INCORPORATING BIOLOGICAL INFORMATION INTO LINEAR MODELS: A BAYESIAN APPROACH TO THE SELECTION OF PATHWAYS AND GENES

By Francesco C. Stingo[†], Yian A. Chen[‡], Mahlet G. Tadesse[§] and Marina Vannucci[†],

Rice University[†], Moffitt Cancer Center[‡] and Georgetown University[§]

The vast amount of biological knowledge accumulated over the years has allowed researchers to identify various biochemical interactions and define different families of pathways. There is an increased interest in identifying pathways and pathway elements involved in particular biological processes. Drug discovery efforts, for example, are focused on identifying biomarkers as well as pathways related to a disease. We propose a Bayesian model that addresses this question by incorporating information on pathways and gene networks in the analysis of DNA microarray data. Such information is used to define pathway summaries, specify prior distributions, and structure the MCMC moves to fit the model. We illustrate the method with an application to gene expression data with censored survival outcomes. In addition to identifying markers that would have been missed otherwise and improving prediction accuracy, the integration of existing biological knowledge into the analysis provides a better understanding of underlying molecular processes.

1. Introduction. DNA microarrays have been used successfully to identify gene expression signatures characteristic of disease subtypes (Golub et al. 1999) or distinct outcomes to therapy (Shipp et al. 2002). Many statistical methods have been developed to select genes for disease diagnosis, prognosis, and therapeutic targets. There is an increased consensus, however, that gene selection alone may not be sufficient. In cancer pharmacogenomics, for instance, cancer drugs are increasingly designed to target specific pathways to account for the complexity of the oncogenic process and the complex relationships between genes (Downward 2006). Metabolic pathways, for example, are defined as a series of chemical reactions in a living cell that can be activated or inhibited at multiple points. If a gene at the top of a signaling cascade is selected as a target, it is not guaranteed that the reaction will be successfully inactivated because multiple genes downstream can still be activated or inhibited. Signals are generally relayed via multiple signaling routes or networks. Even if a branch of the pathway is completely blocked by inhibition

1

or activation of multiple genes, the signal may still be relayed through an alternative branch or even through a different pathway (Bild et al. 2006, Solit et al. 2006). Downward (2006) pointed out that targeting a single pathway or a few signaling pathways might not be sufficient. Thus, the focus is increasingly on identifying both relevant genes and pathways. Genes and/or gene products generally interact with one another and they often function together concertedly. In this paper we propose a Bayesian model that addresses this question by incorporating information of pathway memberships and gene networks in the analysis of DNA microarray data. Such information is used to define pathway summaries, specify prior distributions, and structure the MCMC moves to fit the model.

Several public and commercial databases have been developed to structure and store the vast amount of biological knowledge accumulated over the years into functionally or biochemically related groups. These databases focus on describing signaling, metabolic or regulatory pathways. Some examples include Gene Ontology (GO) (The Gene Ontology Consortium 2000), Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa & Goto 2000), MetaCyc (Krieger et al. 2004), PathDB (www.ncgr.org/pathdb), Reactome KnowledgeBase (Joshi-Tope et al. 2005), Invitrogen iPath (www.invitrogen.com) and Cell Signaling Technology (CST) Pathway (www.cellsignal.com). The need to integrate gene expression data with the biological knowledge accumulated in these databases is well recognized. Several software packages that query pathway information and overlay DNA microarray data on pathways have been developed. Nakao et al. (1999) implemented a visualization tool that color-codes KEGG pathway diagrams to reflect changes in their gene expression levels. GenMAPP (Dahlquist et al. 2002) is another graphical tool that allows visualization of microarray data in the context of biological pathways or any other functional grouping of genes. Doniger et al. (2003) have made use of GenMAPP to view genes involved in specific GO terms. Another widely used method that relates pathways to a set of differentially expressed genes is the gene set enrichment analysis (GSEA) (Subramanian et al. 2005). Given a list of genes GSEA computes an enrichment score to reflect the degree to which a pre-defined pathway is over-represented at the top or bottom of the ranked list. These procedures are useful starting points to observe gene expression changes in the context of known biological processes.

Some recent studies have gone a step further and have focused on incorporating pathway information or gene-gene network information into the analysis of gene expression data. For example, Park et al. (2007) have attempted to incorporate GO annotation to predict survival time, by first grouping genes based on their GO membership, calculating the first principal component to form a super-gene within each cluster then applying a Cox model with L_1 penalty to identify super-genes, i.e., GO terms related to the outcome. Wei & Li (2007) have considered a small set of 33

pre-selected signaling pathways and used the implied relationships among genes to infer differentially expressed genes, and Wei & Li (2008) have extended this work by including a temporal dimension. Li & Li (2008) and Pan et al. (2009) have proposed two different procedures that use the gene-gene network to build penalties in a regression model framework for gene selection. Bayesian approaches have also been developed. Li & Zhang (2010) have incorporated the dependence structure of transcription factors in a regression model with gene expression outcomes. In their approach a network is defined based on the Hamming distance between candidate motifs and used to specify a Markov random field prior for the motif selection indicator. Telesca et al. (2008) have proposed a model for the identification of differentially expressed genes that takes into account the dependence structure among genes from available pathways while allowing for correction in the gene network topology.

These methods use the gene-pathway relationships or gene network information to identify either the important pathways or the genes. Our goal is to develop a more comprehensive method that selects both pathways and genes using a model that incorporates pathway-gene relationships and gene dependence structures. In order to identify relevant genes and pathways, latent binary vectors are introduced and updated using a two-stage Metropolis-Hastings sampling scheme. The gene networks are used to define a Markov random field prior on the gene selection indicators and to structure the Markov chain Monte Carlo (MCMC) moves. In addition, the pathway information is used to derive pathway expression measures that summarize the group behavior of genes within pathways. In this paper we make use of the first latent components obtained by applying partial least squares (PLS) regressions on the selected genes from each pathway. PLS is an efficient statistical regression technique that is frequently used for the analysis of genomic and proteomic data, see Boulesteix & Strimmer (2007). We apply the model to simulated and real data using the pathway structure from the KEGG database.

Our simulation studies show that employing the MRF prior allows us to achieve a better separation of the relevant pathways from the non-relevant ones. In addition, in a simulated setting with fairly small regression coefficients the model with the MRF prior is able to select all the correct genes without any false positive while the model without MRF includes 3 false positives. Other authors have reported similar results. Li & Zhang (2010), in particular, comment on the effect of the MRF prior on the selection power in their linear regression setting. They also notice that adding the MRF prior implies a relatively small increase in computational cost. Wei & Li (2007, 2008) report that their method is quite effective in identifying genes and modified subnetworks and that it has higher sensitivity than commonly used procedures that do not use the pathway structure, with similar and, in some cases, lower false discovery rates. Furthermore, in our formulation of the model we

use the network information not only for prior specification but also to structure the MCMC moves. This is helpful for arriving at promising models in a faster way by proposing relevant configurations. In real data applications the integration of the pathway information may allow the identification of relevant predictors that would have been missed otherwise, aiding the interpretation of the results, in particular for the selected genes that are connected in the MRF, and also improving the prediction accuracy of the selected models.

The paper is organized as follows. In Section 2 we discuss the model formulation and prior specification. Section 3 describes the MCMC procedure to fit the model and strategies for posterior inference. In Section 4 we evaluate the performance of the method using simulated data and illustrate an application of the method to gene expression data with survival outcomes. We conclude with a brief discussion in Section 5.

2. Model specification.   In this Section we describe how we incorporate pathway and gene network information into a Bayesian modeling framework for gene and pathway selection. Figure 1 provides a schematic representation of our proposed approach and model.

2.1. *Regression on latent measures of pathway activity.*   Our goal is to build a model for identifying pathways related to a particular phenotype while simultaneously locating genes from these selected pathways that are involved in the biological process of interest. The data we have available for analysis consist of:

1. Y, an $n \times 1$ vector of outcomes.
2. $\mathbf{X}$, an $n \times p$ matrix of gene expression levels. Without loss of generality, $\mathbf{X}$ is centered so that its columns sum to 0.
3. $\mathbf{S}$, a $K \times p$ matrix indicating membership of genes in pathways, with elements $s_{kj} = 1$ if gene j belongs to pathway k, and $s_{kj} = 0$ otherwise.
4. $\mathbf{R}$, a $p \times p$ matrix describing relationships between genes, with elements $r_{ij} = 1$ if genes i and j have a direct link in the gene network, and $r_{ij} = 0$ otherwise.

The matrices $\mathbf{S}$ and $\mathbf{R}$ are constructed using information retrieved from pathway databases, see the application in Section 4.2 for details.

Since the goal of the analysis is to study the association between the response variable and the pathways, we need to derive a score as a measure of "pathway expression" that summarizes the group behavior of included genes within pathways. We do this by using the latent components from a PLS regression of Y on selected subsets of genes from each pathway. A number of recent studies have, in fact, applied dimension reduction techniques to capture the group behavior of multiple genes. Pittman et al. (2004), for instance, first apply k-means clustering
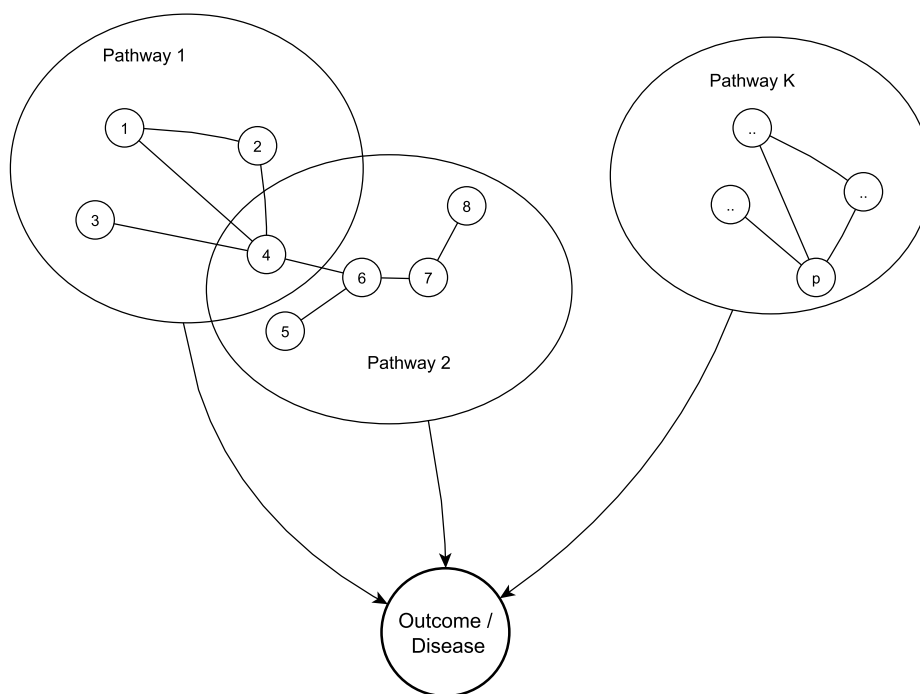
FIG 1. *Schematic representation of our proposed approach. Information on known pathways and gene-gene networks is obtained from available databases. PLS components restricted to known pathways serve as possible regressors to predict a disease outcome, according to model (1). The goal of the inference is to identify the pathways to be included in the model and the genes to be included within those pathways.*

to identify subsets of potentially related genes, then use as regressors the first principal components obtained from applying principal component analysis (PCA) to each cluster. Bair et al. (2006) start by removing genes that have low univariate correlation with the outcome variable then apply PCA on the remaining genes to form clusters or conceptual pathways, which are used as regressors. In our method, instead of attempting to infer conceptual pathways, we use the existing pathway information. We compute a pathway activity measure by applying PLS regression of Y on a subset of selected genes from the pathway. PLS has the advantage of taking into account the covariance between regressors and the response variable Y, whereas PCA focuses solely on the variability in the covariate data. The selection of a subset of gene expressions to form the PLS components is similar in spirit to the sparse PCA method proposed by Zou et al. (2006), which selects variables to be used to form the principal components.

In order to identify both relevant groups and important genes we introduce two binary vector indicators, a $K \times 1$ vector $\boldsymbol{\theta}$ for the inclusion of the groups and a $p \times 1$ vector $\boldsymbol{\gamma}$ for the inclusion of genes, i.e. $\gamma_j = 1$ if gene j is selected for at least one pathway score, and $\gamma_j = 0$ otherwise. Let us assume for the moment a continuous response Y. The linear regression model that relates the response variable to the selected pathways and genes can then be written as

$$
(1) \qquad Y = \mathbf{1}\alpha + \sum_{k=1}^{K_\theta} T_{k(\gamma)}\beta_{k(\gamma)} + \varepsilon, \quad \varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}),
$$

where $K_\theta = \sum_{k=1}^{K} \theta_k$ is the number of selected pathways and where $T_{k(\gamma)}$ corresponds to the first latent PLS component generated based on the expression levels of selected genes belonging to pathway k, that is using the X_j's corresponding to $s_{kj} = 1$ and $\gamma_j = 1$. To be more precise, let pathway k contain $p_k = \sum_{j=1}^{p} s_{kj}$ genes and let $p_{k\gamma} = \sum_{j=1}^{p} s_{kj}\gamma_j$ denote the number of selected genes (i.e., genes included in the model) that belong to pathway k. Then $T_{k(\gamma)}$ corresponds to the first latent PLS component generated by applying PLS to the expression data of the $p_{k\gamma}$ genes, denoted as $\mathbf{X}_{k(\gamma)}$,

$$
T_{k(\gamma)} = \mathbf{X}_{k(\gamma)} U_1,
$$

where U_1 is the $p_{k\gamma} \times 1$ eigenvector corresponding to the largest eigenvalue of $C_{xy}C_{xy}^T$, with $C_{xy} = cov(\mathbf{X}_{k(\gamma)}, Y)$ (see for example Lindgren et al. 1993). Thus, $T_k(\gamma)$ is an $n \times 1$ vector and $\beta_k(\gamma)$ is a scalar. Model (1) can therefore be seen as a PLS regression model with PLS components restricted to available pathways, and where the goal of the inference is to identify the pathways to be included in the model, and the genes to be included within those pathways.

2.2. *Models for categorical or censored outcomes.* In the construction above, we have assumed a continuous response. However, our model formulation can easily be extended to handle categorical or censored outcome variables.

When Y is a categorical variable taking one of G possible values, $0, \ldots, G-1$, a probit model can be used, as done by Albert & Chib (1993), Sha et al. (2004) and Kwon et al. (2007). Briefly, each outcome Y_i is associated with a vector $(p_{i,0}, \ldots, p_{i,G-1})$, where $p_{ig} = P(Y_i = g)$ is the probability that subject i falls in the g-th category. The probabilities p_{ig} can be related to the linear predictors using a data augmentation approach. Let $\mathbf{Z}_i$ be latent data corresponding to the unobserved propensities of subject i to belong to one of the classes. When the observed outcomes Y_i correspond to nominal values, the relationship between Y_i and $\mathbf{Z}_i = (z_{i,1}, \ldots, z_{i,G-1})$ is defined as

$$(2) \quad Y_i = \begin{cases} 0 & \text{if } \max_{1 \leq l \leq G-1}\{z_{i,l}\} \leq 0 \\ g & \text{if } \max_{1 \leq l \leq G-1}\{z_{i,l}\} > 0 \text{ and } z_{i,g} = \max_{1 \leq l \leq G-1}\{z_{i,l}\} \end{cases}.$$

A multivariate normal model can then be used to associate $\mathbf{Z}_i$ to the predictors

$$(3) \quad \mathbf{Z}_i = \mathbf{1}\alpha + \sum_{k=1}^{K_\theta} T_{i,k(\gamma)}\boldsymbol{\beta}_{k(\gamma)} + \boldsymbol{\varepsilon}_i, \quad \boldsymbol{\varepsilon}_i \sim \mathcal{N}(0, \boldsymbol{\Sigma}), \quad i = 1, \ldots, n.$$

If the observed outcomes Y_i correspond, instead, to ordinal categories, the latent variable Z_i is defined such that

$$(4) \quad Y_i = g \quad \text{if } \delta_g < Z_i \leq \delta_{g+1}, \quad g = 0, \ldots, G-1,$$

where the boundaries δ_g are unknown and $-\infty = \delta_0 < \delta_1 < \ldots < \delta_{G-1} < \delta_G = \infty$. The latent variable Z_i is then associated with the predictors through the linear model

$$(5) \quad Z_i = \alpha + \sum_{k=1}^{K_\theta} T_{i,k(\gamma)}\beta_{k(\gamma)} + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \quad i = 1, \ldots, n.$$

For censored survival outcomes, an accelerated failure time (AFT) model can be used (Wei 1992, Sha et al. 2006). In this case, the observed data are $Y_i = \min(T_i, C_i)$ and $\delta_i = I\{Y_i \leq C_i\}$, where T_i is the survival time for subject i, C_i is the censoring time, and δ_i is a censoring indicator. A data augmentation approach can be used and latent variables Z_i can be introduced such that

$$(6) \quad \begin{cases} Z_i = \log(Y_i) & \text{if } \delta_i = 1 \\ Z_i > \log(Y_i) & \text{if } \delta_i = 0 \end{cases}.$$

The AFT model can then be written in terms of the latent Z_i,

$$(7) \qquad Z_i = \alpha + \sum_{k=1}^{K_\theta} T_{i,k(\gamma)} \beta_{k(\gamma)} + \varepsilon_i,$$

where the ε_i's are independent and identically distributed random variables that may take one of several parametric forms. Sha et al. (2006) consider cases where ε_i follows a normal or a t-distribution.

2.3. *Prior for regression parameters.* The regression coefficient β_k in (1) measures the effect of the PLS latent component summarizing the effect of pathway k on the response variable. However, not all pathways are related to the phenotype and the goal is to identify the predictive ones. Bayesian methods that use mixture priors for variable selection have been thoroughly investigated in the literature, in particular for linear models, see George & McCulloch (1993, 1997) for multiple regression, Brown et al. (1998) for extensions to multivariate responses and Sha et al. (2004) for probit models. A comprehensive review on special features of the selection priors and on computational aspects of the method can be found in Chipman et al. (2001). Similarly, here, we use the latent vector $\boldsymbol{\theta}$ to specify a scale mixture of a normal density and a point mass at zero for the prior on each β_k in model (1):

$$(8) \qquad \beta_k | \theta_k, \sigma^2 \sim \theta_k \cdot \mathcal{N}(\beta_0, h\sigma^2) + (1 - \theta_k) \cdot \delta_0(\beta_k), \qquad k = 1, \ldots, K.$$

where $\delta_0(\beta_k)$ is a Dirac delta function. The hyperparameter h in (8) regulates, together with the hyperparameters of $p(\boldsymbol{\theta}, \boldsymbol{\gamma} | \eta)$ defined in Section 2.4 below, the amount of shrinkage in the model. We follow the guidelines provided by Sha et al. (2004) and specify h in the range of variability of the data so as to control the ratio of prior to posterior precision.

For the intercept term, α, and the variance, σ^2, we take conjugate priors

$$(9) \qquad \begin{aligned} \alpha | \sigma^2 &\sim \mathcal{N}(\alpha_0, h_0 \sigma^2) \\ \sigma^2 &\sim \text{Inv-Gamma}(\nu_0/2, \nu_0 \sigma_0^2/2), \end{aligned}$$

where α_0, β_0, h_0, h, ν_0 and σ_0^2 are to be elicited.

2.4. *Priors for pathway and gene selection indicators.* In this section we define the prior distributions for the pathway selection indicator, $\boldsymbol{\theta}$, and gene selection indicator, $\boldsymbol{\gamma}$. These priors are first defined marginally then jointly, taking into account some necessary constraints.

Each element of the latent K-vector $\boldsymbol{\theta}$ is defined as

(10) $$\theta_k = \begin{cases} 1 & \text{if pathway } k \text{ is represented in the model} \\ 0 & \text{otherwise} \end{cases}$$

for $k = 1, \ldots, K$. We assume independent Bernoulli priors for the θ_k's,

(11) $$p(\boldsymbol{\theta}|\varphi_k) = \prod_{k=1}^{K} \varphi_k^{\theta_k}(1 - \varphi_k)^{1-\theta_k},$$

where φ_k determines the proportion of pathways expected *a priori* in the model. A mixture prior can be further specified for φ_k to achieve a better discrimination in terms of posterior probabilities between significant and non-significant pathways by inflating $p(\theta_k = 0)$ toward 1 for the non-relevant pathways, as first suggested by Lucas et al. (2006),

(12) $$p(\varphi_k) = \rho\delta_0(\varphi_k) + (1 - \rho)\mathcal{B}(\varphi_k|a_0, b_0),$$

where $\mathcal{B}(\varphi_k|a_0, b_0)$ is a Beta density function with parameters a_0 an b_0. Since inference on φ_k is not of interest, it can be integrated out to simplify the MCMC implementation. This leads to the following marginal prior for $\boldsymbol{\theta}$

(13) $$p(\boldsymbol{\theta}) = \prod_k \left[\rho \cdot (1 - \theta_k) + (1 - \rho) \cdot \frac{B(a_0 + \theta_k, b_0 + 1 - \theta_k)}{B(a_0, b_0)} \right],$$

where $B(\cdot, \cdot)$ is the Beta function. Prior (13) corresponds to a product of Bernoulli distributions with parameter $\varphi_k^* = \frac{a_0(1-\rho)}{a_0+b_0}$.

For the latent p-vector $\boldsymbol{\gamma}$ we specify a prior distribution that is able to take into account not only the pathway membership of each gene but also the biological relationships between genes within and across pathways, which are captured by the matrix R. Following Li & Zhang (2010) we model these relations using a Markov random field (MRF), where genes are represented by nodes and relations between genes by edges. A MRF is a graphical model in which the distribution of a set of random variables follow Markov properties that can be described by an undirected graph. In particular, a pair of genes that are not connected are considered conditionally independent given all other genes (Besag 1974). Relations on the MRF are represented by the following probabilities

(14) $$p(\gamma_j|\eta, \gamma_i, i \in N_j) = \frac{\exp(\gamma_j F(\gamma_j))}{1 + \exp(F(\gamma_j))},$$

where $F(\gamma_j) = (\mu + \eta \sum_{i \in N_j} \gamma_i))$ and N_j is the set of direct neighbors of gene j in the MRF using only pathways represented in the model, *i.e.*, pathways with $\theta_k = 1$. The corresponding global distribution on the MRF is given by

$$(15) \qquad p(\boldsymbol{\gamma}|\boldsymbol{\theta}, \mu, \eta) \propto \exp(\mu \mathbf{1}_p' \boldsymbol{\gamma} + \eta \boldsymbol{\gamma}' \mathbf{R} \boldsymbol{\gamma})$$

with $\mathbf{1}_p$ the unit vector of dimension p and $\mathbf{R}$ the matrix introduced in section 2.1. The parameter μ controls the sparsity of the model, while η regulates the smoothness of the distribution of $\boldsymbol{\gamma}$ over the graph by controlling the prior probability of selecting a gene based on how many of its neighbors are selected. In particular, higher values of η encourage the selection of genes with neighbors already selected into the model. If a gene does not have any neighbor, then its prior distribution reduces to an independent Bernoulli with parameter $p = \exp(\mu)/[1+\exp(\mu)]$, which is a logistic transformation of μ.

Here, unlike Li & Zhang (2010), who fix both parameters of the MRF prior, we specify a hyperprior for η. We give positive probability to values of η bigger than 0, as negative values of this parameter would favor neighboring genes to have different inclusion status, which is counter-intuitive from a biological point of view. Such restriction on the domain of η also helps with the "phase transition" problem that typically occurs with MRF parameterizations of type (14), where the dimension of the selected model increases massively for small increments of η. When the phase transition occurs the number of selected genes increases substantially and the sparsity of the model gets compromised. Here, after having detected the phase transition value η_{PT}, by simulating from (15) over a grid of η values, we specify a Beta distribution $Beta(c_0, d_0)$ on η/η_{PT}.

Constraints need to be imposed to ensure both interpretability and identifiability of the model. We essentially want to avoid the following scenarios:

1. creation of empty pathways, *i.e.*, selecting a pathway but none of its member genes;
2. creation of orphan genes, *i.e.*, selecting a gene but none of the pathways that contain it;
3. selection of identical subsets of genes by different pathways, a situation that generates identical values $T_{k(\gamma)}$ and $T_{k'(\gamma)}$ to be included in the model.

These constraints imply that some combinations of $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ values are not allowed. The joint prior probability for $(\boldsymbol{\theta}, \boldsymbol{\gamma})$ taking into account these constraints is given by

$$p(\boldsymbol{\theta}, \boldsymbol{\gamma}|\eta) \propto \begin{cases} \prod_{k=1}^{K} \varphi_k^{*\theta_k}(1 - \varphi_k^*)^{1-\theta_k} \exp(\mu \mathbf{1}_p' \boldsymbol{\gamma} + \eta \boldsymbol{\gamma}' \mathbf{R} \boldsymbol{\gamma}) & \text{for valid configurations,} \\ 0 & \text{for invalid configurations.} \end{cases}$$

3. Model fitting. We now describe our MCMC procedure to fit the model and discuss strategies for posterior inference. Clearly, exploring huge posterior spaces, like the one we are dealing with, is a challenging problem. In the Bayesian literature on variable selection for standard linear regression models stochastic search algorithms have been designed to explore the posterior space, and have been successfully employed in genomic applications with prohibitive settings, handling models with thousands of genes. A key to these applications is the assumption of sparsity of the model, i.e., the belief that the response is associated with a small number of regressors. A stochastic search then allows one to explore the posterior space in an effective way, quickly finding the most probable configurations, i.e., those corresponding to the coefficients that have high marginal probabilities, while spending less time in regions with low posterior probability.

We describe below the MCMC algorithm we have designed for our problem. In particular, borrowing from the literature on stochastic searches for variable selection, we work with a marginalized model and design a Metropolis-Hastings algorithm that updates the indicator parameters for the inclusion of pathways and genes with a set of moves that add and/or delete a single gene and a single pathway. Also, we update the parameter η of the MRF from its posterior distribution by employing the general method proposed by Møller et al. (2006). In the Appendix we discuss how our Bayesian stochastic search variable selection kernel generates an ergodic Markov chain over the restricted space. In applications, we have found that a good way to asses if the stochastic exploration can be considered satisfactory is to check the concordance of the posterior probabilities obtained from different chains started from different initial points.

3.1. *Marginal Posterior probabilities.* The model parameters consist of $(\alpha, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma}, \boldsymbol{\theta}, \eta)$. The MCMC procedure can be made more efficient by integrating out some of the parameters. Here, we integrate out the regression parameters, α, $\boldsymbol{\beta}$ and σ^2. This leads to a multivariate t-distribution

$$(16) \quad f(Y|\boldsymbol{T}, \boldsymbol{\theta}, \boldsymbol{\gamma}) \sim \mathcal{T}_{\nu_0}(\alpha_0 \mathbf{1}_n + \boldsymbol{T}_{(\theta,\gamma)}\beta_0, \sigma_0^2(\mathbf{I}_n + h_0 \mathbf{1}_n \mathbf{1}_n' + \boldsymbol{T}_{(\theta,\gamma)}\boldsymbol{\Sigma}_0 \boldsymbol{T}'_{(\theta,\gamma)})),$$

with ν_0 degrees of freedom and $\mathbf{1}_n$ an n-vector of ones, and where $\boldsymbol{\Sigma}_0 = h\mathbf{I}_{K_\theta}$, with $\mathbf{I}_n$ the $n \times n$ identity matrix, and $\boldsymbol{T}_{(\theta,\gamma)}$ the $n \times K_\theta$ matrix derived from the first PLS latent components for the selected pathways using the selected genes. In the notation (16) the two arguments of the t-distribution represent the mean and the scale parameter of the distribution, respectively. The posterior probability distribution of the pathway and gene selection indicators is then given by

$$(17) \quad f(\boldsymbol{\theta}, \boldsymbol{\gamma}, \eta|\boldsymbol{T}, Y) \propto f(Y|\boldsymbol{T}, \boldsymbol{\theta}, \boldsymbol{\gamma}) \cdot p(\boldsymbol{\theta}, \boldsymbol{\gamma}|\eta) \cdot p(\eta).$$

3.2. *MCMC sampling.* The MCMC steps to fit the model consist of: (I) sampling the pathway and gene selection indicators from $p(\boldsymbol{\theta}, \boldsymbol{\gamma}|\text{rest})$; (II) sampling the MRF parameter from $p(\eta|\text{rest})$; (III) sampling additional parameters that would be introduced when fitting a probit model for categorical outcomes or an AFT model for survival outcomes.

(I) The parameters $(\boldsymbol{\theta}, \boldsymbol{\gamma})$ are updated using a Metropolis-Hastings algorithm in a two-stage sampling scheme. The pathway-gene relationships are used to structure the moves and account for the constraints specified in Section 2.4. Details of the MCMC moves for updating $(\boldsymbol{\theta}, \boldsymbol{\gamma})$ are provided in the Appendix. Briefly, they consist of randomly choosing one of the following three move types:

  1. change the inclusion status of gene and pathway by randomly choosing between adding a pathway and a gene or removing them both;

  2. change the inclusion status of gene but not pathway by randomly choosing between adding a gene or removing a gene;

  3. change the inclusion status of pathway but not gene by randomly choosing between adding a pathway or removing a pathway.

(II) At this step we want to draw the MRF parameter η from the posterior density

$$(18) \qquad p(\eta|\boldsymbol{\gamma}) \propto p(\eta)p(\boldsymbol{\gamma}|\eta).$$

The prior distribution on $\boldsymbol{\gamma}$ is of the form

$$(19) \qquad p(\boldsymbol{\gamma}|\eta) = q_\eta(\boldsymbol{\gamma})/Z_\eta$$

with unnormalised density $q_\eta(\boldsymbol{\gamma})$ and a normalizing constant Z_η which is not available analytically. When calculating the Metropolis-Hastings ratio to determine the acceptance probability of a new value η^{new}, that is,

$$(20) \qquad H(\eta^{new}|\eta^{old}) = \frac{p(\eta^{new})q_{\eta^{new}}(\boldsymbol{\gamma})q(\eta^{old}|\eta^{new})}{p(\eta^{old})q_{\eta^{old}}(\boldsymbol{\gamma})q(\eta^{new}|\eta^{old})} \bigg/ \frac{Z_{\eta^{new}}}{Z_{\eta^{old}}} \,,$$

one needs to take into account that $Z_{\eta^{new}}/Z_{\eta^{old}} \neq 1$. Following Møller et al. (2006), we introduce an auxiliary variable w, defined on the same state space as that of $\boldsymbol{\gamma}$, which has conditional density $f(w|\eta, \boldsymbol{\gamma})$ and consider the posterior

$$p(\eta, w|\boldsymbol{\gamma}) \propto f(w|\eta, \boldsymbol{\gamma})p(\eta)q_\eta(\boldsymbol{\gamma})/Z_\eta$$

which of course still involves the unknown Z_η. Obviously, marginalization over w of $p(\eta, w|\boldsymbol{\gamma})$ gives the desired distribution $p(\eta|\boldsymbol{\gamma})$. Now, if (η^{old}, w^{old}) is the current state of the algorithm, we first propose η^{new} with density $q(\eta^{new}|\eta^{old})$ then w^{new} with density $q(w^{new}|w^{old}, \eta^{new}, \eta^{old})$. As usual, the choice of these proposal densities is arbitrary from the point of view of the equilibrium distribution of the chain of η values. The choice of $f(w|\eta, \boldsymbol{\gamma})$ is also arbitrary. The key idea of the method proposed by Møller et al. (2006) is to take the proposal density for the auxiliary variable w to be of the same form as (19), but dependent on η^{new} rather than η^{old}, that is,

$$(21) \quad q(w^{new}|w^{old}, \eta^{new}, \eta^{old}) = p(w^{new}|\eta^{new}) = q_{\eta^{new}}(w^{new})/Z_{\eta^{new}}.$$

Then the Metropolis-Hastings ratio becomes
(22)
$$H(\eta^{new}, w^{new}|\eta^{old}, w^{old}) = \frac{f(w^{new}|\eta^{new}, \boldsymbol{\gamma})p(\eta^{new})q_{\eta^{new}}(\boldsymbol{\gamma})q_{\eta^{old}}(w^{old})q(\eta^{old}|\eta^{new})}{f(w^{old}|\eta^{old}, \boldsymbol{\gamma})p(\eta^{old})q_{\eta^{old}}(\boldsymbol{\gamma})q_{\eta^{new}}(w^{new})q(\eta^{new}|\eta^{old})},$$

and no longer depends on $Z_{\eta^{new}}/Z_{\eta^{old}}$. The new value w^{new} for the auxiliary variable w is drawn from (21) by perfect simulation using the algorithm proposed by Propp & Wilson (1996).

(III) In the case of classification or survival outcomes, the augmented data Z need to be updated from their full conditionals using Gibbs sampling, see Sha et al. (2004), Sha et al. (2006) and Kwon et al. (2007) for details on this step.

3.3. *Posterior Inference.* The MCMC procedure results in a list of visited models with included pathways indexed by $\boldsymbol{\theta}$ and selected genes indexed by $\boldsymbol{\gamma}$, and their corresponding relative posterior probabilities. Pathway selection can be based on the marginal posterior probabilities $p(\theta_k|\boldsymbol{T}, Y)$. A simple strategy is to compute Monte-Carlo estimates by counting the number of appearances of each pathway across the visited models. Relevant pathways are identified as those with largest marginal posterior probabilities. Then relevant genes from these pathways are identified based on their marginal posterior probabilities conditional on the inclusion of a pathway of interest, $p(\gamma_j|\boldsymbol{T}, Y, I\{\theta_k s_{kj} = 1\})$. An alternative inference for gene selection is to focus on a subset of pathways, $\mathcal{P}$, and consider the marginal posterior probability conditional on at least one pathway the gene belongs to being represented in the model, $p(\gamma_j|\boldsymbol{T}, Y, I\{\sum_{k\in\mathcal{P}} \theta_k s_{kj} > 0\})$. We note that Rao-Blackwellized estimates have been employed in standard linear regression models, in place of frequency estimates, by averaging the full conditional posterior probabilities of the inclusion indicators. These estimates are computationally quite

expensive, though they may have better precision, as noted by Guan & Stephens (2009). Because of our strategy for inference, that selects first pathways and then genes conditional on selected pathways, Rao-Blackwellized estimates of marginal probabilities may not be straightforward to derive. In all simulations and examples reported in this paper we have obtained satisfactory results by simply estimating the marginal posterior probabilities with the corresponding relative frequencies of inclusion in the visited models.

Inference for a new set of observations, $(\boldsymbol{X}_f, Y_f)$ can be done via least squares prediction,

$$\widehat{Y}_f = \boldsymbol{1}_n \tilde{\alpha} + \boldsymbol{T}_{f(\theta,\gamma)} \tilde{\boldsymbol{\beta}}_{(\theta,\gamma)}, \tag{23}$$

where $\boldsymbol{T}_{f(\theta,\gamma)}$ is the first principal component based on selected genes from relevant pathways and

$$\tilde{\alpha} = \bar{Y}, \qquad \tilde{\boldsymbol{\beta}}_{(\theta,\gamma)} = (\boldsymbol{T}'_{(\theta,\gamma)} \boldsymbol{T}_{(\theta,\gamma)} + h^{-1} \boldsymbol{I}_{K_\theta})^{-1} \boldsymbol{T}'_{(\theta,\gamma)} Y,$$

with Y the response variable in the training and $\boldsymbol{T}_{(\theta,\gamma)}$ the scores obtained from the training data using selected pathways and genes included in the model. Note that for prediction purposes, since we do not know the future Y_f, a PLS regression cannot be fit. Therefore, we generate $T_{f(\theta,\gamma)}$ by considering the first latent component obtained by applying PCA to each selected pathway using the included genes.

In the case of categorical or censored survival outcomes, the sampled latent variables Z would be used to estimate $\widehat{Z}_f$ then the correspondence between Z and the observed outcome outlined in Section 2.2 can be invoked to predict Y_f (Sha et al. 2004, 2006, Kwon et al. 2007).

4. Application. We assess the performance of the model on simulated data then illustrate an application to a breast cancer data using the KEGG pathway database to define the MRF.

4.1. *Simulation studies.* We investigated the performance of our model using simulated data based on the gene-pathway relations, **S**, and gene network, **R**, of 70 pathways and 1098 genes from the KEGG database. The relevant pathways were defined by selecting 4 pathways at random. For each of the 4 selected pathways, one gene was picked at random and its direct neighbors that belong to the selected pathways were chosen. This resulted in the selection of 4 pathways and 15 genes: 7 out of 30 from the first pathway, 3 out of 35 from the second, 3 out of 105 from the third, and 2 out of 47 from the fourth pathway. Gene expressions for $n = 100$ samples were simulated for these 15 genes using an approach similar to Li & Li (2008). This was accomplished by first creating an ordering among the 15 selected genes by changing the undirected edges in the gene networks into directed edges.

The first node on the ordering, which we denote by X_{F_1}, was selected from each pathway and drawn from a standard normal distribution; note that this node has no parents. Then all child nodes directly connected only to X_{F_1} and denoted by X_{F_2} were drawn from

$$X_{F_2} \sim \mathcal{N}(X_{F_1}\rho, 1).$$

Subsequent child nodes at generation j, X_{F_j}, were drawn using all parents from

$$X_{F_j} \sim \mathcal{N}(\rho\mathbf{X}_{pa(F_j)}\mathbf{1}_{|pa(F_j)|}, 1)$$

where $pa(F_j)$ indicates the set of parents of node j and $\mathbf{X}_{pa(F_j)}$ is a matrix containing the expressions of all the $|pa(F_j)|$ parents for node j. The expression levels of the remaining 1073 genes deemed irrelevant were simulated from a standard normal density. The response variables for the $n = 100$ samples were generated from

$$Y_i = \sum_{j=1}^{15} X_{ij}\beta + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, 1), \ \ i = 1, \ldots, 100.$$

For the first dataset we set $\beta = \pm 0.5$, using the same sign for genes that belong to the same pathway. For the second and third data sets we used $\beta = \pm 1$ and $\beta = \pm 1.5$, respectively. Note that the generating process is different from the model (1) being fit.

We report the results obtained by choosing, when possible, hyperparameters that lead to weakly informative prior distributions. A vague prior is assigned to the intercept parameter α by setting h_0 to a large value tending to ∞. For σ^2, the shape parameter can be set to $\nu_0/2 = 3$, the smallest integer such that the variance of the inverse-gamma distribution is defined, and the scale parameter $\nu_0\sigma_0^2/2$ can be chosen to yield a weakly informative prior. For the vector of regression coefficients, β_k, we set the prior mean to $\beta_0 = 0$ and choose h in the range of variability of the covariates, as suggested in Section 2.3. Specifically, we set $h_0 = 10^6$, $\alpha_0 = \beta_0 = 0$, $\nu_0\sigma_0/2 = 0.5$, and $h = 0.02$. For the pathway selection indicators, θ_k, we set $\varphi_k^* = 0.01$. As for the prior at the gene level, we set $\mu = -3.5$, which corresponds to setting the proportion of genes expected *a priori* in the model to, at least, 3% of the total number of genes. Parameters φ_k^* and μ influence the sparsity of the model and consequently the magnitude of the marginal posterior probabilities. Some sensitivity to these parameters is, of course, to be expected. However, in our simulations we have noticed that the ordering of pathways and genes based on posterior probability remains roughly the same and therefore the final selections are unchanged as long as one adjusts the threshold on the posterior probabilities. Also, for the hyperprior on η, we set $\eta_{PT} = 0.092$, to avoid the phase transition

problem, and $c_0 = 5$ and $d_0 = 2$, to obtain a prior distribution that favors bigger values of η in the interval $0 \leq \eta \leq \eta_{PT}$. In our simulations we did not notice any sensitivity to the specification of c_0 and d_0.

The MCMC sampler was run for 300,000 iterations with the first 50,000 used as burn-in. We computed the marginal posterior probabilities for pathway selection, $p(\theta_k = 1|Y, \boldsymbol{T})$, and the conditional posterior probabilities for gene selection given a subset of selected pathways, $p(\gamma_j|\boldsymbol{T}, Y, I\{\sum_{k \in \mathcal{P}} \theta_k s_{kj} > 0\})$. Figure 2 displays the marginal posterior probabilities of inclusion for all 70 pathways and the conditional posterior probabilities of inclusion for all 1098 genes.

Important pathways and genes can be selected as those with highest posterior probabilities. For example, in all 3 scenarios all four relevant pathways were selected with a marginal posterior probability cut-off of 0.8. Reducing the selection threshold to a marginal posterior probability of 0.5 pulls in two false positive pathways, for all the three simulated scenarios considered. One of the false positives is the pathway with index 17 in Figure 2, which contains more than 100 genes. A closer investigation of the MCMC output reveals that different subsets of its member genes are selected whenever it is included in the model, resulting in a high marginal posterior of inclusion for the pathway but low marginal posterior probabilities for all its member genes. The second false positive pathway appears to be selected often because it contains two or three of the relevant genes that were used to simulate the response variable and were also included in the model with high marginal posterior probabilities; all its other member genes have very low probabilities of selection. As expected, the identification of the relevant genes is easier when the signal-to-noise ratio is higher. Conditional upon the best 4 selected pathways, a marginal posterior probability cut-off of 0.5 on the marginal probability of gene inclusion leads to the selection of 7, 8 and 8 relevant genes, for the three scenarios, respectively, and no false positives. With a marginal probability threshold of 0.1, 14 of the relevant genes are selected with 4 false positives for the scenario with $\beta = \pm 0.5$, while 13 relevant genes are selected with only two false positives for the simulated data with $\beta = \pm 1$. In the simulated setting with regression coefficients $\beta = \pm 1.5$ all the 15 relevant genes are selected without any false positive at a threshold of 0.12.

Generally speaking, the effect of the MRF prior depends on the concordance of the prior network with the data. For the simulated data, we found that the model with the MRF prior, compared to the same model without the MRF, performs better in terms of pathway selection as it provides a clearer separation between relevant and non relevant pathways. In particular, the average difference, over the three scenarios, between the relevant pathway with the lowest posterior probability and the non relevant pathway with the highest posterior probability is 0.28, while without the MRF prior it is only 0.18. In addition, we have observed increased sensitivity of
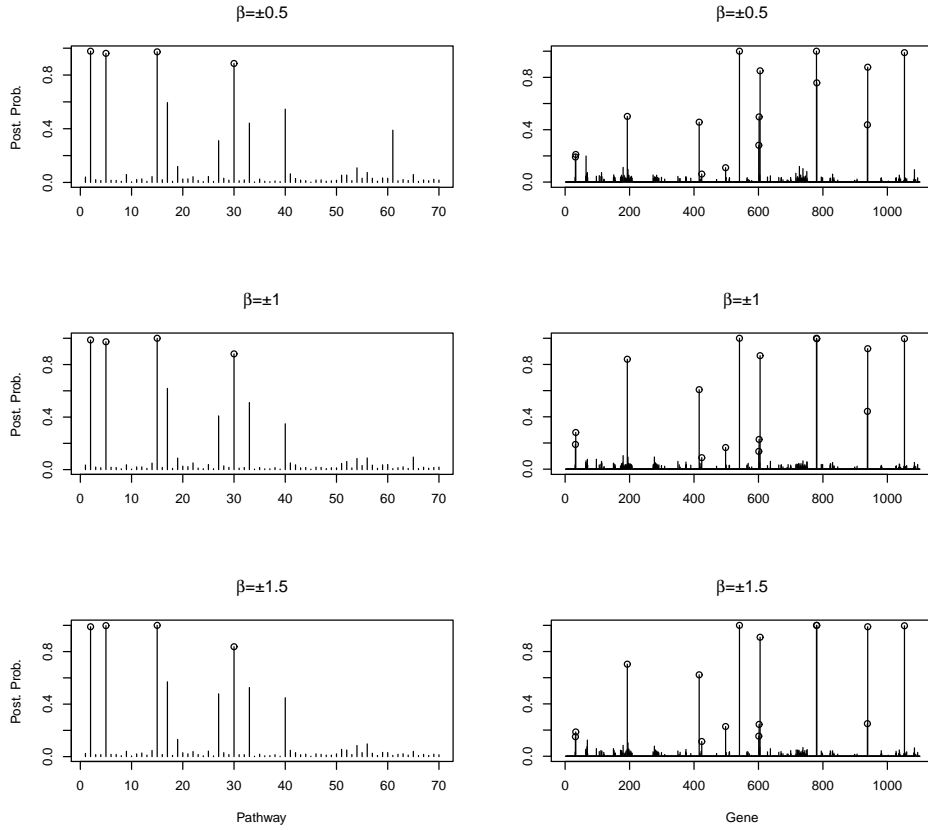
FIG 2. *Simulated data: Marginal posterior probabilities for pathway selection, $p(\theta_k|\boldsymbol{T}, Y)$, and conditional posterior probabilities for gene selection, $p(\gamma_j|\boldsymbol{T}, Y, I\{\sum_{k \in \mathcal{P}} \theta_k s_{kj} > 0\})$, for the three simulated data sets. Open circles indicate pathways and genes used to generate the outcome variable.*

the MRF prior in selecting the true variables. For example, for the simulated case with $\beta \pm 1.5$, in order to select all 15 relevant genes the marginal probability cutoff must be reduced to 0.088 at the expense of including 3 false positives. Other authors have reported similar results (Li & Zhang 2010). In the real data application we describe below, employing information on gene-gene networks aids the interpretation of the results, in particular for those selected genes that are connected in the MRF, and improves the prediction accuracy.

4.2. *Application to microarray data.* We consider the van't Veer et al. (2002) breast cancer microarray data[1]. Gene expression measures were collected on each patient using DNA microarray with 24,481 probes. Missing expression values were imputed using a k-nearest neighbor algorithm with $k = 10$. The procedure consists of identifying the k closest genes to the one with missing expression in array j using the other $n - 1$ arrays, then imputing the missing value by the average expression level of the k neighbors (Troyanskaya et al. 2001). We focus on the 76 sporadic lymph-node-negative patients, 33 of whom developed distant metastasis within 5 years and the remaining 43 did not; the latter are viewed as censored cases. We randomly split the patients into a training set of 38 samples and a test set of the same size. The goal is to identify a subset of pathways and genes that can predict time to distant metastasis for breast cancer patients.

The gene network and pathway information were obtained from the KEGG database. This was accomplished by mapping probes to pathways using the links between pathway node identifiers and LocusLink ID[2].

Using the R package *KEGGgraph* (Zhang & Wiemann 2009) we first downloaded the gene network for each pathway then merged all the networks into a single one with all the genes. A total of 196 pathways and 3,592 corresponding probes were included in the analysis. There is a many-to-many correspondence between pathways and genes, that is each pathway contains multiple genes and most genes are associated with several pathways.

We ran two MCMC chains with different starting numbers of included variables, 50 and 80, respectively. We used 600,000 iterations with a burn-in of 100,000 iterations. We incorporated the first latent vector of the PLS for each pathway into the analysis as described in Section 2.1 and set the number of pathways expected *a priori* in the model to 10% of the total number. For the gene selection, we set the hyperparameter of the Markov random field to $\mu = -3.5$, indicating that *a priori* at least 3% of genes are expected to be selected. As for η, we set $\eta_{PT} = 0.09$, to avoid the phase transition problem, and $c_0 = 1$ and $d_0 = 1$ to obtain a non informative

---

[1]available at www.rii.com/publications/2002/vantveer.htm

[2]provided at ftp://ftp.genome.ad.jp/pub/kegg/pathways/hsa/hsa_gene_map.tab and ftp://ftp.genome.ad.jp/pub/kegg/pathways/map_title.tab

prior distribution. Sensitivity analysis with different choices of these hyperparameters showed that the posterior inference is not affected by different values of c_0 and d_0. For the prior of the regression parameters, we set $\alpha_0 = \beta_0 = 0, h_0 = 10^6$ and $h = 0.1$. A vague prior was specified for σ^2 by choosing $\nu_0/2 = 3$ and $\nu_0\sigma_0^2/2 = 0.5$.

The trace plots for the number of included pathways and the number of selected genes showed good mixing (Figures not shown). The MCMC samplers mostly visited models with 20-45 pathways and 50-90 genes. To assess the agreement of the results between the two chains, we looked at the correlation between the marginal posterior probabilities for pathway selection, $p(\theta_k|\boldsymbol{T}, Y)$, and found good concordance between the two MCMC chains with a correlation coefficient of 0.9933. Concordance among the marginal posterior probabilities was confirmed by looking at a scatter plot of $p(\theta_k|\boldsymbol{T}, Y)$ across the two MCMC chains (Figure not shown).

The model also showed good predictive performance. Sha et al. (2006) already analyzed these data using an AFT model with 3,839 probes as predictors and obtained a predictive MSE of 1.9317 using the 11 probe sets with highest marginal probabilities. Our model incorporating pathway information achieved a predictive MSE of 1.4497 on the validation set, using 12 selected pathways and 41 probe sets with highest posterior probabilities. The selected pathways and genes are clearly indicated in the marginal posterior probability plots displayed in Figures 3 and 4. If we increase the marginal probability thresholds for selection and consider a model with 7 selected pathways and 14 genes, to make the comparison more fair with the results of Sha et al. (2006), we obtain a MSE of 1.7614.

From a practical point of view, researchers can use the posterior probabilities produced by our selection algorithm as a way to prioritize the relevant pathways and genes for further experimental work. For example, the genes corresponding to the best 41 selected probe sets, conditional upon the best 12 selected pathways, are listed in Table 1 divided by islands, which correspond to sets of connected genes in the Markov random field. The islands help with the biological interpretation by locating relevant branches of pathways. A subset of the selected pathways along with islands and singletons are displayed in Figure 5. Several of the identified pathways are known to be involved in tumor formation and progression. For instance, the mitogen-activated protein kinase (MAPK) signaling pathway, which is involved in various cellular functions, including cell proliferation, differentiation and migration, has been implicated in breast cancer metastasis (Lee et al. 2007, Keyse 2008). The KEGG pathway in cancers was also selected with high posterior probability. Other interesting pathways are the insulin signaling pathway, which has been linked to the development, progression and outcome of breast cancer, and purine metabolism, which is involved in nucleotide biosynthesis and affects cell cycle activity of tumor cells.
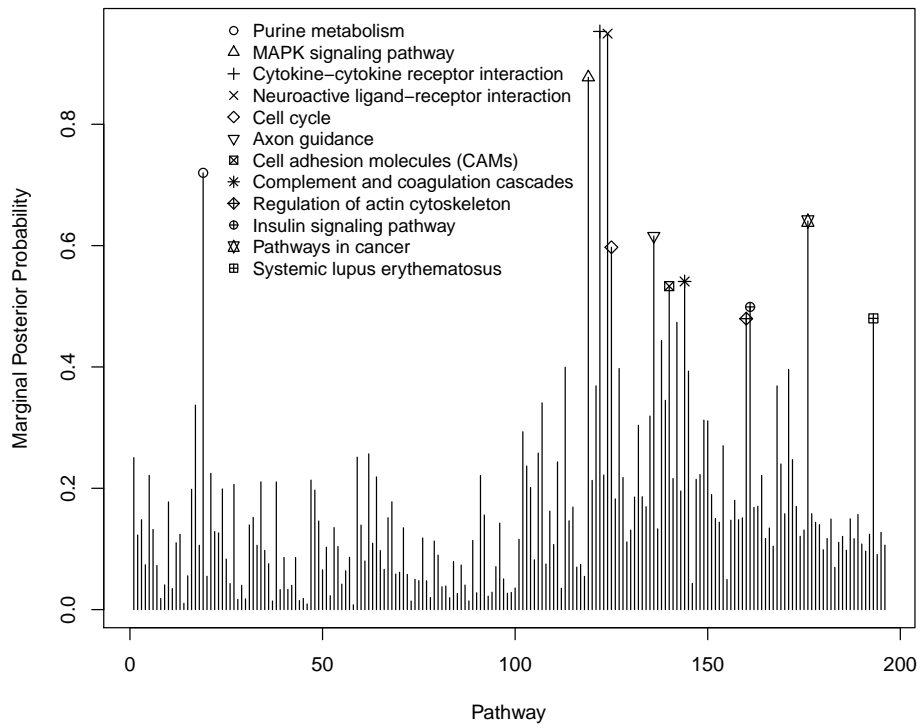
FIG 3. *Microarray data: Marginal posterior probabilities for pathway selection, $p(\theta_k|\boldsymbol{T},Y)$. The 12 pathways with largest probabilities are marked with symbols.*
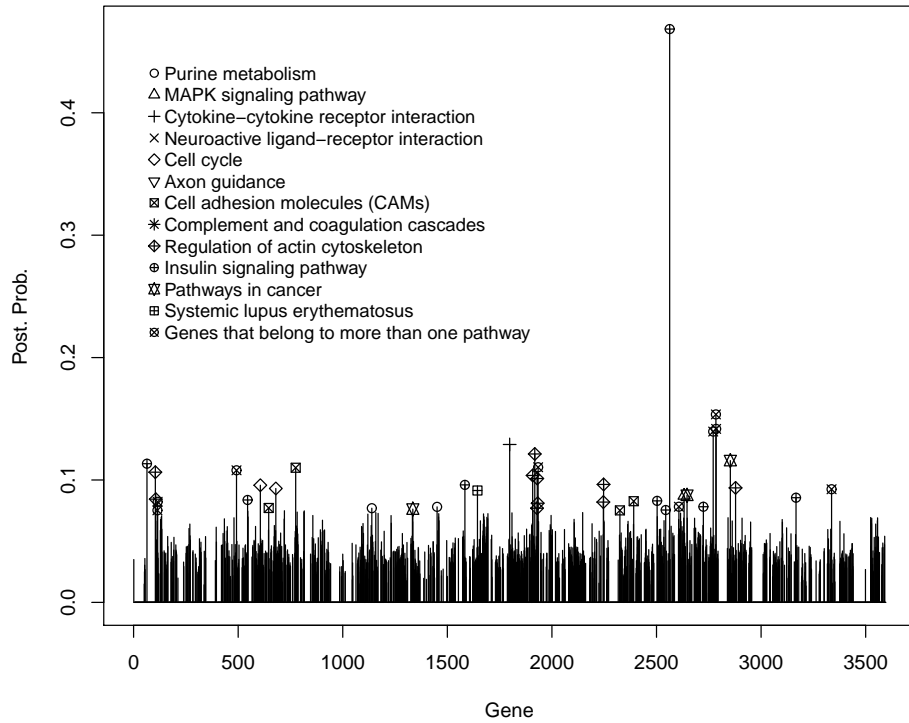
FIG 4. *Microarray data: Conditional posterior probabilities for gene selection,* $p(\gamma_j | \boldsymbol{T}, Y, I\{\sum_{k \in \mathcal{P}} \theta_k s_{kj} > 0\})$. *The 41 probes with largest probability that belong to the 12 selected pathways in Figure 3 are marked with* $\triangle$.
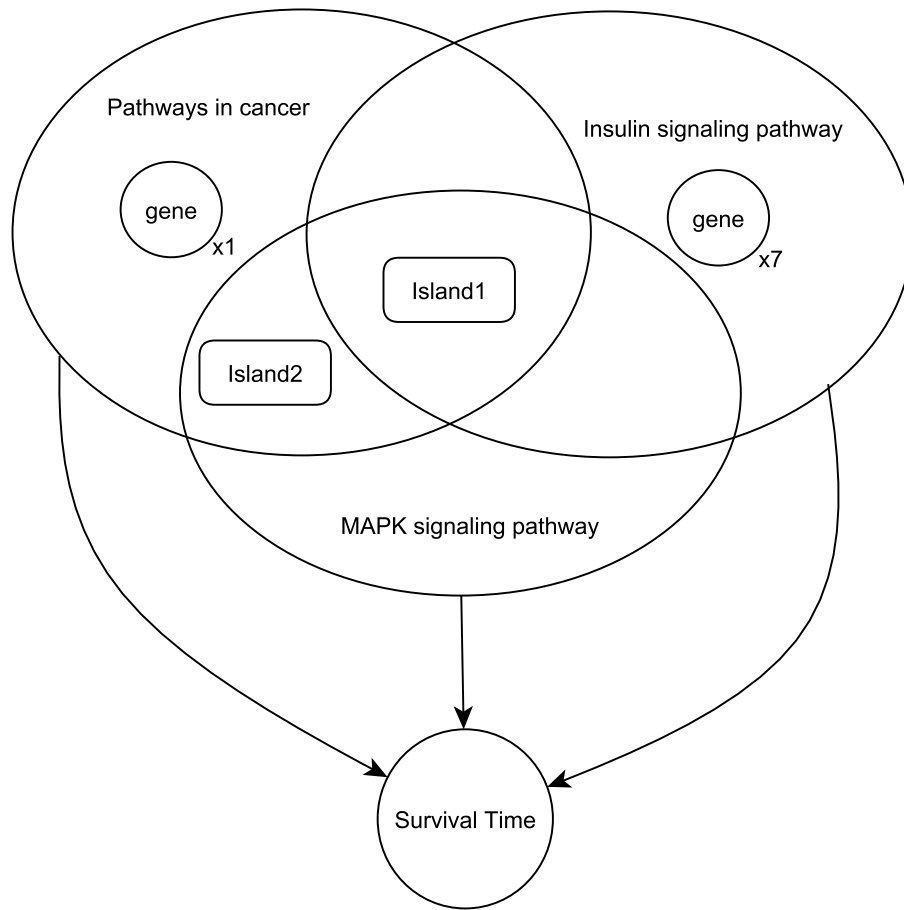
FIG 5. *Microarray data: Graphical representation of a subset of selected pathways with islands and singletons. The genes in the islands are listed in Table 1.*

Singleton genes (no direct neighbor selected)
ACACB (10), C4A (8,12), CALM1 (10), CCNB2 (5), CD4 (7), CDC2 (5), CLDN11 (7), FZD9 (11), GYS2 (10), HIST1H2BN (12), IFNA7 (3), NFASC (7), NRCAM (7), PCK1 (10), PFKP (10), PPARGC1A (10), PXN (9)
Island 1
ACTB (9), ACTG1 (9), ITGA1 (9), ITGA7 (9), ITGB3 (9), ITGB4 (9), ITGB6 (9), ITGB8 (7,10), MYL5 (9), MYL9 (9), PDPK1 (10), PIK3CD (9,10,11), PLA2G4A (2), PLCG1 (11), PRKCA (2,11), PRKY (2,10), PRKY (2,10), PTGS2 (11), SOCS3 (10)
Island 2
ACVR1B (2,3,11), ACVR1B (2,3,11), TGFB3 (2,3,5,11)
Island 3
ENTPD3 (1), GMPS (1)

TABLE 1

The 41 selected genes divided by islands and with associated pathway indices (in parenthesis). The pathway indices correspond to: 1-Purine metabolism, 2-MAPK signaling pathway, 3-Cytokine-cytokine receptor interaction, 4-Neuroactive ligand-receptor interaction, 5-Cell cycle, 6-Axon guidance, 7-Cell adhesion molecules (CAMs), 8-Complement and coagulation cascades, 9-Regulation of actin cytoskeleton, 10-Insulin signaling pathway, 11-Pathways in cancer, 12-Systemic lupus erythematosus.

In addition, several genes with known association to breast cancer were also selected. One of these is Protein kinase C alpha (PKCA), which belongs to the MAPK pathway and the KEGG pathways in cancer. PKCA has been reported to play roles in many different cellular processes, including cell functions associated with breast cancer progression. It has been shown to be overexpressed in some antiestrogen resistant breast cancer cell lines and to be involved in the growth of tamoxifen resistant human breast cancer cells (Frankel et al. 2007). Patients with PKCA-positive tumors have also been shown to have worst survival than patients with PKCA-negative tumors, independently of other factors (Lonne et al. 2010). Prostaglandin-endoperoxide synthase-2 (PTGS2, also known as cyclooxygenase-2 or COX2) has also been related to breast cancer. Denkert et al. (2004) observed COX2 overexpression in breast cancer and strong association with indicators of poor prognosis, such as lymph node metastasis, poor differentiation and large tumor size. This was further confirmed by Gupta et al. (2007), who showed that the expression of COX2 in human breast cancer cells facilitates the assembly of new tumor blood vessels, the release of tumor cells into the circulation, and the breaching of lung capillaries by circulating tumor cells to seed pulmonary metastasis. This is an important finding, as the majority of breast cancer deaths result from metastases rather than from direct effects of the primary tumor itself. Another gene previously shown to be predictive of breast cancer lung metastasis that was also selected by our analysis is integrin, beta-8 (ITGB8) (Landemaine et al. 2008). We

also identified integrin, beta-4 (ITGB4) which regulates key signaling pathways related to carcinoma progression, and has been linked to aggressive tumor behavior and poor prognosis in certain breast cancer subtypes (Guo et al. 2006, Lu et al. 2008).

5. Discussion. We have proposed a model that incorporates biological knowledge from pathway databases into the analysis of DNA microarrays to identify pathways and genes related to a phenotype. Information on pathway membership and gene networks are used to define pathway summaries, specify prior distributions that account for the dependence structure between genes, and define the MCMC moves to fit the model. The gene network prior and the synthesis of the pathway information through PLS bring in additional information that is especially useful in microarray data, where there is low sample size and large measurement error. The performance of the method was evaluated using simulated data and a breast cancer gene expression study with survival outcomes was used to illustrate its application.

Our simulation studies show the effect of the MRF prior on the posterior inference. In general, as expected, the effect of the prior depends on the data and, in particular, on the concordance of the prior network with the data. In our simulations, employing the MRF prior allows us to achieve a better separation of the relevant pathways from those not relevant (in particular, we have found a larger average difference, over three scenarios, between the relevant pathway with the lowest posterior probability and the non relevant pathway with the highest posterior probability). In addition, in the simulated setting with fairly small regression coefficients the model with the MRF prior was able to select all the correct genes without any false positive while the model without MRF includes 3 false positives. Other authors have reported improvements on selection power and sensitivity with respect to commonly used procedures that do not use the pathway structure, with similar, and in some cases, lower false discovery rates. In addition, in our formulation of the model we have used biological information not only for prior specification but also to structure the MCMC moves. This is helpful in arriving at promising models avoiding without visiting invalid configurations. Finally, in real data applications, we have found that employing information on gene-gene networks can lead to the selection of significant genes that would have been missed otherwise, aiding the interpretation of the results, and achieving better prediction results compared to models that do not treat genes as connected elements that work in groups or pathways.

Several MRF priors for gene selection indicators have been proposed in the literature. It is interesting to compare the parametrization of the MRF used in this paper and in Li & Zhang (2010) to the parametrization used in Wei & Li (2007,

2008), where the prior on $\boldsymbol{\gamma}$ is defined as

$$(24) \qquad P(\boldsymbol{\gamma}|\cdot) \propto \exp(d\ n_1 - g\ n_{01}),$$

where n_1 is the number of selected genes and n_{01} is the number of edges linking genes with different values of γ_j, i.e., edges linking included and non-included genes among all pathways,

$$n_1 = \sum_{j=1}^{p} \gamma_j, \qquad n_{01} = \frac{1}{2} \sum_{i=1}^{p} \left[\sum_{j=1}^{p} r_{ij} - \left| \sum_{j=1}^{p} r_{ij}(1 - \gamma_i) - \sum_{j=1}^{p} r_{ij}\gamma_j \right| \right].$$

While d plays the same role as μ in (15), the parametrization using g has a different effect from η on the probability of selection of a gene. This is evident from the conditional probability

$$(25) \qquad P(\gamma_j|\cdot, \gamma_i, i \in N_j) = \frac{\exp(\gamma_j F(\gamma_j))}{1 + \exp(F(\gamma_j))},$$

where $F(\gamma_j) = d + g \sum_{i \in N_j}(2\gamma_i - 1)$. Higher values of g encourage neighboring genes to take on the same γ_j value, and consequently genes with non selected neighbors have lower prior probability of being selected than genes with no neighbors. We felt that parametrization (15) was a better choice for our purposes. First, in a context of sparsity, where only few nodes are supposed to take value 1, a prior that assigns larger probability of inclusion to genes with selected neighbors than to isolated genes seems more appropriate. Second, the exact simulation algorithm of Propp & Wilson (1996) cannot be used to simulate from (24). While any other method of drawing from (24) would be acceptable, as pointed out by Møller et al. (2006), if Markov chain methods are used to sample from a MRF it is then necessary to check, at each step, that the chain has adequately converged to the equilibrium distribution, to avoid introducing additional undesirable stochasticity. On the other hand, one advantage of the parametrization (24) is that there is no phase transition problem associated to the distribution.

Pathway databases are incomplete and the gene network information is often unavailable for many genes. Thus, there may be situations where the dependence structure and the MRF prior specification on the gene selection indicator, $\boldsymbol{\gamma}$, cannot be used for all genes. When the only information available is the pathway membership of genes, the prior on $\boldsymbol{\gamma}$ could be elicited so as to capture other interesting characteristics. For example, we may want a gene to have *a priori* higher probability of being selected when several pathways that contain it are included in the

model. We may also want to avoid favoring the selection of a large pathway just because of its size. In such cases, conditional on $\boldsymbol{\theta}$, independent Bernoulli priors can be specified for γ_j by relating the probability of selection to the proportion of included pathways that contain gene j and adjusting for the pathway sizes, p_k, that is,

$$(26) \qquad \gamma_j|\boldsymbol{\theta} \sim \text{ Bernoulli } \left(c \cdot \frac{\sum_{k=1}^{K} \theta_k s_{kj}/p_k}{\sum_{k=1}^{K} s_{kj}/p_k} \right),$$

where the scalar c is a hyperparameter to be elicited.

In our approach we have made use of PLS components as summary measures of the expression of genes belonging to known pathways and then applied a fully Bayesian approach for the selection of the pathways to be included in the model, and the genes to be included within those pathways. Penalized techniques, including lasso (Tisbhirani 1996), elastic net (Zou & Hastie 2005) and group lasso (Yuan & Lin 2006) have been studied extensively in the literature and have been successfully applied to gene expression data. The group lasso, in particular, defines sets of variables then selects either all the variables in the group or none of them. Recently, a modification of the method was proposed by Friedman et al. (2010) using a more general penalty that yields sparsity at both the group and individual feature levels to select groups and predictors within each group. Our understanding of group lasso is that the method works best in situations where variables belonging to the same group are highly correlated while covariates in different groupings do not exhibit high correlation. However, genes belonging to the same pathway often do not exhibit high correlation in their expression levels. Also, in our case there are genes belonging to different pathways that have high correlation, as well as genes that belong to more than one pathway. It is also worth mentioning that attention has been recently devoted to Bayesian formulations of LASSO techniques and to a comparison of their performances. Initial investigations suggest that, in terms of prediction mean squared error, Bayesian lasso methods perform similarly to and, in some cases, better than the frequentist lasso (see for example Kyung et al. 2010). Particularly relevant to the approach presented in our paper is the work of Guan & Stephens (2009), who apply Bayesian variable selection (BVS) and stochastic search methods in a regression model for genome-wide data. In simulations they find that, in spite of the apparent computational challenges, the BVS method produces better power and predictive performance compared with standard lasso techniques.

References.

Albert, J. & Chib, S. (1993), 'Bayesian analysis of binary and polychotomous response data', *Journal of American Statistical Association* **88**, 669–679.

Bair, E., Hastie, T., Paul, D., & Tibshirani, R. (2006), 'Prediction by supervised principal components', *Journal of the American Statistical Association* **101**, 119–137.

Besag, J. (1974), 'Spatial interaction and the statistical analysis of lattice systems', *Journal of the Royal Statistical Society, Ser. B* **36**, 192–225.

Bild, A. H., Yao, G., Chang, J. T., Wang, Q., Potti, A., Chasse, D., Joshi, M., Harpole, D., Lancaster, J. M., Berchuck, A., Jr, J. A. O., Marks, J. R., Dressman, H. K., West, M. & Nevins, J. R. (2006), 'Oncogenic pathway signatures in human cancers as a guide to targeted therapies', *Nature* **439**, 353–357.

Boulesteix, A. & Strimmer, K. (2007), 'Partial least squares: a versatile tool for the analysis of high-dimensional genomic data', *Briefings in Bioinformatics* **8(1)**, 32–44.

Brown, P. J., Vannucci, M. & Fearn, T. (1998), 'Multivariate Bayesian variable selection and prediction', *Journal of the Royal Statistical Society, Series B* **60**, 627–641.

Chipman, H., George, E. & R.E. McCulloch, R. (2001), *The Practical Implementation of Bayesian Model Selection*, IMS Lecture Notes - Monograph Series Volume 38.

Dahlquist, K., Salomonis, N., Vranizan, K., Lawlor, S. & Conklin, B. (2002), 'GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways', *Nature Genetics* **31**, 19–20.

Denkert, C., Winzer, K. & Hauptmann, S. (2004), 'Prognostic impact of cyclooxygenase-2 in breast cancer', *Clinical Breast Cancer* **4**, 428–33.

Doniger, S., Salomonis, N., Dahlquist, K., Vranizan, K., Lawlor, S. & Conklin, B. (2003), 'MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile for microarray data', *Genome Biology* **41**, R7.

Downward, J. (2006), 'Signatures guide drug choice', *Nature* **439**, 274–275.

Frankel, L., Lykkesfeldt, A., Hansen, J. & Stenvang, J. (2007), 'Protein Kinase C alpha is a marker for antiestrogen resistance and is involved in the growth of tamoxifen resistant human breast cancer cells', *Breast Cancer Res. Treat.* **104**, 165–179.

Friedman, J., Hastie, T. & Tibshirani, R. (2010), A note on the group lasso and a sparse group lasso, Technical report, Dept of Stat, Stanford University.

George, E. I. & McCulloch, R. E. (1997), 'Approaches for Bayesian variable selection', *Statistica Sinica* **7**, 339–373.

George, E. & McCulloch, R. (1993), 'Variable selection via Gibbs sampling', *J. Am. Statist. Assoc.* **88**, 881–9.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. & Lander, E. (1999), 'Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring', *Science* **286**, 531–537.

Guan, Y. & Stephens, M. (2009), 'Bayesian variable selection regression for genome-wide association studies, and other large-scale problems', *submitted* .

Guo, W., Pylayeva, Y., Pepe, A., Yoshioka, T., Muller, W., Inghirami, G. & Giancotti, F. (2006), 'beta4 integrin amplifies erbb2 signaling to promote mammary tumorigenesis', *Cell* **126**, 489502.

Gupta, G., Nguyen, D., Chiang, A., Bos, P., Kim, J., Nadal, C., Gomis, R., Manova-Todorova, K. & Massague, J. (2007), 'Mediators of vascular remodelling co-opted for sequential steps in lung metastasis', *Nature* **446**, 765–770.

Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, G., Wu, G., Matthews, L., Lewis, S., Birney, E. & Stein, L. (2005), 'Reactome: a knowledgebase of biological pathways', *Nucleic Acids Research* **33**, D428–32.

Kanehisa, M. & Goto, S. (2000), 'Kyoto encyclopedia of genes and genomes', *Nucleic Acid Res* **28**, 27–30.

Keyse, S. (2008), 'Dual-specificity MAP kinase phosphatases (MKPs) and cancer', *Cancer and Metastasis Reviews* **27**, 253–261.

Krieger, C., Zhang, P., Mueller, L., Wang, A., Paley, S., Arnaud, M., Pick, J., Rhee, S. & Karp, P.

(2004), 'MetaCyc: a multiorganism database of metabolic pathways and enzymes', *Nucleic Acids Research* **32**, D438–442.

Kwon, D., Tadesse, M., Sha, N., Pfeiffer, R. & Vannucci, M. (2007), 'Identifying biomarkers from mass spectrometry data with ordinal outcome', *Cancer Informatics* **3**, 19–28.

Kyung, M., Gill, J., Ghosh, M. & Casella, G. (2010), 'Penalized regression, standard errors, and bayesian lassos', *Bayesian Analysis* **5(2)**, 369–412.

Landemaine, T., Jackson, A., Bellahcene, A., Rucci, N., Sin, S., Abad, B., Sierra, A., Boudinet, A., Guinebretiere, J.-M., Ricevuto, E., Nogues, C., Briffod, M., Bieche, I., Cherel, P., Garcia, T., Castronovo, V., Teti, A., Lidereau, R. & Driouch, K. (2008), 'A six-gene signature predicting breast cancer lung metastasis', *Cancer Research* **68**, 60929.

Lee, S., Jeong, Y., Im., H. G., Kim, C., Chang, Y. & Lee, I. (2007), 'Silibinin suppresses PMA-induced MMP-9 expression by blocking the AP-1 activation via MAPK signaling pathways in MCF-7 human breast carcinoma cells', *Biochemical and Biophysical Research Communications* **354(1)**, 65–171.

Li, C. & Li, H. (2008), 'Network-constrained regularization and variable selection for analysis of genomics data', *Bioinformatics* **24**, 1175–1182.

Li, F. & Zhang, N. (2010), 'Bayesian variable selection in structured high-dimensional covariate space with application in genomics', *Journal of the American Statistical Association* **to appear**.

Lindgren, F., Geladi, P. & Wold, S. (1993), 'The kernel algorithm of PLS', *Journal of Chemometrics* **7**, 45–59.

Lonne, G., Cornmark, L., Zahirovic, I., Landberg, G., Jirström, K. & Larsson, C. (2010), 'PKCalpha expression is a marker for breast cancer aggressiveness', *Molecular Cancer* **14(9)**, 76.

Lu, S., Simin, K., Khan, A. & Mercurio, A. (2008), 'Analysis of integrin beta4 expression in human breast cancer: association with basal-like tumour and prognostic significance', *Clinical Cancer Research* **14**, 1050–8.

Lucas, J., Carvalho, C., Wang, Q., Bild, A., Nevins, J. & West, M. (2006), Sparse statistical modelling in gene expression genomics, *in* K. Do, P. Mueller & M. Vannucci, eds, 'Bayesian Inference for Gene Expression and Proteomics', Cambridge University Press, pp. 155–176.

Møller, J., Pettitt, A., Reeves, R. & Berthelsen, K. (2006), 'An efficient markov chain monte carlo method for distributions with intractable normalising constants', *Biometrika* **92 (2)**, 451–458.

Nakao, M., Bono, H., Kawashima, S., Kamiya, T., Sato, K., Goto, S. & Kanehisa, M. (1999), 'Genome-scale gene expression analysis and pathway reconstruction in KEGG', *Genome Informatics Series: Workshop on Genome Informatics* **10**, 94–103.

Pan, W., Xie, B. & Shen, X. (2009), 'Incorporating predictor network in penalized regression with application to microarray data', *Biometrics* p. to appear.

Park, M. Y., Hastie, T. & Tibshirani, R. (2007), 'Averaged gene expressions for regression', *Biostatistics* **8**, 212–27.

Pittman, J., Huang, E., Dressman, H., Horng, C., Cheng, S., Tsou, M., Chen, C., Bild, A., Iversen, E., Huang, A., Nevins, J. & West, M. (2004), 'Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes', *Proceedings of the National Academy of Sciences* **101**, 8431–6.

Propp, J. & Wilson, D. (1996), 'Exact sampling with coupled markov chains and applications to statistical mechanics', *Random Structures and Algorithms* **9 (1)**, 223–252.

Robert, C. & Casella, G. (2004), *Monte Carlo Statistical Methods*, Springer Verlag.

Sha, N., Tadesse, M. & Vannucci, M. (2006), 'Bayesian variable selection for the analysis of microarray data with censored outcomes', *Bioinformatics* **22**, 2262–2268.

Sha, N., Vannucci, M., Tadesse, M. G., Brown, P. J., Dragoni, I., Davies, N., Roberts, T. C., Contestabile, A., Salmon, N., Buckley, C. & Falciani, F. (2004), 'Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage', *Biometrics* **60**, 812–9.

Shipp, M. A., Ross, K. N., Tamayo, P., Weng, A. P., Kutok, J. L., Aguiar, R. C., Gaasenbeek, M.,

Angelo, M., Reich, M., Pinkus, G. S. & *e*t al. (2002), 'Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning', *Nature Medicine* **8**, 68–74.

Solit, D. B., Garraway, L. A., Pratilas, C. A., Sawai, A., Getz, G., Basso, A., Ye, Q., Lobo, J. M., She, Y., Osman, I., Golub, T. R., Sebolt-Leopold, J., Sellers, W. R. & Rosen, N. (2006), 'BRAF mutation predicts sensitivity to MEK inhibition', *Nature* **439**, 358–362.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S. & Mesirov, J. P. (2005), 'Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles', *Proceedings of the National Academy of Sciences* **102**, 15545–50.

Telesca, D., Muller, P., Parmigiani, G. & Freedman, R. (2008), Modeling dependent gene expression, Technical report, University of Texas M.D. Anderson Cancer Center, Department of Biostatistics.

The Gene Ontology Consortium (2000), 'Gene Ontology: tool for the unification of biology', *Nature Genetics* **25**, 25–29.

Tisbhirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society, Series B* **58**, 267–288.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. & Altman, R. (2001), 'Missing value estimation methods for DNA microarrays', *Bioinformatics* **17**, 520–525.

van't Veer, L., Dai, H., van de Vijver, M., He, Y., Hart, A., Mao, M., Peterse, H., van der Kooy, K., Marton, M., Witteveen, A., Schreiber, G., Kerkhoven, R., Roberts, C., Linsley, P., Bernards, R. & Friend, S. (2002), 'Gene expression profiling predicts clinical outcome of breast cancer', *Nature* **415**, 530–536.

Wei, L. (1992), 'The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis', *Statistics in Medicine* **11**, 1871–1879.

Wei, Z. & Li, H. (2007), 'A Markov random field model for network-based analysis of genomic data', *Bioinformatics* **23**, 15371544.

Wei, Z. & 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* **02(1)**, 408–429.

Yuan, M. & Lin, Y. (2006), 'Model selection and estimation in regression with grouped variables', *Journal of the Royal Statistical Society, Series B* **68**, 49–67.

Zhang, J. & Wiemann, S. (2009), 'KEGGgraph: a graph approach to KEGG PATHWAY in R and Bioconductor', *Bioinformatics* **25**, 1470–1471.

Zou, H. & Hastie, T. (2005), 'Regularization and variable selection via the elastic net', *J. Royal Statistical Society, Series B.* **67**, 301–320.

Zou, H., Hastie, T. & Tibshirani, R. (2006), 'Sparse principal component analysis', *Journal of Computational and Graphical Statistics* **15**, 265–286.

APPENDIX A: MCMC SCHEME FOR SAMPLING $(\boldsymbol{\theta}, \boldsymbol{\gamma})$

We now describe the MCMC steps for $(\boldsymbol{\theta}, \boldsymbol{\gamma})$ in more detail. As previously described, no empty pathways or orphan genes are proposed during sampling and, for identifiability, selecting the same set of genes for different pathways is not allowed. At each iteration, only one pathway and/or a gene are proposed to be added or removed.

(1) Change inclusion status of both gene and pathway – randomly choose between addition (move 1.i) or removal (move 1.ii):

(1.i) Add a pathway and a gene:
First select a pathway that is not included in the model and has none of its member genes in the model ($\theta_k^{old} = 0$ and $p_{k\gamma}^{old} = 0$). Randomly choose one gene from the pathway ($\gamma_j^{old} = 0$) and propose including both the pathway and the gene, i.e., set $\theta_k^{new} = 1$, $\gamma_j^{new} = 1$. The move is accepted with probability

(27)
$$\min\left\{1, \frac{f(\boldsymbol{\theta}^{new}, \boldsymbol{\gamma}^{new}|\boldsymbol{T}, Y)}{f(\boldsymbol{\theta}^{old}, \boldsymbol{\gamma}^{old}|\boldsymbol{T}, Y)} \cdot \frac{p_k \cdot \sum_{r=1}^{K} I\{\theta_r^{old} = 0, p_{r\gamma}^{old} = 0\}}{\sum_{r=1}^{K} I\{\theta_r^{new} = 1, p_{r\gamma}^{new} = 1, cond1, condId1\}}\right\},$$

where *cond1* and *condId1* are explained in move type (1.ii) below.

(1.ii) Remove a pathway and a gene:
This move is the reverse of (1.i) described above. We first select a pathway that is included in the model and has only one of its member genes in the model ($\theta_k^{old} = 1$ and $p_{k\gamma}^{old} = 1$). In addition, this included gene ($\gamma_j^{old} = 1$) may not be the sole representative for other included pathways, to ensure that no empty pathway is created. Furthermore, identical sets of genes from different selected pathways cannot be created. These constraints correspond, respectively, to *cond1* and *condId1* in the proposal ratios (27) and (28). We attempt to remove both the pathway and the gene, i.e., set $\theta_k^{new} = 0$, $\gamma_j^{new} = 0$ and accept the move with probability

(28)
$$\min\left\{1, \frac{f(\boldsymbol{\theta}^{new}, \boldsymbol{\gamma}^{new}|\boldsymbol{T}, Y)}{f(\boldsymbol{\theta}^{old}, \boldsymbol{\gamma}^{old}|\boldsymbol{T}, Y)} \cdot \frac{\sum_{r=1}^{K} I\{\theta_r^{old} = 1, p_{r\gamma}^{old} = 1, cond1, condId1\}}{p_k \cdot \sum_{r=1}^{K} I\{\theta_r^{new} = 0, p_{r\gamma}^{new} = 0\}}\right\},$$

(2) Change the inclusion status of gene but not pathway – randomly choose between addition (2.i) or removal (2.ii):

(2.i) Add a gene in an already included pathway:
First select a pathway already included in the model and that has some member genes that could potentially be added ($\theta_k^{old} = 1$ and $p_k > p_{k\gamma}^{old}$). Let G be the set of pathways that satisfy these conditions. Choose one of the non-included genes from this pathway ($\gamma_j^{old} = 0$) and attempt to add it, i.e, set $\theta_k^{new} = \theta_k^{old} = 1$, $\gamma^{new} = 1$. The proposal is accepted with probability

(29)

$$\min\left\{1, \frac{f(\boldsymbol{\theta}^{new}, \boldsymbol{\gamma}^{new}|\boldsymbol{T}, Y)}{f(\boldsymbol{\theta}^{old}, \boldsymbol{\gamma}^{old}|\boldsymbol{T}, Y)} \cdot \frac{\sum_{r=1}^{K} I\{\theta_r^{old} = 1, p_r > p_{r\gamma}^{old}\} \cdot \sum_{r\in G} \frac{1}{p_{r\gamma}^{new(cond2\gamma, condId2\gamma)}}}{\sum_{r=1}^{K} I\{\theta_r^{new} = 1, p_{r\gamma}^{new} > 1, cond2_\theta, condId2_\theta\} \cdot \sum_{r\in G} \frac{1}{p_r - p_{r\gamma}^{old}}}\right\},$$

where 'cond2$_\theta$', 'cond2$_\gamma$', 'condId2$_\theta$' and 'condId2$_\gamma$' are explained in move type (2.ii) below.

(2.ii) Remove a gene from an already included pathway:
This move is the reverse of (2.i) described above. We first select a pathway already included in the model and that has more than one of its member genes included in the model ($\theta_k^{old} = 1, p_{k\gamma}^{old} > 1$). In addition, at least one of the included genes from this pathway may not be the sole representative for other included pathways and its removal would not create an identifiability problem – this corresponds to constraints 'cond2$_\theta$' and 'condId2$_\theta$'in the proposal ratios of (29) and (30). Once the pathway is selected, choose a gene among the eligible candidates, that is, an included member gene ($\gamma_j^{old} = 1$) which is not the sole representative for other included pathways and whose removal does not create an identifiability problem – this corresponds to constraints 'cond2$_\gamma$' and 'condId2$_\gamma$'. Constraints 'cond2$_\theta$' for pathways, and 'cond2$_\gamma$' for genes, will ensure that no empty pathways are created after the proposed move. Leave the pathway status unchanged and attempt to remove the selected gene, i.e., set $\theta_k^{new} = \theta_k^{old} = 1$, $\gamma_j^{new} = 0$. The proposed move is accepted with probability

(30)

$$\min\left\{1, \frac{f(\boldsymbol{\theta}^{new}, \boldsymbol{\gamma}^{new}|\boldsymbol{T}, Y)}{f(\boldsymbol{\theta}^{old}, \boldsymbol{\gamma}^{old}|\boldsymbol{T}, Y)} \cdot \frac{\sum_{r=1}^{K} I\{\theta_r^{old} = 1, p_{r\gamma}^{old} > 1, cond2_\theta, condId2_\theta\} \cdot \sum_{r\in G} \frac{1}{p_r - p_{r\gamma}^{new}}}{\sum_{r=1}^{K} I\{\theta_r^{new} = 1, p_r > p_{r\gamma}^{new}\} \cdot \sum_{r\in G} \frac{1}{p_{r\gamma}^{old(cond2\gamma, condId2\gamma)}}}\right\}.$$

(3) Change inclusion status of pathway but not gene – randomly choose between addition (3.i) or removal (3.ii):

(3.i) Add a pathway but leave genes' status unchanged:
First select a pathway that is not included in the model but has some of its member genes included in the model through other pathways ($\theta_k^{old} = 0$ and $p_{k\gamma}^{old} \geq 1$). Attempt to add the pathway but leave the status of its member genes unchanged, i.e., set $\theta_k^{new} = 1$. The proposed move is accepted with probability

(31)
$$\min\left\{1, \frac{f(\boldsymbol{\theta}^{new}, \boldsymbol{\gamma}^{new}|\boldsymbol{T}, Y)}{f(\boldsymbol{\theta}^{old}, \boldsymbol{\gamma}^{old}|\boldsymbol{T}, Y)} \cdot \frac{\sum_{r=1}^{K} I\{\theta_r^{old} = 0, p_{r\gamma}^{old} \geq 1, condId3\}}{\sum_{r=1}^{K} I\{\theta_r^{new} = 1, p_{r\gamma}^{new} \geq 1, cond3\}}\right\},$$

where *condId3* means that it is not possible to select a pathway whose selected genes form the entire set of selected genes for another selected pathway, and *cond3* is explained in move type (3.ii) below.

(3.ii) Remove a pathway but leave genes' status unchanged:
This move is the reverse of (3.i) described above. First select a pathway included in the model that has all of its $p_{k\gamma}^{old}$ included member genes associated with other included pathways ($\theta_k^{old} = 1$ and 'cond3'). This will ensure that no orphan gene is created. Attempt to remove the pathway but leave the status of the genes unchanged, i.e., set $\theta_k^{new} = 0$ and accept the move with probability

(32)
$$\min\left\{1, \frac{f(\boldsymbol{\theta}^{new}, \boldsymbol{\gamma}^{new}|\boldsymbol{T}, Y)}{f(\boldsymbol{\theta}^{old}, \boldsymbol{\gamma}^{old}|\boldsymbol{T}, Y)} \cdot \frac{\sum_{r=1}^{K} I\{\theta_r^{old} = 1, p_{r\gamma}^{old} \geq 1, cond3\}}{\sum_{r=1}^{K} I\{\theta_r^{new} = 0, p_{r\gamma}^{new} \geq 1, condId3\}}\right\}.$$

It is easy to see that our Bayesian stochastic search variable selection kernel generates an ergodic Markov chain over the restricted space. First note that the chain produced by our MCMC has the following properties:

- It is aperiodic and has an invariant probability distribution (by definition of the M-H kernel);
- It is irreducible (noting that every move is equipped with its reverse, that it is possible to reach any valid configuration in the parameter space starting from the configuration where no pathways and no genes are selected, and that the probability of moving will never be zero, i.e., the probability of moving from one point to another in n steps is bigger then zero);
- Properties above imply that the chain is recurrent and therefore ergodic.

Let $\operatorname{supp} q(\cdot|x)$ indicate the support of our proposal distribution, i.e., the set of possible configurations that can be generated from the previously visited configuration x. We say that the supports for two different x's are connected if they share at least one configuration. We need to check that the union of all connected supports is equal to the entire support of the posterior distribution, i.e,

$$\bigcup_{x \in \operatorname{supp} f} \operatorname{supp} q(\cdot|x) \supset \operatorname{supp} f$$

where f is the target density and $q(\cdot|x)$ the proposal distribution (Robert & Casella 2004). It is easy to verify that our MCMC satisfies this condition because, starting from the configuration where no pathways or genes are selected, it is possible to reach every admissible configuration, and because the union of these points is exactly equal to the support of the target distribution.

MOFFITT CANCER CENTER
TAMPA, FL 33612, USA.
E-MAIL: Ann.Chen@moffitt.org

DEPARTMENT OF STATISTICS
RICE UNIVERSITY
HOUSTON, TX 77005, USA.
E-MAIL: marina@rice.edu

DEPARTMENT OF MATHEMATICS & STATISTICS
GEORGETOWN UNIVERSITY
WASHINGTON, DC 20057, USA.
E-MAIL: mgt26@georgetown.edu