Libertas Academica
FREEDOM TO RESEARCH

Open Access: Full open access to this and thousands of other papers at http://www.la-press.com.

Cancer Informatics

Supplementary Issue: Classification, Predictive Modelling, and Statistical Analysis of Cancer Data (A)

Bayesian Joint Selection of Genes and Pathways: Applications in Multiple Myeloma Genomics

Lin Zhang[1], Jeffrey S. Morris[2], Jiexin Zhang[3], Robert Z. Orlowski[4] and Veerabhadran Baladandayuthapani[5]

[1]Postdoctoral fellow, Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA. [2]Professor, Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA. [3]Principal Statistical Analyst, Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA. [4]Professor, Department of Lymphoma & Myeloma, and of Experimental Therapeutics, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA. [5]Associate Professor, Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA.

ABSTRACT: It is well-established that the development of a disease, especially cancer, is a complex process that results from the joint effects of multiple genes involved in various molecular signaling pathways. In this article, we propose methods to discover genes and molecular pathways significantly associated with clinical outcomes in cancer samples. We exploit the natural hierarchal structure of genes related to a given pathway as a group of interacting genes to conduct selection of both pathways and genes. We posit the problem in a hierarchical structured variable selection (HSVS) framework to analyze the corresponding gene expression data. HSVS methods conduct simultaneous variable selection at the pathway (group level) and the gene (within-group) level. To adapt to the overlapping group structure present in the pathway–gene hierarchy of the data, we developed an overlap-HSVS method that introduces latent partial effect variables that partition the marginal effect of the covariates and corresponding weights for a proportional shrinkage of the partial effects. Combining gene expression data with prior pathway information from the KEGG databases, we identified several gene–pathway combinations that are significantly associated with clinical outcomes of multiple myeloma. Biological discoveries support this relationship for the pathways and the corresponding genes we identified.

KEYWORDS: Bayesian variable selection, hierarchical variable selection, multiple myeloma, overlapping group

SUPPLEMENT: Classification, Predictive Modelling, and Statistical Analysis of Cancer Data (A)

CITATION: Zhang et al. Bayesian Joint Selection of Genes and Pathways: Applications in Multiple Myeloma Genomics. *Cancer Informatics* 2014:13(S2) 113–123 doi: 10.4137/CIN.S13787.

RECEIVED: March 31, 2014. **RESUBMITTED:** July 16, 2014. **ACCEPTED FOR PUBLICATION:** July 17, 2014.

ACADEMIC EDITOR: JT Efird, Editor in Chief

TYPE: Methodology

FUNDING: The research of VB and JSM is partially supported by NIH grant R01 CA160736 and the Cancer Center Support Grant (CCSG) (P30 CA016672). VB and RZO were also partially supported by the The MD Anderson Cancer Center SPORE in Multiple Myeloma (P50 CA142509) and The MD Anderson Cancer Center Support Grant (P30 CA016672). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute or the National Institutes of Health. The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: lzhang9@mdanderson.org

Paper subject to independent expert blind peer review by minimum of two reviewers. All editorial decisions made by independent academic editor. Prior to publication all authors have given signed confirmation of agreement to article publication and compliance with all applicable ethical and legal requirements, including the accuracy of author and contributor information, disclosure of competing interests and funding sources, compliance with ethical requirements relating to human and animal study participants, and compliance with any copyright requirements of third parties.

Introduction

Multiple myeloma (MM) is a cancer of plasma cells, which is the second most common hematological malignancy in the United States.[1] It is characterized by malignant, neoplastic transformation of terminally differentiated B cells in the bone marrow known as plasma cells, the principal function of which is to produce antibodies, also known as immunoglobulins, which play an important role in immune surveillance. Immunoglobulins are normally composed of small molecules known as heavy chains and light chains. There are five types of heavy chains, immunoglobulin G (IgG), IgA, IgM, IgD, and IgE, and two types of light chains, kappa and lambda, each combination forming one type of immunoglobulin complex. When a single abnormal clone of plasma cells results in an excessive number of light chains, these do not attach to the heavy chains to form the normal immunoglobulin complex, but rather enter the bloodstream as unattached light chains (and thus are labeled as serum free light chains). Myeloma progression is often seen when one type of immunoglobulin is excessively produced, causing a monoclonal protein spike.[2] Correspondingly, there will often be a large amount of one type of light chain (kappa or lambda) produced as a consequence, leading to an abnormally increased or decreased value in the free light chain (kappa/lambda) ratio in the serum. Hence, this ratio is an important indicator for the diagnosis, monitoring, and prognosis of MM.[2–4] The degree to which the ratio deviates from the normal range indicates the extent of monoclonal gammopathy, which relates to the severity of MM. Therefore, the identification of specific genomic markers with expression levels that are associated with the extent of monoclonal gammopathy could potentially elucidate the molecular mechanisms underlying the progression of MM.

Advances in microarray technologies have increased the availability of high-throughput gene expression datasets, allowing for genome-wide investigations of molecular activities underlying diseases, including MM. A common interest in such studies is the identification of relevant genetic markers, eg, genes, that are associated with the development or progression of diseases. Traditional studies have mainly relied on univariate analysis, in which each gene is modeled independently, as in the work of Dryja[5] and Golub et al.[6] among others. However, we assume that the development of a disease is a complex process that results from the joint effects of multiple genes. Thus, it is of great interest to model the joint effects of the expression levels of genes over the whole genome as measured by microarray assays, and select genes whose expression levels exhibit significant associations with the clinical outcomes of patients with a specific disease or condition, such as MM. This is essentially a problem of *variable or feature selection*.

Inferential challenges for variable selection based on gene expression datasets from microarray assays include not only high-dimensionality (relative to sample size), but also the presence of a structured hierarchy induced by biological mechanisms. Genes typically do not influence the disease state by themselves, but act through their involved pathway(s), which allows us to consider genes related to a given pathway as a natural group of interacting genes. Studies have indicated that although many genes may be related to a complex disease such as cancer, relatively few pathways play a role in cancer development.[7] In addition, therapeutic interventions based on the inhibition of targeted pathways have been approved by the U.S. Food and Drug Administration for a variety of cancer types.[8] Hence, it is of equal interest for us to identify significant pathways as well as individual genes that are associated with the clinical outcomes of cancers. The Kyoto Encyclopedia of Genes and Genomes (KEGG) database is a popular public database that provides information on discovered pathways and their involved genes.[9] The pathway information available from the KEGG database allows us to assign genes into groups based on the specific pathways in which they are involved, and conduct analyses at the pathway level. In this article, we aim to select both pathways and individual genes mapped to these pathways that are significantly associated with clinical outcomes of MM based on gene expression profiles obtained from microarray assays.

The problem of simultaneously selecting MM-associated pathways and genes within those pathways involves variable selection at two hierarchical levels: the group level (pathways) and the within-group level (genes within a pathway). Many variable selection methods have recently been developed to incorporate grouping structures in datasets and conduct variable selection at the group level. Yuan and Lin[10] proposed the group lasso method, in which a lasso penalty function is applied to the L_2-norm of the coefficients within each group. This method was subsequently extended by Raman et al.[11] for the Bayesian setting. Zhao et al.[12] generalized the group lasso method by replacing the L_2-norm of the coefficients in each group with the L_γ-norm for $1 < \gamma \leq \inf$. In the extreme case where $\gamma = 1$, the coefficient estimates within a group are encouraged to be exactly the same. These model selection methods focus on group selection without due consideration of selection at the within-group level; that is, they only allow the variables within a group to be "all in or all out" of the model. More recently, some methods have been developed in the Bayesian and frequentist frameworks that apply to making selections at both the group and within-group levels. Wang et al.[13] reparameterized the predictor coefficients and selected variables by maximizing the penalized likelihood with two penalizing terms. Ma et al.[14] proposed a clustering threshold gradient-directed regularization method for genetic association studies. Stingo et al.[15] used two sets of binary indicators for making respective selections at the group and within-group levels. Zhang et al.[16] developed a Bayesian approach, the hierarchical structured variable selection (HSVS) method, which utilizes a generalized "spike and slab" selection prior for simultaneous group selection and within-group shrinkage. They demonstrated through simulations a superior performance of the HSVS method in high-dimensional data analysis as a strong variable selector at both the group and within-group levels.

In this paper, we generalize the HSVS method of Zhang et al.[16] and apply it to our gene expression datasets from MM cell samples to identify significant pathways as well as genes within those pathways that are associated with the free light chain kappa/lambda ratio in serum while controlling for the demographic/clinical covariates. We recognize that individual genes may play multiple roles in cellular functions and belong to more than one biological pathway, as illustrated in Figure 1, ie, two groups of variables may overlap. Thus, we generalize the HSVS method to a more flexible overlap-HSVS method to accommodate the overlapping group structures in our pathway-based data analysis. Similar to our process for employing the HSVS method, we employ a selection prior imposed on a latent binary indicator for each group for the group-level selection, combined with a robust shrinkage distribution, such as a Laplace prior, for each coefficient in a group to represent the within-group-level shrinkage. For an individual variable that is mapped to multiple pathways, we introduce latent variables to represent the parts of the regression of the variable on the response that contribute through different pathways, with the shrinkage strength on each partial effect proportional to the regularization parameter of the corresponding pathway. Although the latent variables representing the partial effects of the variable on the response through each pathway are unidentifiable, the marginal effect of the variable, which equals the sum of the latent variables, can be uniquely estimated, which provides coherent inference. In addition, the proportional shrinkage on each part leads to a relatively stable Bayesian estimation of the model. We used the overlap-HSVS method to analyze the MM dataset of 208 patients, and identified both pathways and genes within the pathways that are potentially associated with the progression of MM.

The rest of the paper is structured as follows. In Methods, we introduce the overlap-HSVS hierarchical model in detail and describe the algorithm for posterior inference. We present the data and the results of our application of the overlap-HSVS method to the MM genetic association study in detail in the subsequent sections.

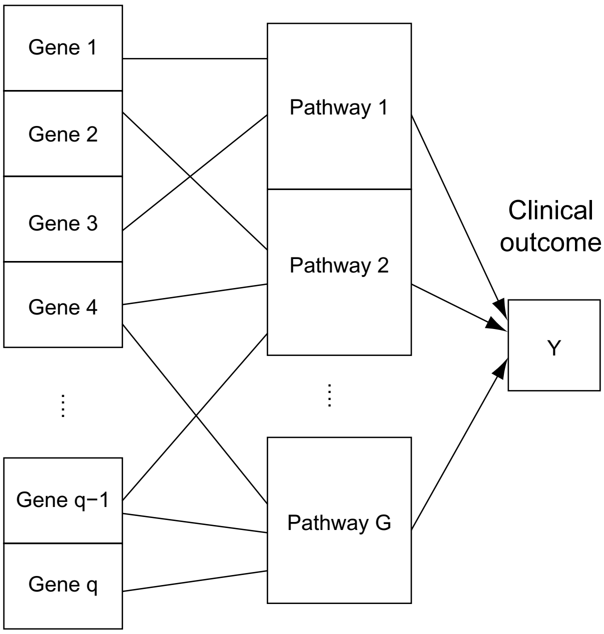

Figure 1. Schematic plot showing the overlapping group structures present in the gene expression data. Each gene in the left column can belong to one or multiple pathways, the activities of which are associated with the clinical outcome.

Methods

Notations. Let $\boldsymbol{Y} = (y_1,\ldots,y_n)^T$ denote the independent observations of the continuous clinical outcomes/responses of interest from n patients/samples and X denote the $n \times q$-dimensional covariate matrix of the gene expression profiles with

$$X = \left(x_1,\ldots,x_q\right) = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1q} \\ x_{21} & x_{22} & \cdots & x_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nq} \end{pmatrix},$$

where x_{ij} denotes the expression level of the jth gene for the ith subject.

Coding of pathway information. Let G be the total number of groups/pathways to which the q explanatory covariates/genes belong, where grouping of the genes is based on the pathway information obtained via KEGG. We use $\mathcal{G}_j = \{g1(j),\cdots,g_{k_j}(j)\}$ to denote the set of pathway indexes that include gene j as a group member for $j = 1,\ldots,q$, and $\Psi_g = \{j{:}g \in \Gamma_j\}$ to denote the set of genes that lie in the gth pathway for $g = 1,\ldots,G$.

Bayesian hierarchical model for gene/pathway selection. We assume that the continuous responses relate to the covariates through the linear regression model

$$Y = Z\alpha + X\beta + \varepsilon,$$

where Z denotes the demographic covariates, age and gender, in our MM dataset, with their associated parameters α, and X are the genomic predictors with associated parameters $\boldsymbol{\beta}$. When the q variables can be partitioned into G groups, ie, each variable j belongs to only one group $g(j)$, the HSVS method assigns a "spike and slab" prior to each coefficient β_j as

$$\begin{aligned} \beta_j \mid \gamma_{g(j)},\sigma^2,\tau_j^2 &\sim (1-\gamma_{g(j)})\delta_0 + \gamma_{g(j)}\mathcal{N}(0,\sigma^2,\tau_j^2), \text{ for } j = 1,\ldots,q \\ \gamma_g(j) \mid p &\sim \text{Bernoulli}(p), \\ \tau_j^2 \mid \lambda_{g(j)} &\sim \text{Exp}(\lambda_{g(j)}/2), \end{aligned} \tag{1}$$

where δ_0 is a degenerate distribution that places all its mass at 0. Here $\gamma_{g(j)}$ is a latent binary indicator variable for group $g(j)$ for the group-level selection. When $\gamma_{g(j)} = 0$, all the predictors in group $g(j)$ are excluded from the regression model with the coefficients $\beta_{j}^{*} = 0$ for all $j^{*} \in \Psi_{g(j)}$ including β_j; conversely, when $\gamma_{g(j)} = 1$, group $g(j)$ remains in the model, and β_j is then assigned an independent scale mixture distribution of normals, allowing for independent shrinkage of each individual coefficient within the group and achieving shrinkage at the within-group level. By setting the scale mixing distribution to be an exponential distribution with a rate parameter $\lambda_{g(j)}/2$ for group $g(j)$, we achieve a Bayesian lasso estimator of β_j with a group-specific regularization parameter $\lambda_{g(j)}$.[17]

In the presence of overlapping groups where variable j belongs to multiple groups, $g1(j),\ldots,g_{kj}(j)$ for some j, we modify the HSVS method above by introducing latent variables $\beta_{jg1}(j),\ldots,\beta_{jgk_j}(j)$ such that

$$\beta_j = \beta_{jg1(j)} + \beta_{jg2(j)} + \cdots + \beta_{jgk(j)}(j), \qquad (2)$$

where β_j represents the marginal regression coefficient of covariate X_j on the response, and $\beta_{jgk(j)}$ represents the partial effect of X_j on the response imposed through the functioning of pathway $g_k(j)$. We then assign a "spike and slab" prior to each partial effect $\beta_{jgk(j)}$ similar to that in the HSVS method

$$\begin{aligned}
&\beta_{jgk(j)} \sim (1-\gamma_{gk(j)})\delta_0 + \gamma_{gk(j)}\mathcal{N}(0,\sigma^2 w_{jgk(j)}\tau_j^2),\\
&\gamma_{gk(j)} \mid p \sim \text{Bernoulli}(p),\\
&\tau_j^2 \sim \text{Exp}\left(\frac{1}{2\sum_{gk(j):gk(j)\in\mathrm{G}_j} 1/\lambda_{gk(j)}}\right),
\end{aligned} \qquad (3)$$

where $w_{jgk(j)}$ is a set of weights taking values between [0,1], and $\sum_{g:g\in g_i} w_{jg} = 1$ for each j. We note that we can estimate the sum, the marginal effect β_j which is our focus in posterior inferences; however, its latent components $\beta_{jgk(j)}$ are unidentifiable. This is because any subset of coefficients, β_j, is identifiable only if $X_s' X_s$ is invertible where X_s are the columns of the data matrix associated with the coefficients. For the vector of the latent components, $\beta_{jg(j)} = (\beta_{jg1(j)}, \beta_{j2g(j)},\ldots,\beta_{jgk(j)})$, their corresponding covariates, $X_{jg(j)} = (X_j, X_j,\ldots,X_j)$, are repetitive columns of the vector X_j. Therefore, $X'_{jg(j)} X_{jg(j)}$ is not invertible and the latent components $\beta_{jg(j)}$ are unidentifiable. In other words, there is no unique partition of β_j, since all solutions with $\sum_{j\in g_j} \beta_{jgk(j)} = \beta_j$ are equivalent. The prior specification ensures that the latent additive components of β_j shrink proportionally toward zero, with the shrinkage strength of each part $\beta_{jgk(j)}$ determined by the corresponding weight. This leads to a relatively stable Bayesian estimation for the latent components. In addition, when $\gamma_{gk(j)} = 1$ for all $g_k(j) \in \Psi_j$, we have

$$\beta_j = \sum_{g_k(j)} \beta_{jgk(j)} \sim \mathcal{N}\left(0, \sigma^2 \tau_j^2\right),$$

with the marginal scale τ_j^2 from an exponential mixing distribution. The exponential scale mixture of zero-centered normals is equivalent to a Bayesian lasso prior distribution that leads to shrinkage on coefficient estimation toward zero. Therefore, we achieve a Bayesian lasso estimation of the marginal effect β_j, as in the HSVS method, where the regularization parameter of the Bayesian lasso is equal to $1/\left(\sum_{gk(j):gk(j)\in gj} 1/\lambda_{gk(j)}\right)$. Note that $2/\lambda_{gk(j)}$ is the scale parameter of the exponential mixing prior for the normal scales of the coefficients in group $g_k(j)$ for $g_k(j) \in \mathcal{G}_j$, and the scale parameter of the exponential prior on τ_j^2 is the sum of them. The strength of shrinkage on the marginal effect β_j is thus jointly determined by the shrinkage strengths of all its involved groups. We call this extended HSVS method that is adaptive to overlapping group structures the *overlap-HSVS* method. When there is a perfect partition of individual variables without overlap, the overlap-HSVS prior defaults to an HSVS prior.

Choice of weights. One way to specify the weights in the prior of equation (3) is based on prior biological knowledge of the degree to which each gene contributes to the different pathways. When no such information is available, which is assumed for the application of this paper, we specify the weights as

$$w_{jg(j)} = \frac{\theta_{gk(j)}}{\sum_{g:g\in\mathcal{G}_j} \theta g}, \qquad (4)$$

where $\theta_g = 1/\lambda_g$ This is a natural choice as it leads to

$$w_{jgk(j)}\tau_j^2 \sim \text{Exp}\left(\frac{1}{2\theta_{gk(j)}}\right),$$

given $\theta_1,\ldots,\theta_G$, which guarantees that the partial effect $\beta_{jgk(j)}$ has the same extent of shrinkage as the other coefficients in the group $g_k(j)$. For ease of exposition in this paper, hereafter, we reparameterize the overlap-HSVS prior by using $\theta g = 1/\lambda g$.

Prior specifications. We complete the prior specifications by assigning hyperpriors similar to those in the HSVS model. We use a diffuse Gaussian prior $N(0, cI)$ for the coefficients for fixed effects b, where c is some large value. For the parameter p that controls the group-level selection, we use a conjugate Beta hyperprior: Beta(a, b) with (fixed) parameters a and b. We assign a common inverse gamma distribution Inv-Gamma(r, s) on the regularization parameters $\theta_1,\ldots,\theta_G$, ensuring their positivity. We use the improper prior density $\pi(\sigma^2) = 1/\sigma^2$ on the error variance. These choices of prior distributions lead to closed-form full conditional distributions for most of the parameters, facilitating a Bayesian Markov chain Monte Carlo (MCMC) computation via the Gibbs sampling algorithm. Please refer to Zhang et al.[16] for a detailed study and discussion of the sensitivity of the prior choices for the HSVS type selection methods. Our full hierarchical model for the overlap-HSVS model can thus be succinctly written as

$$\begin{aligned}
&\text{Likelihood}:\\
&\qquad \mathrm{Y} \mid Z, X, \beta, \sigma^2 : \mathcal{N}(Z\alpha + X\beta, \sigma^2 I_n),\\
&\text{Priors}:\\
&\qquad \alpha \sim N(0, cI),\\
&\qquad \beta_{jgk(j)} \sim (1-\gamma_{gk(j)})\delta_0 + \gamma_{gk(j)}\mathcal{N}(0,\sigma^2 w_{jgk(j)}\tau_j^2),
\end{aligned}$$

Hyperpriors :

$$\gamma_g \mid p \sim \text{Bernoulli}(p), \quad \tau_j^2 \sim \text{Exp}\left(1/(2\sum_{gk(j):gk(j)\in\mathcal{G}j}\theta_{gk(j)})\right),$$
$$p \sim \text{Beta}(a,b), \quad \theta_g^2 \sim \text{Inv}-\text{Gamma}(\tau,\beta),$$
$$\sigma^2 \sim 1/\sigma^2,$$

where the weights $w_{jgk(j)}$ are defined as in equation (4).

Posterior inference and summaries. We conduct the MCMC computation using a Gibbs sampling algorithm to generate posterior samples of all the parameters based on the full conditional posterior distributions, with block updating of $(\beta_1,\ldots,\beta_q)$, $(\gamma_1,\ldots,\gamma_G)$, $\boldsymbol{\alpha}$, σ^2, p, $\left(\tau_1^2,\ldots,\tau_q^2\right)$ and $(\theta_1,\ldots,\theta_G)$ in sequence. The full conditional distributions based on the overlap-HSVS hierarchical model presented above are all in closed form except for the regularization parameters $\boldsymbol{\theta}$, which were updated using a Metropolis–Hastings–within-Gibbs algorithm.

Pathway selection. In particular, the MCMC computation generates posterior samples of the latent binary indicator variables γ_g, indicating the inclusion/exclusion of pathway g in/from the model in each MCMC iteration. Briefly, at the tth MCMC iteration, we have the group indicator $\gamma_g^{(t)} = 1$ if pathway g is selected, and $\gamma_g^{(t)} = 0$ otherwise for each $g = 1,\ldots,G$. We then can estimate the posterior probability of including pathway g in the regression model across the MCMC samples by

$$\hat{p}_g = \frac{1}{T}\sum_{t=1}^{T} I\{\gamma_g^{(t)}\}, \tag{5}$$

where T is the total number of posterior samples collected. Such posterior probabilities of pathway inclusion give us a measure of the significance of the candidate pathways that are associated with the outcome of interest.

Gene selection. At the within-group (pathway) level, significant genes within a pathway can be selected based on the 95% credible intervals of the regression coefficients obtained from their posterior samples. Another method for evaluating the significance of the genes is based on the posterior probabilities of gene-specific effects. Suppose $\phi > 0$ denotes an effect size such that a gene j with $|\beta_j| > \phi$ is considered to have a practically significant impact on the response. Then the significance of gene j is indicated by the posterior probability of gene j's effect as estimated by

$$\hat{\pi}_j = \frac{1}{T}\sum_{t=1}^{T} I\{|\beta_j^{(t)}| > \phi\}, \tag{6}$$

where $\beta_j^{(t)}$ is the posterior sample of the regression coefficient of gene j, β_j, in the tth MCMC iteration.

Given these posterior probabilities of pathway inclusion or gene effect obtained above, we could identify significant pathways or genes using a false-discovery-rate-based thresholding method. That is, we could flag a pathway or gene as significant if the corresponding posterior probability is greater than some significant threshold, which is determined to control the overall false discovery rate at a desired level.[18,19]

Analysis of Gene Expression Datasets from MM Cancer Cells

Dataset and pre-processing. The MM data are from the Multiple Myeloma Research Consortium Reference Collection, containing a total of 304 MM patient samples, the gene expression profiles of whom were measured using Affymetrix U133 Plus 2.0 microarrays. We used Robust Multichip Average for quantification of the data. Excluding samples without appropriate clinical information, we obtained a dataset including 208 patients with MM for further genetic association analysis. Since our interest lies in simultaneous selection of pathways and genes that are associated with MM, we only considered those genes, the pathway information of which is available, and excluded those not belonging to any pathway. Hence, the resulting dataset consists of the results of microarray assays of gene expression levels for 6,323 genes from the 208 patients with MM. In addition to the gene expression profiles, non-genetic clinical information is also contained in the database, including the patient's age, gender, and measurements of clinical outcomes such as the serum free light chain ratio. We took the absolute values of the logarithm of the serum free light chain ratios as continuous responses, which measure the deviation of the ratios away from 1. They give the discrepancy between the amounts of the two types of free light chains (kappa and lambda) in logarithmic scale. Plots of data show that controlling for the numbers of kappa-type and lambda-type MM patients present in the data, the distribution of the log ratios is approximately symmetric about 0. Since the amounts of the two types of free light chains are extremely unbalanced in MM, this works as an indicator of the severity of MM. We use age and gender as the demographic covariates in our model.

Considering that the dimension of the gene expression dataset is too high compared to the sample size, we conducted a pre-screening of the gene variables for the purpose of dimension reduction. We first selected 198 individual genes by using the univariate t-test and thresholding at a p-value of 0.001. We retrieved the pathway information of the set of 198 genes from the KEGG database and found that these genes were involved in 134 biological pathways. There are a total of 4,347 genes in the original gene expression profiles that are mapped to these 134 pathways. After excluding the KEGG pathways with biological functions that are irrelevant to myeloma diseases, we further reduced the dimension and kept $G = 21$ pathways for which there were at least 10 gene members having a p-value less than 0.05. The final dataset for HSVS includes $q = 1{,}387$ genes mapped to the 21 pathways. The numbers of genes lying in each pathway range from 26 to 260. Among the 1,387 genes, 277 are mapped to multiple biological pathways.

Finally, we conducted gene and pathway analysis of the MM dataset using the overlap-HSVS method, which included the 21 candidate pathways as explanatory covariate groups and two demographic covariates, age and gender. We collected 5,000 posterior samples after a burn-in of 3,000 MCMC iterations. The total computational time was 1 hour and 10 minutes using a computer with a single core 3.4 GHz CPU and 4 GB of memory.

Results. Figure 2 depicts the posterior probabilities of pathway selection via our overlap-HSVS method based on the MCMC samples of the binary group indicators γ_g. In essence, the posterior probabilities of pathway were obtained by taking the average of the values of γ_g from MCMC as shown in equation (5). We see that three KEGG pathways have the posterior probabilities of being selected approximately 1, indicating their significance, whereas other pathways have the posterior probabilities approximately zero, indicating no significance. The candidate KEGG pathways are listed in Table 1, with the three significant pathways indicated in bold italics. The galactose metabolism pathway has been shown to be related to a defect in mitochondrial function in cancer cells.[20] The cell cycle pathway controls the commitment of cells to transition from the G1 phase to the DNA synthesis S phase, the misregulation of which has been known to lead to uncontrolled proliferation of tumor cells in different types of cancers, including MM.[21,22] The Wnt signaling pathway is functionally related to developmental processes such as cell-fate specification, cell proliferation, and cell migration, which plays a role in carcinogenesis in multiple organs.[23]

Figure 3 shows the 95% credible interval estimates of the individual coefficients for the genes lying in the three significant pathways, with panel (a) corresponding to the gene set in the galactose metabolism pathway, panel (b) to that in the cell cycle pathway, and panel (c) to that in the Wnt signaling pathway. We considered a gene to be significant if the 95% credible interval of its coefficient did not include zero. Based on the posterior samples of the coefficients, we flagged 28 genes as significant; we list them in Table 2. Note that the gene PFKM (phosphofructokinase) that was flagged as a member of the galactose metabolism pathway is also involved in glucose metabolism. Hence, further examination of the two metabolism pathways is potentially needed to reveal the oxidative energy metabolism in MM cells. In addition, we calculated the posterior probabilities of gene-specific effects as described in Section 2.3, with the effect size chosen as $\phi = 0.69$, which corresponds to the log ratio of a two-fold change in the response when the predictor increases or decreases by a unit. Figure 4 plots 30 genes with the highest posterior probabilities of gene effects, most of which were flagged as significant by the 95% credible intervals, as expected.

Biological interpretations. Among the selected genes, *ANAPC7* encodes a protein that is required for the proper ubiquitination function of the complex APC/C, a large E3 ubiquitin ligase that controls cell cycle progression by targeting a number of cell cycle regulators, and has been shown to be related to carcinogenesis in organs such as the breast.[24] The genes *CCNA1*, *CCND2*, and *CCND3* encode proteins that belong to the highly conserved cyclin family, whose members are characterized by a dramatic periodicity in protein abundance throughout the cell cycle. These proteins have been shown to interact with and/or be involved in the phosphorylation of the tumor suppressor protein Rb, and have been identified as being related to the occurrence of MM.[25] The genes *WNT11* and *WNT5B* belong to the *WNT* gene family that encodes secreted signaling proteins implicated in oncogenesis. These genes, as well as *FZD8*, the gene encoding receptors for the signaling proteins, have been identified in biological studies as being associated with MM.[25,26] Other genes that are of potential interest include *CDC14B* and *MAPK9*, which have been shown to regulate the well-known tumor suppressor protein p53; *SFRP4*, which has been identified as being associated with MM[27]; and *PPP2R1B* and *RAC1*, which have also been identified as playing important roles in carcinogenesis in multiple organs.[28,29]

We analyzed these 28 genes identified as significant through the use of Ingenuity Pathway Analysis (IPA) software

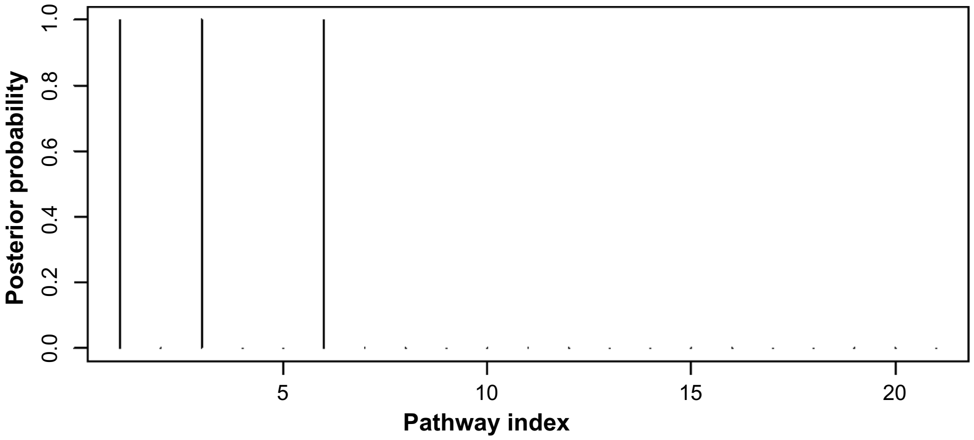

Figure 2. Posterior probability of including each pathway in the model for the MM data.

Table 1. Significant KEGG pathways selected for the MM data.

NO.	KEGG ID	KEGG PATHWAY	# OF GENES SELECTED	TOTAL # OF GENES
1	*hsa00052*	*galactose metabolism*	*1*	*26*
2	hsa04010	MAPK signaling pathway	0	265
3	*hsa04110*	*cell cycle*	*12*	*117*
4	hsa04115	p53 signaling pathway	0	68
5	hsa04210	Apoptosis	0	87
6	*hsa04310*	*Wnt signaling pathway*	*18*	*148*
7	hsa03018	RNA degradation	0	57
8	hsa03030	DNA replication	0	36
9	hsa03040	Spliceosome	0	116
10	hsa03420	Nucleotide excision repair	0	44
11	hsa04512	ECM reception interaction	0	84
12	hsa04620	Toll like reception signaling	0	102
13	hsa04621	NOD like reception signaling	0	62
14	hsa04622	RIG-I-like reception signaling	0	71
15	hsa05120	Epithelial cell signaling in *Helicobacter pylori* infection	0	68
16	hsa00310	Lysine degradation	0	44
17	hsa00330	Arginine and proline metabolism	0	54
18	hsa03010	Ribosome	0	87
19	hsa04910	Insulin signaling pathway	0	135
20	hsa05211	Renal cell carcinoma	0	70
21	hsa04540	Gap junction	0	87

(Ingenuity® Systems, www.ingenuity.com) in order to gain insight into the cellular functions associated with this set of genes. Our analysis found this set of genes to be related to three gene regulatory networks, as shown in Figure 5. These networks involve ERK1/2, Rb, TP53, NFKB, ER, and AKT as their main downstream targets or upstream regulators. These genes are known to play important roles in various cancers of different organs. One of the networks also involves the production of immunoglobulins, validating their association with the response variable. In addition, the IPA identified multiple canonical pathways that are significant to this set of genes, the top five of which include the cell cycle pathway, the Wnt signaling pathway, and two pathways that are related to rheumatoid arthritis. This agrees with the biological discovery that occurrence of MM may be increased in patients with rheumatoid arthritis.[30] These biological discoveries partially support our inferential results of the MM data analysis based on the overlap-HSVS method. The agreement between our inference and the biological discoveries also validates the serum free light chain ratio as a strong indicator for diagnosing and monitoring MM.

We also compared the inferential results to those of a popular variable selection method, the lasso, which carries out a selection process at only the individual gene level. The lasso method identified 39 genes that were considered to be related to the response. Without considering the group structure in variable selection, the lasso method identified 19 KEGG pathways in which the genes are involved to be relevant to the response out of the 21 total pathways under consideration. In summary, the overlap-HSVS method results in a more parsimonious model than the lasso method, with a smaller pool of significant genes and pathways selected for further biological investigation via future functional validations.

Real data based simulation. We conducted a simulation study based on the inferential results to examine the performance of our method in the analysis of the MM data. We considered a data matrix of 1,387 predictors that have the same grouping structure as that in the MM dataset. The posterior median estimates of the parameters obtained by the overlap-HSVS method for the MM dataset were taken as the true values of the coefficients β and the error variance σ^2. Hence, the coefficient vector includes 28 nonzero individual coefficients from three significant groups. Each column of the data matrix, X, was generated from a normal distribution $\mathcal{N}(0, s_j^2)$ with the variance s_j^2 the same as that in the gene expression data matrix from the MM patients.

We generated 20 simulated datasets and applied the overlap-HSVS method as well the lasso method, as a comparison, to each dataset for simultaneous selection of pathways and genes. At the group level, the overlap-HSVS method has an average true positive rate (TPR) of 0.82 (≈2.45 pathways correctly selected) and an average false positive rate (FPR) of 0.03

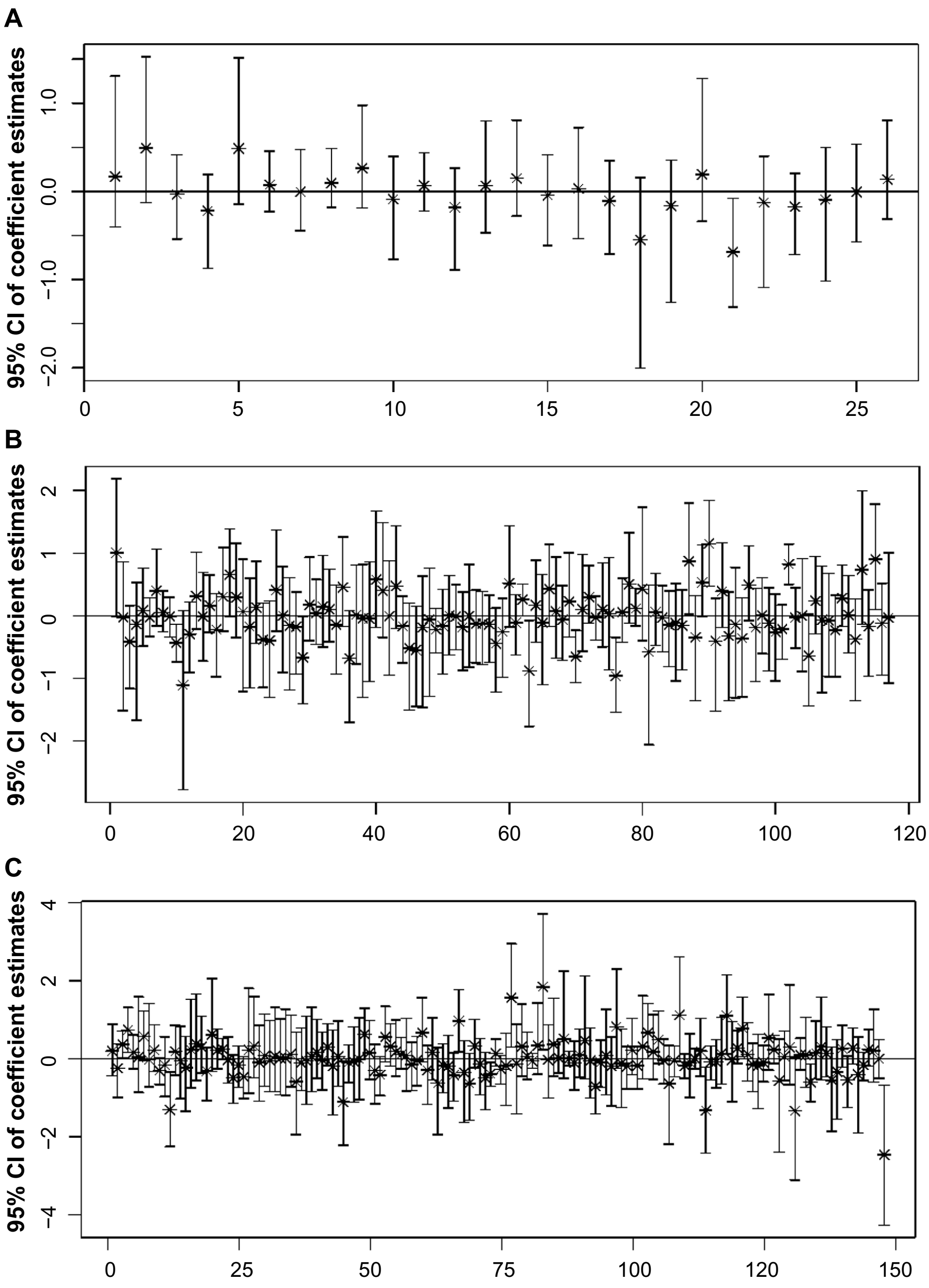

Figure 3. The 95% posterior credible intervals of the coefficients for the gene variables in the selected pathways: (**A**) galactose metabolism pathway; (**B**) cell cycle pathway; (**C**) Wnt signaling pathway.

(≈0.60 pathways falsely selected), with the respective standard errors being 0.20 and 0.04. In contrast, the lasso method always includes all groups in the model for all simulated datasets, with TPRs and FPRs both being 1.00. At the within-group level, the overlap-HSVS method has an average TPR of 0.60 (≈16.75 genes correctly selected) and an average FPR of 0.02 (≈30.80 genes falsely selected), with the respective standard errors being 0.34 and 0.02. In contrast, the lasso method has an average TPR of 0.94 (≈26.3 genes correctly selected) and an average FPR of 0.08 (≈103.75 genes falsely selected), with the respective standard errors being 0.03 and 0.01. The results show that our overlap-HSVS method can identify most groups as well as within-group individual variables while having significantly lower FPRs at both levels compared to the lasso method. The simulation results are consistent with the results of a detailed simulation study on the HSVS methods in Zhang et al.[16] suggesting that the overlap-HSVS/HSVS method is a strong variable selector at both the group and within-group levels.

Conclusion

In this paper, we developed the overlap-HSVS method for simultaneous selection of groups and variables within the

Table 2. Significant genes selected for the MM data.

NO.	GENE SYMBOL	GENE NAME
1	ANAPC7	anaphase promoting complex subunit 7
2	CAMK2G	calcium/calmodulin-dependent protein kinase II gamma
3	CCNA1	cyclin A1
4	CCND2	cyclin D2
5	CCND3	cyclin D3
6	CCNE2	cyclin E2
7	CDC14B	cell division cycle 14B
8	CDKN1C	cyclin-dependent kinase inhibitor 1C (p57, Kip2)
9	CSNK2B	casein kinase 2, beta polypeptide
10	CTNNB1	catenin (cadherin-associated protein), beta 1, 88kDa
11	CUL1	cullin 1
12	DBF4	DBF4 homolog
13	FZD7	frizzled family receptor 7
14	FZD8	frizzled family receptor 8
15	MAPK9	mitogen-activated protein kinase 9
16	MCM7	minichromosome maintenance complex component 7
17	NKD2	naked cuticle homolog 2 (Drosophila)
18	PCNA	proliferating cell nuclear antigen
19	PFKM	phosphofructokinase, muscle
20	PPP2R1B	protein phosphatase 2, regulatory subunit A, beta
21	PRKCA	protein kinase C, alpha
22	RAC1	ras-related C3 botulinum toxin substrate 1 (rho family, small GTP binding protein Racl)
23	ROCK1	Rho-associated, coiled-coil containing protein kinase 1
24	SFN	stratifin
25	SFRP4	secreted frizzled-related protein 4
26	WIF1	WNT inhibitory factor 1
27	WNT11	wingless-type MMTV integration site family, member 11
28	WNT5B	wingless-type MMTV integration site family, member 5B

groups when group overlap exists in the data structure. We applied the method to a MM dataset to select significant pathways and genes that are associated with clinical outcomes of MM. The method can be applied to a variety of structured data such as RNA-Seq data or data from patients of other cancer types, in which overlapping group structure exists among the predictor variables and group selection is of interest. We extended the HSVS method by introducing latent partial effect variables and corresponding weights to proportionally shrink the partial effects toward zero. The key idea underlying the model specification is that if the group of variables has an overall significant impact on the response, with less shrinkage, then the variables that are shared with other groups have more influence on the response through the contribution of this significant group. One weakness of the method brought about from this setup is that if the members of a group predominantly overlap with those of other groups, it may lead to instability in MCMC computations.

In coordination with the performance of the HSVS method, the overlap-HSVS method as an extension for analyzing overlapping group structures is a strong variable selection method, at both the group and within-group levels. Applied to the MM dataset, the method identified three significant pathways, including 28 significant genes, which were considered to be associated with the outcome measurements (the differences in the measurements of the two serum free light chain types in log ratio). Some of the genes and pathways we identified have been determined in biological investigations to be important biomarkers of MM. Our results support the hypothesis that the serum free light chain ratio can be an indicative measurement of the severity of MM, and is useful in diagnosing and monitoring this disease. We applied the method to the MM dataset for pathway and gene selection. The method can also be applied to a variety of structured data, in which overlapping group structure exists among the predictor variables and group selection is of interest. Examples include RNA-Seq data or gene expression

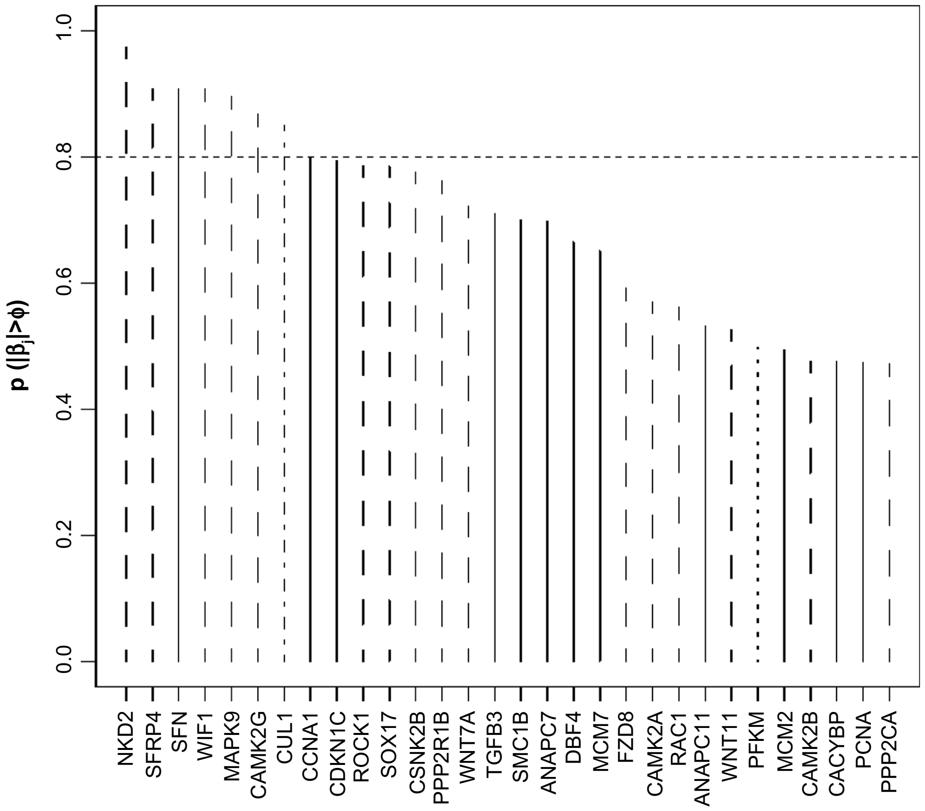

Figure 4. The 30 genes with highest posterior probabilities of regression coefficients being greater than $\phi = 0.5$ in the absolute value. The line patterns correspond to the pathways to which the genes belong: dotted lines for genes in the galactose metabolism pathway; solid lines for genes in the cell cycle pathway; dashed lines for genes in the Wnt signaling pathway; dot-dash lines for genes in both the cell cycle and Wnt signaling pathways.

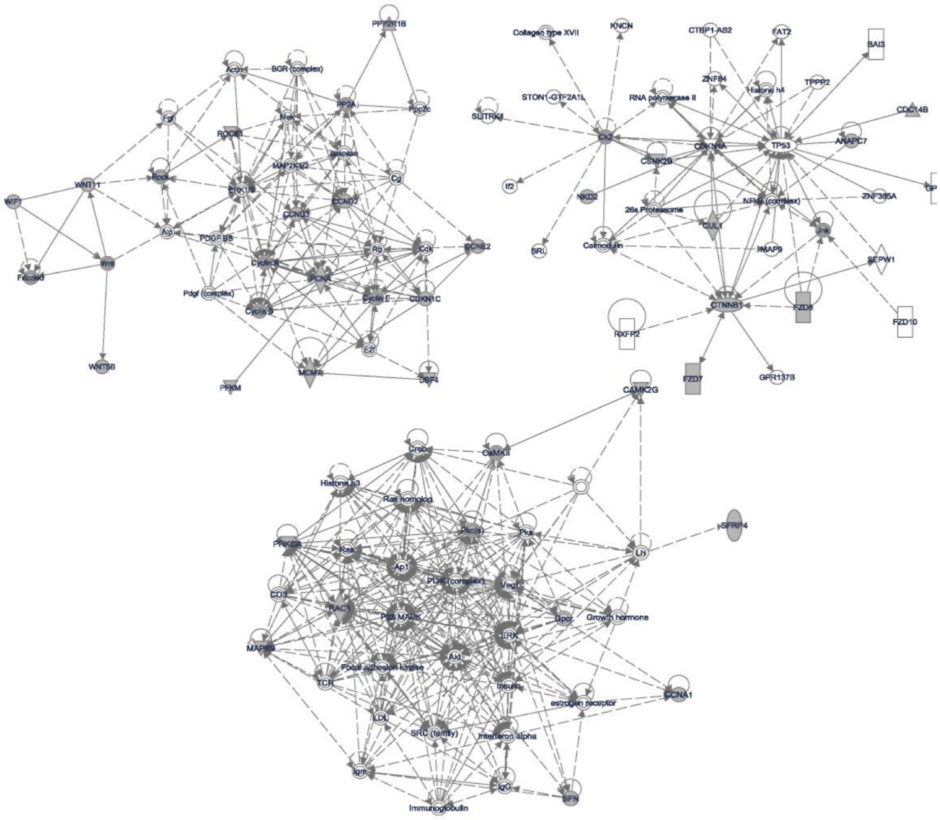

Figure 5. Three regulatory networks that involve the 28 flagged genes as identified by IPA. The nodes with filled color correspond to the flagged significant genes.

data from patients of other cancer types, which we leave as potential future projects.

Although the overlap-HSVS method is able to deal with the overlapping group structures in the pathway-based analysis of MM gene expression data, the model treats the genes within a pathway as independent, assigning an independent normal-exponential prior to each coefficient. However, in reality, genes work in an interactive fashion. With developments in biological study, we have increased knowledge of the regulatory networks between genes within a pathway. In this situation, we seek to incorporate such prior biological information of the regulatory relationships between genes into the HSVS/overlap-HSVS methods. This will allow us to borrow strength in inference, not only from samples, but also from interactive genes. In addition, the incorporation of prior biological information of gene regulatory networks can take the high correlations among variables into consideration and thus result in more robust model selection and estimation. We leave this topic for our future studies.

Author Contributions

Conceived and designed the experiments: LZ, VB. Analyzed the data: LZ. Wrote the first draft of the manuscript: LZ, VB. Contributed to the writing of the manuscript: JM, JZ, RO, VB. Agree with manuscript results and conclusions: LZ, JM, JZ, RO, VB. Jointly developed the structure and arguments for the paper: LZ, VB. Made critical revisions and approved final version: LZ, JM, JZ, VB. All authors reviewed and approved of the final manuscript.

REFERENCES

1. Raab MS, Podar K, Breitkreutz I, Richardson PG, Anderson KC. Multiple myeloma. *Lancet*. 2009;374(9686):324–39.
2. Dispenzieri A, Kyle R, Merlini G, et al; International Myeloma Working Group. International Myeloma Working Group guidelines for serum free light chain analysis in multiple myeloma and related disorders. *Leukemia*. 2009; 23(2):215–24.
3. Siegel D, Bilotti E, Hoeven KH. Serum free light chain analysis for diagnosis, monitoring, and prognosis of monoclonal gammopathies. *Lab Med*. 2009;40:363–6.
4. Iwama K, Chihara D, Tsuda K, et al. Normalization of free light chain kappa/lambda ratio is a robust prognostic indicator of favorable outcome in patients with multiple myeloma. *Eur J Haematol*. 2013;90(2):134–41.
5. Dryja TP. Gene-based approach to human gene-phenotype correlations. *Proc Natl Acad Sci USA*. 1997;94:12117–121.
6. Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*. 1999;286:531–7.
7. Vogelstein B, Kinzler KW. *The Genetic Basis of Human Cancer*. Toronto, Ontario, Canada: McGraw-Hill; 2001.
8. Davies M, Hennessy B, Mills GB. Point mutations of protein kinases and individualised cancer therapy. *Expert Opin Pharmacother*. 2006;7:2243–61.
9. Kanehisa M, Araki M, Goto S, et al. KEGG for linking genomes to life and the environment. *Nucleic Acids Res*. 2008;36(Database issue):D480–4.
10. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. *J R Stat Soc Series B Stat Methodol*. 2006;68:49–67.
11. Raman S, Fuchs T, Wild P, Dahl E, Roth V. The Bayesian group-lasso for analyzing contingency tables. In: Proceedings of the 26th International Conference on Machine Learning, Montreal, Canada. 2009: 881–8.
12. Zhao P, Rocha G, Yu B. Grouped and hierarchical model selection through composite absolute penalties. *Ann Stat*. 2009;37:3468–97.
13. Wang S, Nan B, Zhou N, Zhu J. Hierarchically penalized Cox regression for censored data with grouped variables. *Biometrika*. 2009;96:307–22.
14. Ma S, Zhang Y, Huang J, et al. Identification of non-Hodgkin's lymphoma prognosis signatures using the CTGDR method. *Bioinformatics*. 2010;26:15–21.
15. Stingo FC, Chen YA, Tadesse MG, Vannucci M. Incorporating biology information into linear models: a Bayesian approach to the selection of pathways and genes. *Ann Appl Stat*. 2011;5(3):1978–2002.
16. Zhang L, Baladandayuthapani V, Mallick BM, et al. Bayesian hierarchical structured variable selection methods with application to molecular inversion probe studies in breast cancer. *J R Stat Soc Ser C Appl Stat*. 2014;March, Epub ahead of print. Available at http://onlinelibrary.wiley.com/doi/10.1111/rssc.12053/pdf
17. Park T, Casella G. The Bayesian lasso. *J Am Stat Assoc*. 2008;103:681–6.
18. Müller P, Parmigiani G, Robert C, Rousseau J. Optimal sample size for multiple testing: the case of gene expression microarrays. *J Am Stat Assoc*. 2004;99:990–1001.
19. Baladandayuthapani V, Ji Y, Talluri R, Neito-Barajas LE, Morris J. Bayesian random segmentation models to identify shared copy number aberrations for array CGH data. *J Am Stat Assoc*. 2010;105:390–400.
20. Rossignol R, Gilkerson R, Aggeler R, Yamagata K, Remington SJ, Capaldi RA. Energy substrate modulates mitochondrial structure and oxidative capacity in cancer cells. *Cancer Res*. 2004;64(3):985–93.
21. Malumbres M, Barbacid M. Cell cycle, CDKs and cancer: a changing paradigm. *Nat Rev Cancer*. 2009;9(3):153–66.
22. Hideshima T, Mitsiades C, Tonon G, Richardson PG, Anderson KC. Understanding multiple myeloma pathogenesis in the bone marrow to identify new therapeutic targets. *Nat Rev Cancer*. 2007;7:585–98.
23. Logan CY, Nusse R. The Wnt signaling pathway in development and disease. *Annu Rev Cell Dev Biol*. 2004;20:781–810.
24. Park KH, Choi SE, Eom M, Kang Y. Downregulation of the anaphase-promoting complex (APC)7 in invasive ductal carcinomas of the breast and its clinicopathologic relationships. *Breast Cancer Res*. 2006;7(2):R238–47.
25. Zhan F, Huang Y, Colla S, et al. The molecular classification of multiple myeloma. *Blood*. 2006;108(6):2020–8.
26. Mahtouk K, Moreaux J, Hose D, et al. Growth factors in multiple myeloma: a comprehensive analysis of their expression in tumor cells and bone marrow environment using Affymetrix microarrays. *BMC Cancer*. 2010;10:198.
27. He B, Lee AY, Dadfarmay S, et al. Secreted frizzled-related protein 4 is silenced by hypermethylation and induces apoptosis in B-catenin-deficient human mesothelioma cells. *Cancer Res*. 2005;65:743–8.
28. Wang SS, Esplin ED, Li JL, et al. Alterations of the PPP2R1B gene in human lung and colon cancer. *Science*. 1998;282(5387):284–7.
29. Azab AK, Azab F, Blotta S, et al. RhoA and Rac1 GTPases play major and differential roles in stromal cell-derived factor-1-induced cell adhesion and chemotaxis in multiple myeloma. *Blood*. 2009;114(3):619–29.
30. Eriksson M. Rheumatoid arthritis as a risk factor for multiple myeloma: a case-control study. *Eur J Cancer*. 1993;29:259–63.