0.67	0.05	0.27	0.14	0.72
		Posterior	0.04	0.15	0.20	0.69	0.05	0.24	0.16	0.73	0.05	0.33	0.11	0.80
	1.20	*p* value	0.04	0.21	0.16	0.68	0.05	0.33	0.12	0.77	0.05	0.45	0.09	0.82
		Posterior	0.04	0.24	0.14	0.76	0.06	0.40	0.11	0.84	0.06	0.54	0.08	0.89
	1.25	*p* value	0.05	0.27	0.13	0.73	0.05	0.48	0.08	0.84	0.04	0.62	0.05	0.91
		Posterior	0.04	0.33	0.11	0.80	0.05	0.57	0.06	0.90	0.05	0.73	0.05	0.95
	1.30	*p* value	0.05	0.37	0.12	0.79	0.05	0.61	0.05	0.90	0.05	0.78	0.04	0.94
		Posterior	0.05	0.46	0.09	0.87	0.05	0.73	0.05	0.94	0.05	0.87	0.04	0.97
0.2	1.05	*p* value	0.06	0.06	0.50	0.52	0.05	0.08	0.39	0.54	0.06	0.09	0.39	0.55
		Posterior	0.05	0.06	0.49	0.55	0.04	0.08	0.33	0.59	0.05	0.09	0.36	0.59
	1.10	*p* value	0.06	0.09	0.39	0.53	0.05	0.13	0.26	0.60	0.06	0.17	0.24	0.64
		Posterior	0.05	0.09	0.36	0.58	0.04	0.15	0.22	0.66	0.04	0.20	0.18	0.71
	1.15	*p* value	0.05	0.15	0.24	0.63	0.04	0.24	0.14	0.71	0.04	0.37	0.10	0.78
		Posterior	0.05	0.18	0.20	0.69	0.05	0.28	0.13	0.79	0.05	0.44	0.10	0.85
	1.20	*p* value	0.05	0.24	0.18	0.70	0.05	0.43	0.10	0.81	0.04	0.55	0.06	0.89
		Posterior	0.05	0.28	0.15	0.78	0.05	0.52	0.08	0.88	0.05	0.67	0.06	0.93
	1.25	*p* value	0.05	0.35	0.12	0.78	0.05	0.57	0.06	0.88	0.04	0.73	0.04	0.94
		Posterior	0.05	0.43	0.09	0.84	0.05	0.68	0.05	0.94	0.05	0.82	0.04	0.97
	1.30	*p* value	0.05	0.46	0.08	0.83	0.05	0.73	0.05	0.93	0.05	0.86	0.04	0.97
		Posterior	0.06	0.57	0.08	0.90	0.06	0.84	0.05	0.97	0.05	0.93	0.04	0.99

doi:10.1371/journal.pgen.1001353.t001

0.2, 0.01), whose average prior probability $Pr(S_i=1)$ is approximately 0.2, and average prior probabilities $Pr(S_i=S_j=1)$ for $<i,j>\in\mathcal{E}$ are roughly 0.13, 0.11, and 0.08, respectively. For the proposed method, we consider three decision rules. The first one (PM1) uses the posterior mean with a cut-off value $m_1^*=0.5$ as in (5), the second one is MCP as in (4), and the third one (PM2) is the method to control the FDR at 0.1 as in (6). Then we compare them with the *p* value method (P value) with a cut-off value set at 0.05 and the correction method (BH) of Benjamini & Hochberg (1995) [21]. For each scenario we simulate 100 data sets, and run a Gibbs sampler with 100 restarts where each start contains 100 iterations. For each simulated data set, we calculate the FPR, sensitivity (Sens.), FDR, and AUC as before. Table 3 lists the average values of the 100 simulation runs. In general PM1 and MCP control the FPR below the 0.05 level and have lower FDR than the *p* value while achieving better or similar power as the *p* value method. In terms of controlling FDR, PM2 controls the FDR around 0.1, and it has smaller FPR or better power than the BH method in most cases when it achieves similar or better FDR.

Crohn's disease data

We use one Crohn's disease [15] data set to further evaluate the performance of the proposed method. Details of this data can be found in the Materials and Methods section.

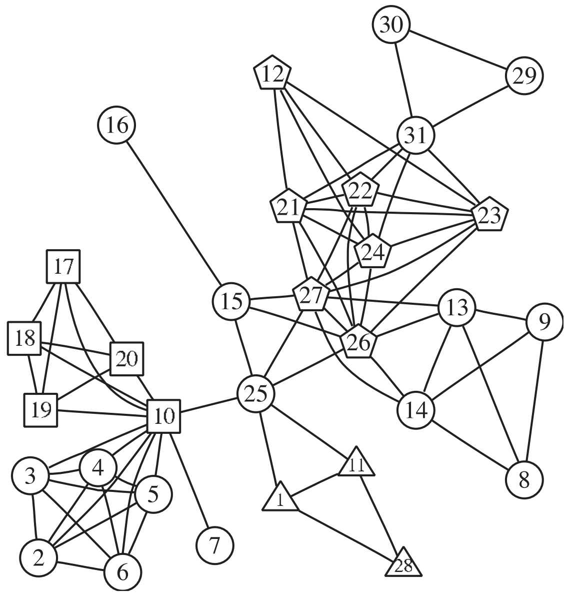

Figure 3. A 31-node network adapted from BioCarta "Human Rho cell motility signaling pathway." Triangles, rectangles and pentagons denote three different sets of truly associated genes, each of which contains three, five, and seven nodes, respectively.
doi:10.1371/journal.pgen.1001353.g003

We run our algorithm on 289 pathways that have at least 20 genes with non-missing p values. The hyperparameters (h,τ_1,τ_0) are chosen such that the average prior mean is roughly between 0.2 and 0.4 based on the simulation findings. To evaluate the performance, we consider 32 target genes that are confirmed to be related to the Crohn's disease [22]. Among these genes, 10 genes can be mapped to 66 pathways. In Figure 6 we plot the AUC values of the rankings by p values on the y axis and posterior means on the x axis for pathways containing three or more target genes. A majority of AUC values are improved if genes are rank ordered by the posterior mean. The average AUC based on p values is 0.568 while on posterior means is 0.613. To see what causes the rank changes of genes in the posterior probability, in Figure 7 we show the Human IL-2 Receptor Beta Chain in T cell Activation pathway from BioCarta. Genes in this pathway are densely connected. To aid visualization, we randomly remove some edges. Significant genes whose p values are below 0.05 are colored in cyan, genes with improved ranks are colored in light blue and others are colored in pink. It can be clearly seen that genes colored in light blue have more connections with the significant genes, and are more heavily linked among themselves, compared to other genes in the pathway. Genes that have many interactions with each other may play important roles in the biological processes in the pathway. When they are connected to many significant genes, it might be reasonable that they are more likely to associate with the disease than other genes.

Discussion

In this article we introduced a Bayesian method to incorporate prior knowledge of biological pathways into GWAS. This approach uses a MRF as a prior distribution to model the interactions among genes that participate in the same pathway. We showed that the posterior distribution is also a MRF and can be sampled via a Gibbs sampler. Inferences based on the posterior distribution allow us to combine data from the association study with prior information of biological pathways. In particular, this framework considers the topology of all genes in a pathway, which has not been fully utilized in many of the existing methods. The simulation studies and real data example suggest that the proposed method has higher power to identify genes associated with disease.

One limitation of the MRF model is that the Gibbs sampler tends to move around local maxima for a long time and thus can be slow in convergence to the posterior distribution. We recommend to run the Markov chain Monte Carlo with multiple random restarts, and examine the sampling distribution of network statistics, like the number of genes labeled with +1 and the proportion of edges linking genes with identical labels. In our studies, we found that a Markov chain initially moves very rapidly from its starting state, usually within the first 10 to 20 steps, before it reaches some steady states and stabilizes for a long period thereafter. We suggest running 100 Gibbs steps for each random starting state, and conducting the simulation with 100 restarts. The computing time of this scheme typically takes a few minutes on a PC for a pathway of about 30 genes. We should also mention that the characteristics of the MRF defined in (1) depend on both the hyperparameters and the structure of the network under consideration. Consequently there does not exist a set of hyperparameters that can be suitable for all pathways. To assist the specification of hyperparameters, we provide an algorithm of estimating hyperparameters based on a

Table 2. Eight priors.

		Hyperparameters			**Estimates**		
Group	**Subgroup**	h	τ_1	τ_0	$E[Pr(S_i=1)]$	$E[Pr(S_i=S_j=1)]$	$E[Pr(S_i=S_j=-1)]$
1	*a*	−3.00	0.10	0.01	0.044	0.003	0.917
	b	−3.00	0.25	0.10	0.049	0.043	0.923
2	*a*	−2.00	0.10	0.01	0.156	0.047	0.710
	b	−2.50	0.20	0.05	0.141	0.119	0.776
3	*a*	−1.25	0.05	0.01	0.250	0.081	0.563
	b	−3.00	0.25	0.05	0.254	0.264	0.602
4	*a*	−1.50	0.10	0.01	0.355	0.227	0.402
	b	−2.00	0.25	0.10	0.405	0.412	0.466

doi:10.1371/journal.pgen.1001353.t002

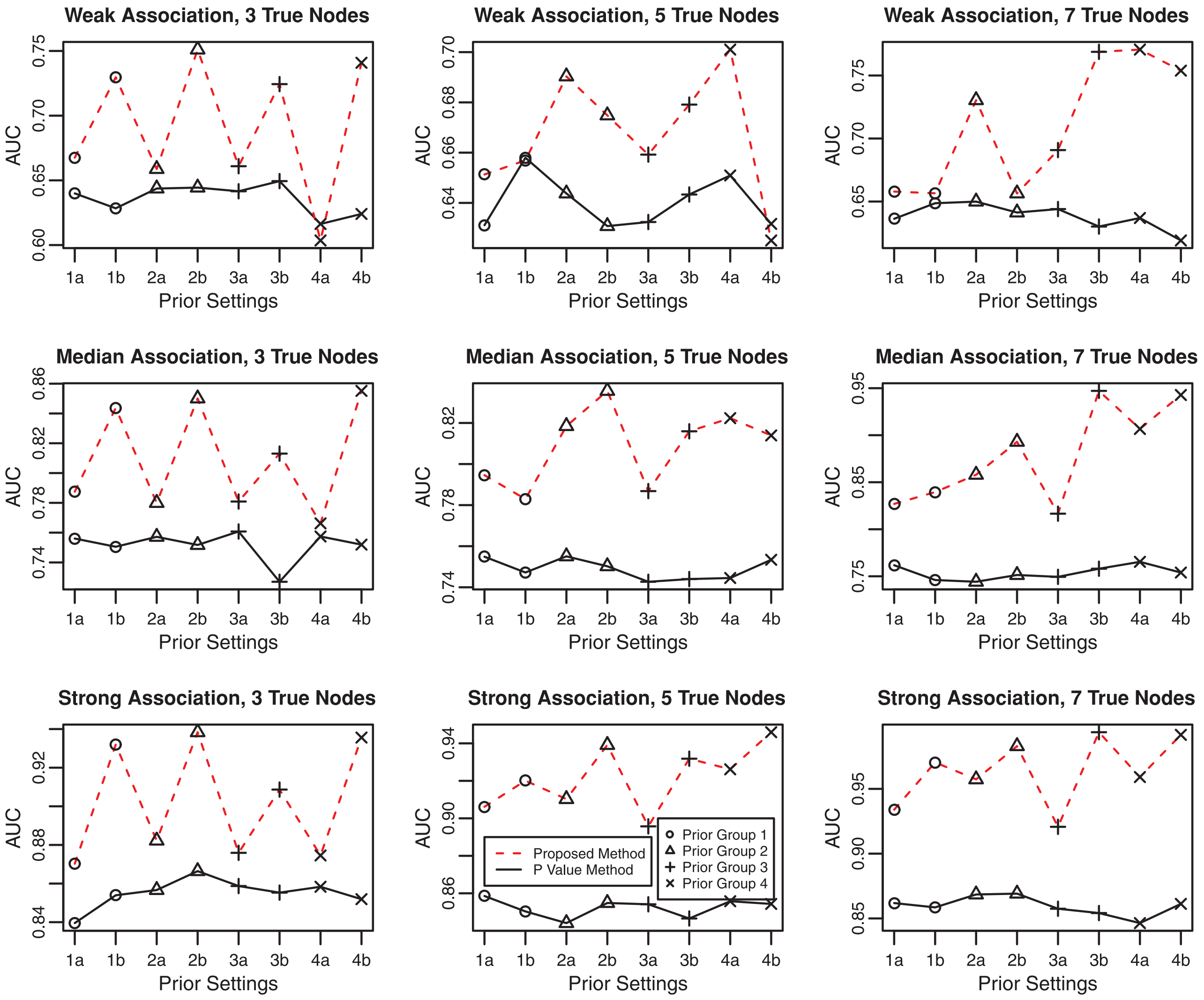

Figure 4. Comparison under different priors. The three rows of panels are for weak (top panel), median (middle panel), and strong (bottom panel) association signals. The three columns of panels are for three sets of truly associated genes corresponding to three (left), five (middle), and seven (right) nodes, respectively. The dotted lines link AUC of the proposed method and the solid lines connect AUC using p values. Circles, triangles, plus signs and crosses denote prior parameter groups 1, 2, 3, and 4, respectively.
doi:10.1371/journal.pgen.1001353.g004

conditional empirical Bayes approach in Text S2. It is recommended that these values would be used in initial attempts and it would be better to test several other variants of hyperparameters, possibly through fine-tuning the initial values. It is helpful to draw samples from the prior distribution to assess the effects o f different prior settings. One limitation of pathway-based analysis is that not all the genes can be associated with pathways. It is likely with knowledge accumulation, more genes will be mapped to pathways. An R package is under construction and will be made publically available soon.

Materials and Methods

The MRF property of the prior distribution on pathways

The nearest neighbor Gibbs measure on gene pathways in formula (1) defines a MRF and its conditional distribution has a logistic regression form as shown below.

Proposition 1. The Gibbs measure in (1) is Markovian and thus defines a MRF

$$Pr(S_i|S_{\mathcal{V}-i},\boldsymbol{\theta}_0)=Pr(S_i|S_{N_i},\boldsymbol{\theta}_0).$$

Moreover, the conditional distribution has a logistic form:

$$logit\,Pr(S_i|S_{N_i},\boldsymbol{\theta}_0)=h+\tau_1\left(w_iJ_i^{(1)}+\sum_{k\in N_i}w_kI_1(S_k)\right)-\tau_0\left(w_iJ_i^{(-1)}+\sum_{k\in N_i}w_kI_{-1}(S_k)\right),\ i=1,\cdots,n,\qquad(7)$$

where $J_i^{(l)}=\sum_{k\in N_i}I_l(S_k)$, $l=\pm1$. Equivalently, (7) can be rewritten as a system of linear equations:

$$\mathrm{logit}Pr(S_i|S_{N_i},\boldsymbol{\theta}_0)=\beta_{i0}+\beta_{i1}S_1+\cdots+\beta_{in}S_n,\ i=1,\cdots,n,\qquad(8)$$

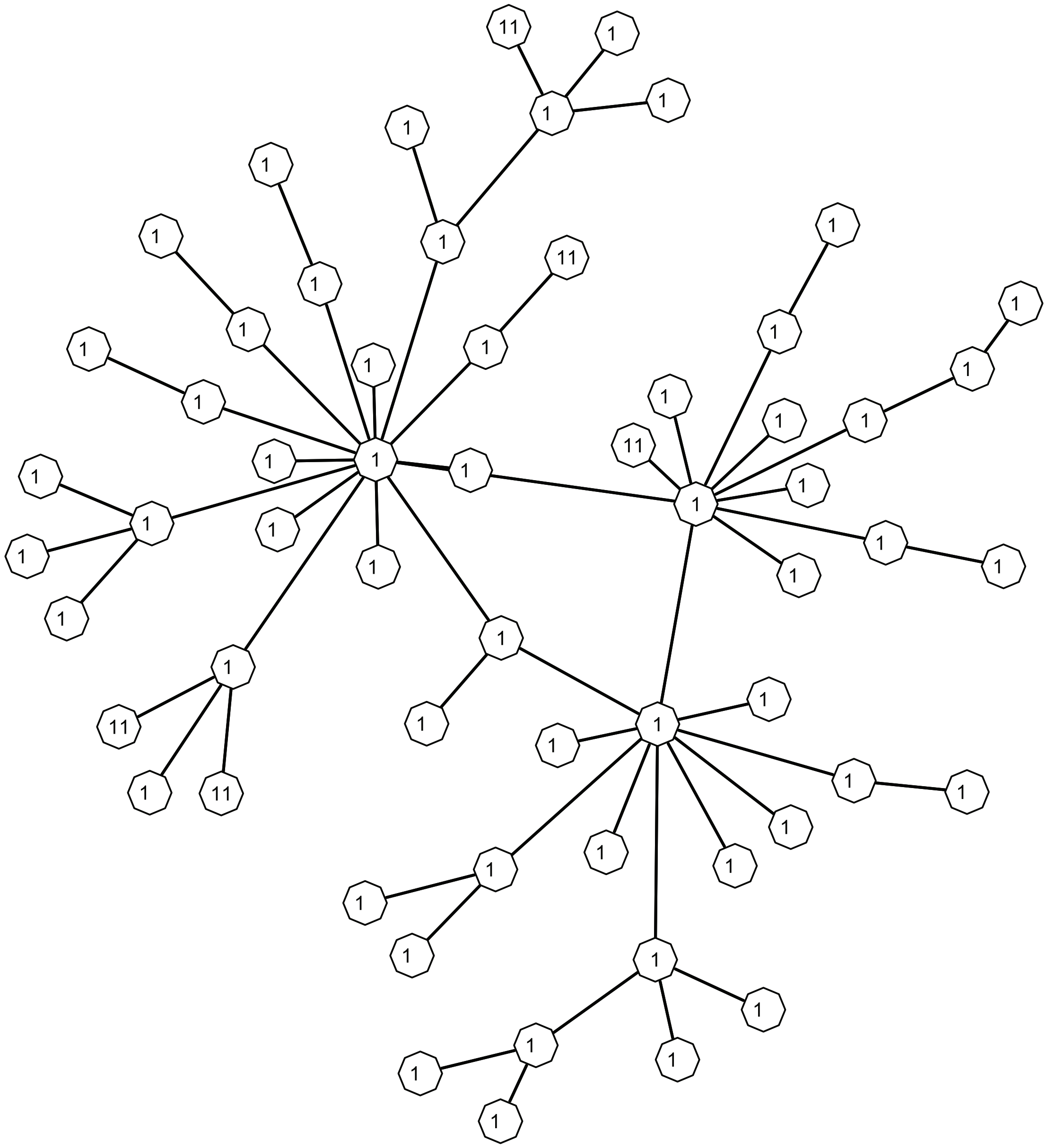

Figure 5. A simulated 60-node network.
doi:10.1371/journal.pgen.1001353.g005

where

$$\beta_{i0} = h,$$

$$\beta_{ij} = \begin{cases} 0 & \text{if } i = j \text{ or } <i,j> \notin \mathcal{E} \\ (w_i + w_j)\{\tau_1 I_1(S_j) - \tau_0 I_{-1}(S_j)\} & \text{if } <i,j> \in \mathcal{E}. \end{cases}$$

Proof. See Text S1.

This result shows the Markov property that the conditional distribution of S_i, given all other node labels in the network, is equal to the conditional distribution of S_i given all its neighbors. It follows immediately from (8) that if S_i and S_j are not neighbors, then they are conditionally independent.

Now we give an interpretation of w_i. From (8), it is clear that the conditional distribution of S_i depends on the weighted sum of labels of its neighbors, with weight $(w_i + w_j)\tau_1$ if $S_j = 1$ and $-(w_i + w_j)\tau_0$ if $S_j = -1$. Here $(w_i + w_j)$ is the sum of weights on both ends of a linking edge. We set w_i to be the square root of of d_i, which is the degree of gene i. As a result, a gene that interacts with many other genes in the pathway has a large weight because it may play a central role in a biological process and thus it is likely to have a large influence.

The Markovian property of (1) can be derived directly from a more general result [18], which states that a nearest neighbor Gibbs measure is equivalent to a MRF. Our proof that is specific to (1) and is needed to derive the logistic model in (7). We note that under the setting of rectangular lattice systems, Besag [23,24] presented a general logistic model called the autologistic model.

Table 3. Average FPR, sensitivities, FDR, and AUC of the 60-node simulated network.

			Weak Association				Median Association				Strong Association			
Model	**Method**		**FPR**	**Sens.**	**FDR**	**AUC**	**FPR**	**Sens.**	**FDR**	**AUC**	**FPR**	**Sens.**	**FDR**	**AUC**
1		P Value	0.0470	0.157	0.817	0.629	0.0516	0.297	0.745	0.739	0.0523	0.533	0.621	0.863
		BH	0.0021	0.037	0.085		0.0014	0.043	0.070		0.0032	0.187	0.098	
	Prior 1	PM1	0.0412	0.167	0.790	0.657	0.0468	0.297	0.695	0.776	0.0496	0.567	0.591	0.899
		MCP	0.0370	0.150	0.768		0.0418	0.253	0.696		0.0454	0.523	0.593	
		PM2	0.0018	0.030	0.087		0.0014	0.050	0.075		0.0026	0.187	0.120	
	Prior 2	PM1	0.0404	0.163	0.792	0.653	0.0437	0.290	0.689	0.768	0.0472	0.543	0.585	0.890
		MCP	0.0375	0.150	0.779		0.0416	0.277	0.689		0.0449	0.523	0.583	
		PM2	0.0016	0.030	0.085		0.0021	0.090	0.110		0.0025	0.177	0.115	
	Prior 3	PM1	0.0300	0.143	0.688	0.690	0.0326	0.293	0.592	0.795	0.0377	0.577	0.483	0.907
		MCP	0.0253	0.140	0.648		0.0270	0.247	0.594		0.0337	0.500	0.480	
		PM2	0.0023	0.040	0.107		0.0012	0.053	0.065		0.0019	0.187	0.100	
2		P Value	0.0457	0.173	0.730	0.629	0.0514	0.330	0.668	0.738	0.0505	0.465	0.565	0.840
		BH	0.0018	0.018	0.090		0.0027	0.035	0.100		0.0038	0.178	0.094	
	Prior 1	PM1	0.0389	0.168	0.694	0.659	0.0450	0.340	0.621	0.788	0.0446	0.508	0.523	0.879
		MCP	0.0370	0.145	0.701		0.0416	0.318	0.618		0.0411	0.490	0.515	
		PM2	0.0020	0.018	0.110		0.0016	0.050	0.085		0.0016	0.178	0.075	
	Prior 2	PM1	0.0379	0.170	0.692	0.653	0.0439	0.323	0.624	0.775	0.0430	0.490	0.522	0.869
		MCP	0.0370	0.153	0.694		0.0413	0.305	0.626		0.0413	0.468	0.530	
		PM2	0.0020	0.018	0.110		0.0016	0.048	0.085		0.0020	0.158	0.095	
	Prior 3	PM1	0.0289	0.143	0.639	0.683	0.0364	0.320	0.583	0.803	0.0352	0.485	0.469	0.888
		MCP	0.0252	0.125	0.610		0.0321	0.288	0.577		0.0293	0.455	0.454	
		PM2	0.0014	0.018	0.080		0.0013	0.048	0.065		0.0027	0.205	0.108	
3		P Value	0.0478	0.148	0.756	0.648	0.0476	0.326	0.591	0.744	0.0458	0.524	0.468	0.856
		BH	0.0038	0.028	0.129		0.0015	0.050	0.051		0.0029	0.166	0.052	
	Prior 1	PM1	0.0402	0.136	0.744	0.675	0.0425	0.336	0.544	0.794	0.0409	0.584	0.422	0.897
		MCP	0.0364	0.124	0.737		0.0387	0.320	0.545		0.0373	0.568	0.403	
		PM2	0.0027	0.022	0.150		0.0011	0.056	0.053		0.0020	0.230	0.060	
	Prior 2	PM1	0.0404	0.134	0.738	0.669	0.0411	0.318	0.559	0.779	0.0398	0.556	0.427	0.886
		MCP	0.0380	0.124	0.749		0.0393	0.302	0.559		0.0375	0.546	0.414	
		PM2	0.0025	0.026	0.140		0.0009	0.052	0.045		0.0018	0.206	0.062	
	Prior 3	PM1	0.0282	0.120	0.654	0.700	0.0333	0.310	0.486	0.806	0.0315	0.556	0.381	0.904
		MCP	0.0247	0.100	0.620		0.0280	0.270	0.479		0.0276	0.540	0.358	
		PM2	0.0020	0.014	0.110		0.0011	0.046	0.055		0.0015	0.216	0.052	

doi:10.1371/journal.pgen.1001353.t003

The MRF property of the posterior distribution

To see that the posterior distribution is also a MRF, note that for node i,

$$Pr(S_i=+1|\boldsymbol{y},S_{\mathcal{V}-i},\boldsymbol{\theta}_0,\boldsymbol{\theta}_1) \propto f_1(y_i|\boldsymbol{\theta}_1)Pr(S_i=+1|S_{\mathcal{V}-i},\boldsymbol{\theta}_0)$$

$$=f_1(y_i|\boldsymbol{\theta}_1)Pr(S_i=+1|S_{N_i},\boldsymbol{\theta}_0).$$

Thus, the conditional posterior distribution of S_i given all other nodes only depends on its neighbors, which means the posterior distribution is also a MRF. The conditional posterior log odds of S_i is

$$h+\log LR(y_i;\boldsymbol{\theta}_1)+\tau_1\left(w_iJ_i^{(1)}+\sum_{k\in N_i}w_kI_1(S_k)\right)-\tau_0\left(w_iJ_i^{(-1)}+\sum_{k\in N_i}w_kI_{-1}(S_k)\right), \qquad (9)$$

where

$$LR(y_i;\boldsymbol{\theta}_1)=\frac{f_1(y_i|\boldsymbol{\theta}_1)}{f_0(y_i)}$$

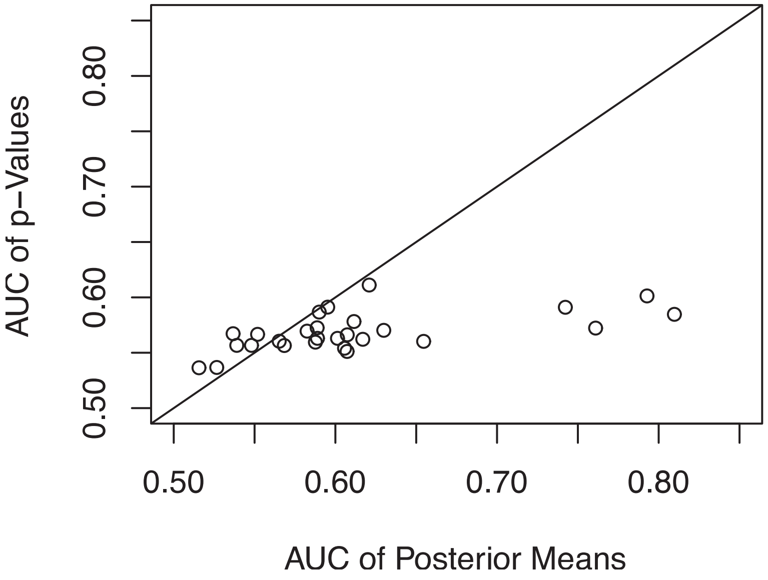

Figure 6. AUC comparison of rankings by p values and posterior means for Crohn's disease data. AUC values of the rankings by p values are on the y axis and that of the posterior means are on the x axis for pathways containing three or more target genes.
doi:10.1371/journal.pgen.1001353.g006

is the marginal likelihood ratio. Therefore, (9) is the product of the marginal likelihood ratio, reflecting the evidence from the data for association with the disease, and the conditional prior odds, reflecting the effect from interactions among neighboring genes from the biological pathway.

To make it clear, we can rewrite (9) in the form of a system of auto-logistic regression equations:

$$\text{logit}\,Pr(S_i|\boldsymbol{y},S_{\mathcal{V}-i},\boldsymbol{\theta}_0,\boldsymbol{\theta}_1)=\beta^{'}_{i0}+\beta_{i1}S_1+\cdots+\beta_{in}S_n,\ i=1,\cdots,n,\quad (10)$$

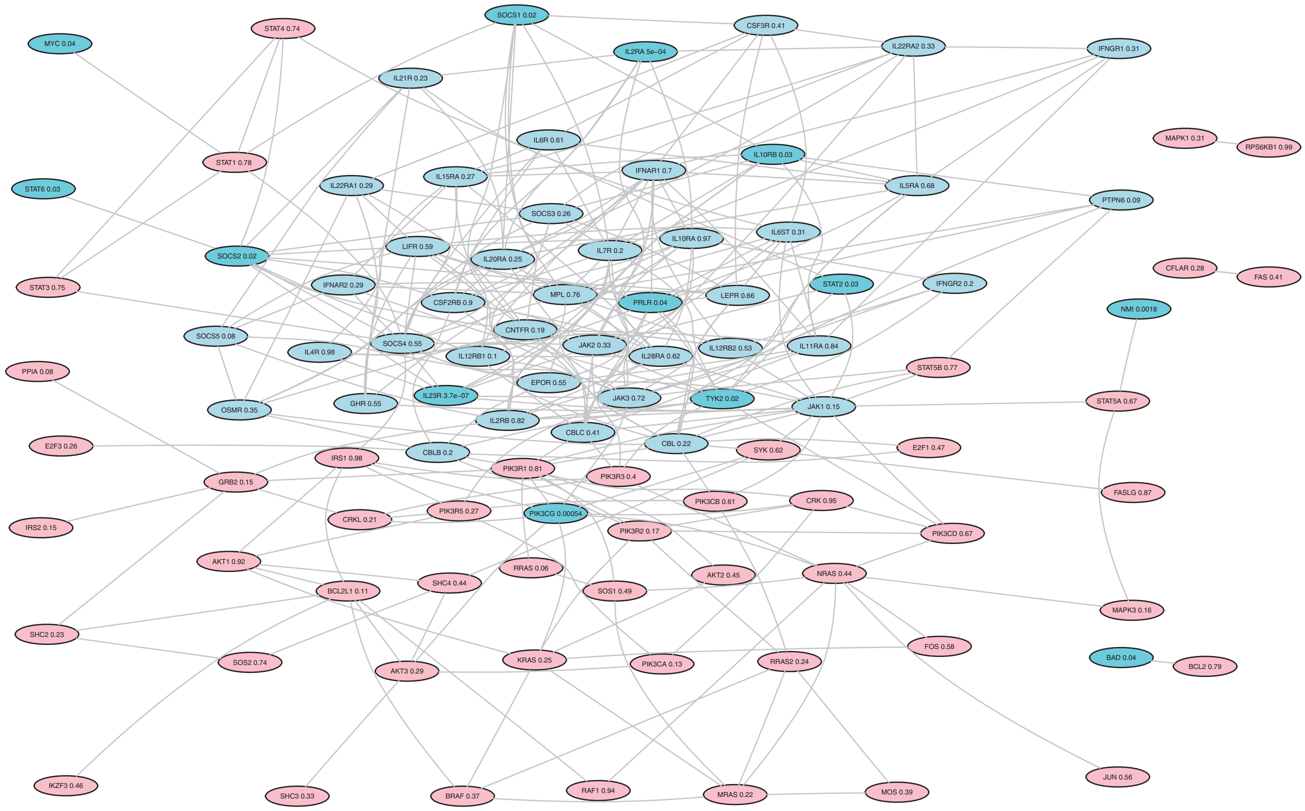

Figure 7. IL-2 receptor Beta chain in T cell activation pathway. Significant genes whose p values are below 0.05 are colored in cyan, genes with improved ranks are colored in light blue and others are colored in pink.
doi:10.1371/journal.pgen.1001353.g007

where

$$\beta'_{i0} = h + \log LR(y_i;\boldsymbol{\theta}_1),$$

$$\beta_{ij} = \begin{cases} 0 & \text{if } i=j \text{ or } <i,j> \notin \mathcal{E} \\ (w_i+w_j)\{\tau_1 I_1(S_j) - \tau_0 I_{-1}(S_j)\} & \text{if } <i,j> \in \mathcal{E} \end{cases}$$

There are a few observations. First, it is easy to see that the posterior conditional logit form in (9) is the same as the prior conditional logit in (8) except its intercept is $h + \log LR(y_i;\boldsymbol{\theta}_1)$. Thus, the observed log likelihood ratio provides a fixed additive effect to the prior logit. Second, the coefficient matrix is symmetric, i.e., $\beta_{ij} = \beta_{ji}$. If gene i and j are not neighbors, then $\beta_{ij} = \beta_{ji} = 0$ and they are conditionally independent. On the other hand, if they are neighbors, then the impact between each other is equal. Third, genes i and j are in general correlated in their joint posterior distribution, even if they are not neighbors and are conditionally independent. Moreover, the more common neighbors they share with each other, the stronger the correlation between the two.

The MCMC algorithm

To sample from the posterior distribution, here we implement a Gibbs sampler that is well suited for a MRF. The algorithm is described as follows. First we set an initial value for $\boldsymbol{S}$, say $\boldsymbol{s}^{(0)}$. Then in step k, we update the labels sequentially for $i = 1,\cdots,n$ according to (10):

$$\text{logit}Pr(s_i^{(k)}|\boldsymbol{y}, s_1^{(k)},\cdots,s_{i-1}^{(k)}, s_{i+1}^{(k-1)},\cdots,s_n^{(k-1)},\boldsymbol{\theta}_0,\boldsymbol{\theta}_1)$$

$$= \beta'_{i0} + \beta_{i1}s_1^{(k)} + \cdots + \beta_{i,i-1}s_{i-1}^{(k)} + \beta_{i,i+1}s_{i+1}^{(k-1)} + \cdots + \beta_{in}s_n^{(k-1)},$$

to obtain $\boldsymbol{s}^{(k)}$ from $\boldsymbol{s}^{(k-1)}$. In each cycle we may want to randomize the order in which the nodes are updated.

Crohn's disease data

The Crohn's disease [15] data set is used to evaluate the performance of the proposed method in the Results section. Crohn's disease is a type of inflammatory bowel disease characterized by chronic inflammation of discontinuous segments of the intestine. The disease is found to be related to the interaction of several factors including genetic susceptibility, the intestinal microbial flora of the patient, the patient's immune response to these microbiota, and environmental triggers [25]. It has been well established that Crohn's disease has a strong genetic component [26].

The cohort used in the analysis includes 401 cases and 433 controls. SNPs with a call rate greater than 0.9, minor allele frequency greater than 0.01, and HWE p value greater than 0.001 are kept, while subjects with a call rate less than 0.95 are removed from the analysis. Finally 397 cases and 431 controls remain in the analysis. SNPs are considered being mapped to a gene if their physical locations are within ±10 kb from the start or end point of the gene as given by Refseq annotation at the NCBI website. Gene level p values are obtained by regressing disease status on PCA components that account for at least 85% of the variation [27–29]. The pathways and genes in each pathway as well as the gene-level p values can be found at http://bioinformatics.med.yale.edu/group/software.html.

Supporting Information

Figure S1 Estimation of h via an empirical Bayes method.
Found at: doi:10.1371/journal.pgen.1001353.s001 (0.57 MB EPS)

Figure S2 Estimation of τ via an empirical Bayes method.
Found at: doi:10.1371/journal.pgen.1001353.s002 (0.56 MB EPS)

Text S1 Proof of Proposition 1.
Found at: doi:10.1371/journal.pgen.1001353.s003 (0.12 MB PDF)

Text S2 Estimating hyperparameters through a conditional empirical Bayes approach.
Found at: doi:10.1371/journal.pgen.1001353.s004 (0.15 MB PDF)

Acknowledgments

We thank Dr. David Ballard for providing gene level p values of the Crohn's disease data set.

Author Contributions

Conceived and designed the experiments: JC HZ. Performed the experiments: JC. Analyzed the data: MC. Wrote the paper: MC HZ.

References

1. Wang K, Li M, Bucan M (2007) Pathway-based approaches for analysis of genomewide association studies. Am J Hum Genet 81: 1278–1283.
2. Peng G, Luo L, Siu H, Zhu Y, Hu P, et al. (2010) Gene and pathway-based second-wave analysis of genome-wide association studies. Eur J Hum Genet 18: 111–117.
3. Ballard DH, Cho J, Zhao H (2010) Comparisons of multi-marker association methods to detect association between a candidate region and disease. Genet Epidemiol 34: 201–212.
4. Chun H, Ballard D, Cho J, Zhao H (2011) Identification of association between disease and multiple markers within a candidate region via sparse partial least squares regression. Submitted.
5. Franke L, van Bakel H, Fokkens L, de Jong ED, Egmont-Petersen M, et al. (2006) Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. Am J Hum Genet 78: 1011–1025.
6. Aerts S, Lambrechts D, Maity S, Loo PV, Coessens B, et al. (2006) Gene prioritization through genomic data fusion. Nat Biotechnol 24: 537–544.
7. Ma X, Lee H, Wang L, Sun F (2007) CGI: a new approach for prioritizing genes by combining gene expression and protein-protein interaction data. Bioinformatics 23: 215–221.
8. Hutz JE, Kraja AT, McLeod HL, Province MA (2008) Candid: a flexible method for prioritizing candidate genes for complex human traits. Genet Epidemiol 32: 779–790.
9. Köhler S, Bauer S, Horn D, Robinson PN (2008) Walking the interactome for prioritization of candidate disease genes. Am J Hum Genet 82: 949–958.
10. Wu X, Jiang R, Zhang MQ, Li S (2008) Network-based global inference of human disease genes. Mol Syst Biol 4: 189.
11. Saccone SF, Saccone NL, Swan GE, Madden PAF, Goate AM, et al. (2008) Systematic biological prioritization after a genome-wide association study: an application to nicotine dependence. Bioinformatics 24: 1805–1811.
12. Wei P, Pan W (2008) Incorporating gene networks into statistical tests for genomic data via a spatially correlated mixture model. Bioinformatics 24: 404–411.
13. Wei Z, Li H (2007) A Markov random field model for network-based analysis of genomic data. Bioinformatics 23: 1537–1544.
14. Li H, Wei Z, Maris JM (2010) A hidden markov random field model for genome-wide association studies. Biostatistics 11: 139–150.
15. Duerr RH, Taylor KD, Brant SR, Rioux JD, Silverberg MS, et al. (2006) A genome-wide association study identifies IL23R as an inflammatory bowel disease genes. Science 314: 1461–1463.
16. Salomonis N, Hanspers K, Zambon AC, Vranizan K, Lawlor SC, et al. (2007) Genmapp 2: new features and resources for pathway analysis. BMC Bioinformatics 8: 217.
17. Kanehisa M, Goto S (2000) Kegg: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28: 27–30.
18. Kindermann R, Snell JL (1980) Markov random fields and their applications. American Mathematical Society, ISBN: 0-8218-3381-2.
19. Besag J (1986) On the statistical analysis of dirty pictures. Journal of the Royal Statistical Society Series B (Methodological) 48: 259–302.

20. Müller P, Parmigiani G, Robert C, Rousseau J (2004) Optimal sample size for multiple testing: the case of gene expression microarrays. Journal of the American Statistical Association 99: 990–1001.
21. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B (Methodological) 57: 289–300.
22. Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, et al. (2008) Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat Genet 40: 955–962.
23. Besag J (1972) Nearest-neighbour systems and the auto-logistic model for binary data. Journal of the Royal Statistical Society Series B (Methodological) 34: 75–83.
24. Besag J (1974) Spatial interaction and the statistical analysis of lattice systems. Journal of the Royal Statistical Society Series B (Methodological) 36: 192–236.
25. Sartor RB (2006) Mechanisms of disease: pathogenesis of Crohn's disease and ulcerative colitis. Nat Clin Pract Gastroenterol Hepatol 3: 390–407.
26. Peeters M, Nevens H, Baert F, Hiele M, de Meyer AM, et al. (1996) Familial aggregation in Crohn's disease: increased age-adjusted risk and concordance in clinical characteristics. Gastroenterology 111: 597–603.
27. Gauderman WJ, Murcray C, Gilliland F, Conti DV (2007) Testing association between disease and multiple SNPs in a candidate gene. Genet Epidemiol 31: 383–395.
28. Wang K, Abbott D (2008) A principal components regression approach to multilocus genetic association studies. Genet Epidemiol 32: 108–118.
29. Ballard DH (2009) Integration of Genomic Data to Identify Genes and Pathways Associated with Disease. Ph.D. thesis, Yale University.