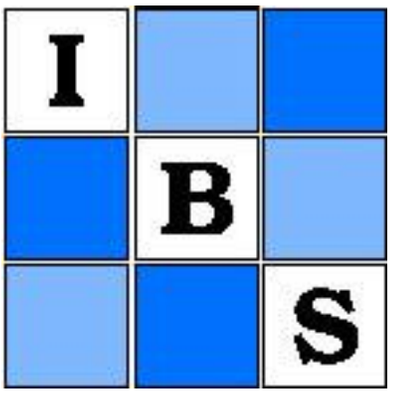

BIOMETRICS 73, 615–624
June 2017

DOI: 10.1111/biom.12587

A Bayesian Integrative Approach for Multi-Platform Genomic Data: A Kidney Cancer Case Study

Thierry Chekouo,[1] Francesco C. Stingo,[2,*] James D. Doecke,[3] and Kim-Anh Do[4]

[1]Department of Mathematics and Statistics, University of Minnesota Duluth, Duluth, MN 55812, USA
[2]Dipartimento di Statistica, Informatica, Applicazioni "G.Parenti", University of Florence, 50134 Florence, Italy
[3]CSIRO Health and Biosecurity/Australian e-Health Research Center Level 5, Queensland 4029, Australia
[4]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA
*email: f.stingo@disia.unifi.it

SUMMARY. Integration of genomic data from multiple platforms has the capability to increase precision, accuracy, and statistical power in the identification of prognostic biomarkers. A fundamental problem faced in many multi-platform studies is unbalanced sample sizes due to the inability to obtain measurements from all the platforms for all the patients in the study. We have developed a novel Bayesian approach that integrates multi-regression models to identify a small set of biomarkers that can accurately predict time-to-event outcomes. This method fully exploits the amount of available information across platforms and does not exclude any of the subjects from the analysis. Through simulations, we demonstrate the utility of our method and compare its performance to that of methods that do not borrow information across regression models. Motivated by The Cancer Genome Atlas kidney renal cell carcinoma dataset, our methodology provides novel insights missed by non-integrative models.

KEY WORDS: Bayesian variable selection; Integrating multi-regressions; Markov random field; Multiplatform genomic data; Non-local prior.

1. Introduction

Rapid advances in genomic technologies and a combined international effort to collaborate in the search for informative cancer biomarkers have created large, well-characterized sets of genomic and clinical data that are publicly available. The efforts to define risk factors for cancer pathogenesis and progression have focused on individual sets of markers, and have evolved from investigating single nucleotide polymorphisms (SNPs) to analyses of transcriptomics and miRNA expression and DNA methylation. While risk variants from SNP studies have been used to identify SNPs associated with disease risk (Bush and Moore., 2012), studies assessing transcriptomics, DNA methylation and miRNA markers have led to discoveries in both disease mechanisms and treatment options (Lin and Gregory, 2015). In the search for potential "oncogenes," many early studies focused on individual genomic platforms; however, such studies failed to incorporate the potential regulation by both DNA methylation and non-coding RNAs. Although single platform studies have provided great insight into cancer biology, the integration of genomic data from multiple platforms has the additional capability to assess the co-regulation of mRNA expression in the disease model.

Genomic data integration has been defined as the process of statistically combining evidence from different data resources or platforms to uncover synergistic mechanisms previously unseen via individual platform analyses (Lu et al., 2005). In an effort to examine the utility of genomic data integration, investigators such as Stingo et al. (2010); Wang et al. (2013); Srivastava et al. (2013) and Chekouo et al. (2015) used full Bayesian methods to investigate pairwise interactions between mRNA and miRNA expression, and mRNA expression and DNA methylation. Daemen et al. (2008) and more recently Wu et al. (2012) developed a two-step frequentist method that combines mRNA expression and proteomic data, but few have truly combined information from three or more molecular platforms. Integration of diverse genomic data from many platforms has the potential to increase precision, accuracy, and statistical power for the identification of combinations of important biomarkers associated with clinical outcomes. In fact, multi-factorial diseases such as cancer are thought to result from random genetic alterations via different mechanisms (Hamid et al., 2009), which can be revealed only by the integrative analysis of data from multiple platforms.

A fundamental problem observed in many large-scale genomic studies is the loss of sample size while integrating complete data over multiple platforms. Assessing 11 tumor types from The Cancer Genome Atlas (TCGA) with reasonably large sample sizes, we observed large sample size reductions (greater than 50%) when combining mRNA, miRNA and DNA methylation platforms. Specifically, only 50%, 10%, 26% and 41% of the respective samples of breast cancer, glioblastoma, kidney renal clear cell carcinoma (KIRC) and uterine cancer had complete data for all three platforms. Thus, using a complete sample approach to multi-platform genomic data decreases the potential sample size and decreases the capability to identify potentially important and clinically relevant genomic biomarkers. In addition, standard integrative models often fail to predict survival times

for patients whose data are missing for one or more of the platforms analyzed.

To overcome these drawbacks, we developed novel Bayesian methodology that integrates multi-regression models (IMR) to identify small sets of potential biomarkers to use in predicting survival times in the context of accelerated failure time (AFT) models. We use a combination of the sample sets analyzed through each platform to build each regression model, which allows us to use the complete sample set measured from each platform.

Using a biological model in which biomarkers with large effect sizes will be selected from different regression models, we link the selection of biomarkers via Markov random fields (MRFs), which encourages the selection of common biomarkers from multiple regression models through informative priors. The posterior probabilities of inclusion provide the importance of biomarkers from each platform in each involved regression model, and then summarize the similarity between biomarkers from each platform. We apply our approach to samples from TCGA KIRC data, and specifically focus on mRNA, miRNA and DNA methylation data, along with clinical information and time-to-event outcomes.

2. Motivating Dataset

Prognostic models based on gene expression (mRNA) from either microarrays or RNA sequencing data have been extensively studied in cancer research. Here, we aim to define prognostic models that integrate mRNA expression with two well-known regulatory factors, DNA methylation and miRNA markers. DNA methylation consists of adding a methyl group ($-CH_3$) to the cytosine (C) DNA nucleotides, and is often found where the DNA sequence has a high proportion of cytosine and guanine (G) nucleotides. Over-representation of DNA methylation across genomic regions within gene promoters and transcription start sites has been associated with significant down-regulation of gene expression, and has proven to be a useful source of biomarkers for risk stratification and cancer pathophysiology (Laird, 2003). miRNAs are small non-coding RNAs that bind the 3' end of the mRNA transcript, actively degrading the mRNA and modifying mRNA translation. miRNA are traditionally 18-25 nucleotides in length and can bind multiple genes, thereby providing a means for large-scale gene regulation. Many miRNA markers have been identified in cancer pathways linked to clinical outcomes, and are considered to be potential oncogenes (Qin, 2008; Xu et al., 2014).

In this paper, we consider experimental data from the TCGA KIRC database. We imported level-3 genomic data, including 20,532 mRNA markers (Illumina HiSeq2000 platform) from 543 samples, 825 miRNA markers (Illumina miSeq platform) from 320 samples, and 473,929 DNA methylation markers (probes, Illumina Infinium HumanMethylation450 BeadChip array) from 497 samples. For each of the three platforms, we used the same principles of data cleaning and pre-processing, including variable reduction via a) minimal variance (or trimmed variance); b) minimal range (or trimmed range), c) proportion of missing data (greater than 40% missing data) and d) minimal association with the time-to-event outcome of interest, via univariable Cox proportional hazards (CPH) regression analysis. For DNA methylation, we included in our dataset only those probes at the 5' (including promoter, transcription start site and first exon) and 3' ends of each gene, that were highly correlated with the genes included in the final dataset to enrich for maximal effects on mRNA expression. To avoid multi-collinearity issues, we used the *gene shaving* algorithm of Hastie et al. (2000) to effectively identify subsets of biomarkers with correlated expression patterns across samples or conditions and assessed these along with individual biomarkers. All pre-processing steps are detailed in Web Appendix A. Our pre-processing steps resulted in the selection of $p_1 = 776$ mRNA transcripts (genes), $p_2 = 91$ miRNA markers and $p_3 = 729$ methylation probes from 448, 198 and 248 samples, respectively.

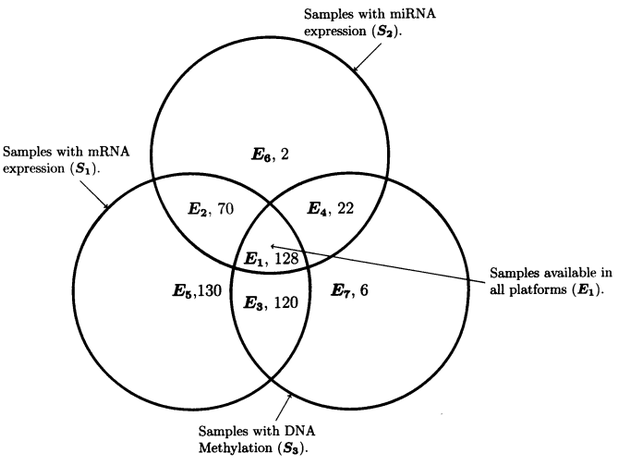

Figure 1. A 3-set Venn diagram for three platforms with the number of samples in each distinct group $\boldsymbol{E}_k$, $k = 1, \ldots, 7$. For instance, "$\boldsymbol{E}_6$, 2" means group $\boldsymbol{E}_6$ has 2 samples.

The common approach used to construct a multivariable model in the presence of missing blocks of data (e.g. when gene measurements for any platform are missing for a large number of units) is to restrict the analysis to the set of units with fully observed data. In this study, we aimed to build a prognostic model that could jointly analyze the samples for which there were complete data across all platforms (130 samples) in addition to the other samples with missing data. To do this, we subdivided the 478 samples into 7 groups ($\boldsymbol{E}_k$, $k = 1, ..., 7$), which align with the distinct regions of the 3-set Venn diagram represented in Figure 1.

The number of samples, and/or events, in group $\boldsymbol{E}_4$, $\boldsymbol{E}_6$ or $\boldsymbol{E}_7$ is clearly too small to build an informative statistical model. Therefore for this study, we focused our attention on only the 448 samples from groups $\boldsymbol{E}_1$, $\boldsymbol{E}_2$, $\boldsymbol{E}_3$ and $\boldsymbol{E}_5$. Thus, mRNAs, miRNAs and methylation probes, respectively, are potential markers in 4, 2, and 2 statistical models. Patient characteristics by group are described in Web Appendix A, Table 1.

3. Model Specification

We assume a total of K types of biomarkers (or platforms) that are likely to be associated with a time-to-event outcome of interest. As described in Section 2, for several samples,

the availability of genomic data is restricted to a subset of the K platforms. Let $\boldsymbol{S_k}$ be the set of samples with measurements available from platform k. Therefore, we have a total of $2^K - 1$ distinct regions within the K-set Venn diagram of the sets $\boldsymbol{S_k}$, $k = 1, ..., K$. For example, if we aim to integrate mRNA and miRNA data ($K = 2$), we have three distinct subsets consisting of samples with both mRNA and miRNA measurements, only mRNA, or only miRNA. In our case study, we integrate mRNA, miRNA, and DNA methylation ($K = 3$) data; the data structure can be represented by a 3-set Venn diagram (Figure 1), defined by 7 distinct groups of samples $\boldsymbol{E_s}$, $s = 1, ..., 7$. For each group s, $y_n^{(s)} = \min(t_n^{(s)}, c_n^{(s)})$ and $\delta_n^{(s)} = I\{y_n^{(s)} \leq c_n^{(s)}\}$ $(n = 1, \ldots, N_s)$ denote the observed clinical outcome, where $t_n^{(s)}$ is the survival time for subject n, $c_n^{(s)}$ is the censoring time, and $\delta_n^{(s)}$ is a censoring indicator.

Let $\boldsymbol{P}^{(s)}$ and $\boldsymbol{X}^{(s)}$ be the matrices of standardized biomarker measurements and the clinical prognostics factors (such as age, gender, tumor stage and grade), respectively. Hence, data from different platforms are on a comparable scale. We can define an AFT model with $\boldsymbol{P}^{(s)}$ and $\boldsymbol{X}^{(s)}$ as covariates. Following Tanner and Wong (1987), we introduce a latent variable $\boldsymbol{y}^{\star(s)}$ such that for any sample n in group s, $y_n^{\star(s)} = \log(t_n^{(s)})$ if $\delta_n^{(s)} = 1$, and $y_n^{\star(s)} > \log(y_n^{(s)})$, otherwise. The proposed AFT model can be written as

$$\boldsymbol{y}^{\star(s)} = \mathbb{1}_{N_s}\beta_0^{(s)} + \boldsymbol{P}^{(s)}\boldsymbol{\beta}^{(s)} + \boldsymbol{X}^{(s)}\boldsymbol{\beta}'^{(s)} + \boldsymbol{\nu}^{(s)}, \quad \boldsymbol{\nu}^{(s)} \sim \mathcal{N}(\boldsymbol{0}, \sigma_s^2 I_{N_s}), \tag{1}$$

where $\beta_0^{(s)}$ is the intercept, and $\boldsymbol{\beta}^{(s)}$ and $\boldsymbol{\beta}'^{(s)}$ are the regression coefficients for biomarkers and prognostic factors, respectively. For our approach, we define $s = 1, ..., 2^K - 1$ regression models, one for each distinct group $\boldsymbol{E}_s$.

Let p_k be the number of biomarkers of type k, e.g., the number of genes with measured mRNA expression. Each type of biomarker is part of 2^{K-1} distinct groups of samples. For example, when $K = 3$, gene expression data are obtained for a sample that can belong to one of the following 4 groups according to Figure 1: (i) $\boldsymbol{E_1}$ (mRNA, miRNA and methylation), (ii) $\boldsymbol{E_2}$ (mRNA and miRNA), (iii) $\boldsymbol{E_3}$ (mRNA and methylation), and (iv) $\boldsymbol{E_5}$ (mRNA). For each regression/group, we envision that only a subset of biomarkers is associated with the time-to-event endpoint. We introduce K binary matrices, $\boldsymbol{\Gamma_k}$, to perform the selection of the genomic markers; these are $p_k \times 2^{K-1}$ matrices, where each component $\boldsymbol{\gamma}_k^{(p)}$ is a p_k binary vector, with $p \in \{\boldsymbol{E}_1, \ldots, \boldsymbol{E}_{2^K-1}\}$. The generic element $\gamma_{k,g}^{(p)}$ is set to 1 if biomarker g of type k is associated with the outcome of interest in equation p, and set to 0 otherwise. For example when $K = 3$, from Figure 1 we can define $\boldsymbol{\Gamma_1} = (\boldsymbol{\gamma}_1^{(E_1)}, \boldsymbol{\gamma}_1^{(E_2)}, \boldsymbol{\gamma}_1^{(E_3)}, \boldsymbol{\gamma}_1^{(E_5)})$ for the selection of genes, $\boldsymbol{\Gamma_2} = (\boldsymbol{\gamma}_2^{(E_1)}, \boldsymbol{\gamma}_2^{(E_2)}, \boldsymbol{\gamma}_2^{(E_4)}, \boldsymbol{\gamma}_2^{(E_6)})$ for the selection of miRNAs, and $\boldsymbol{\Gamma_3} = (\boldsymbol{\gamma}_3^{(E_1)}, \boldsymbol{\gamma}_3^{(E_3)}, \boldsymbol{\gamma}_3^{(E_4)}, \boldsymbol{\gamma}_3^{(E_7)})$ for the selection of methylation probes. We assume that only the genomic biomarkers are subject to variable selection; whereas all clinical prognostic factors are ubiquitous and are always included in the regression models.

3.1. *Biomarker Selection via Non-Local Priors*

The variable selection indicators $\boldsymbol{\Gamma_k}$ can be used to define a mixture distribution on the regression coefficients $\beta_{k,g}^{(p)}$ from model (1). These prior probabilities are set to be independent *a priori* and are defined by the following mixture model:

$$p(\beta_{k,g}^{(p)}|\gamma_{k,g}^{(p)}, \tau_k, \sigma_s^2, r) = \gamma_{k,g}^{(p)}\mathcal{PM}(\beta_{k,g}^{(p)}; r, \tau_k, \sigma_s^2) + (1 - \gamma_{k,g}^{(p)})\mathcal{I}_0(\beta_{k,g}^{(p)}), \tag{2}$$

where $\mathcal{PM}(\beta_{k,g}^{(p)}; r, \tau_k, \sigma_s^2)$ is a product moment distribution (pMOM) density defined by the parameters r, τ_k and σ_s^2; τ_k and σ_s^2 are scale parameters that determine the dispersion of $\beta_{k,g}^{(p)}$ around 0 and $r \in \mathbb{N}^+$ is the order of the density. pMOM distributions are symmetric at zero, defined on $\mathbb{R}$ and can be written as $\mathcal{PM}(\beta_{k,g}^{(p)}; r, \tau_k, \sigma_s^2) \propto (\beta_{k,g}^{(p)})^{2r} \exp\left\{-\frac{1}{2\tau_k\sigma_s^2}(\beta_{k,g}^{(p)})^2\right\}$. Such pMOM distributions have been efficiently used for Bayesian model selection (Johnson and Rossell, 2012; Chekouo et al., 2015). They are called *non-local* priors since they give low probabilities to coefficients close to zero and efficiently eliminate regression models that contain unnecessary explanatory variables. We also assume a pMOM prior on the regression coefficients $\boldsymbol{\beta}'^{(s)}$ and the intercept terms $\beta_0^{(s)}$ with parameters τ_0 and r. We assume a conjugate inverse gamma prior for σ_s^2, with parameters α and ψ.

3.2. *MRF Prior to Borrowing Strength Across Groups*

We envision that a strong prognostic marker will likely induce a non-negligible regression coefficient for the majority of the regression models in which this biomarker is included as a covariate. To capture these probabilistic dependencies, we define a MRF prior on the variable selection prior indicators $\gamma_{k,g}^{(p)}$'s that encourages the selection of the same biomarkers across the regression models. MRFs have previously been used to model the relationships among covariates (Li and Zhang, 2010; Stingo et al., 2011, 2012) or to encourage the selection of the same edges in related graphs (Peterson et al., 2015). The prior probability of the vector of binary indicators of the inclusion of biomarker g (of type k) $\boldsymbol{\gamma_{k,g}} = (\gamma_{k,g}^{(1)}, ..., \gamma_{k,g}^{(2^{K-1})})$ is defined as

$$p(\boldsymbol{\gamma_{k,g}}|\nu_k, \boldsymbol{\Theta_k}) = Z(\nu_k, \boldsymbol{\Theta_k})^{-1} \exp(\nu_k \boldsymbol{1}_{2^{K-1}}^T \boldsymbol{\gamma_{k,g}} + \boldsymbol{\gamma_{k,g}}^T \boldsymbol{\Theta_k} \boldsymbol{\gamma_{k,g}}), \tag{3}$$

where ν_k is a parameter that controls the prior inclusion probability and is set to a fixed value that encourages a sparse selection of biomarkers of type k. The parameter $\boldsymbol{\Theta_k} = (\theta_{pp'}^{(k)})$ is a $2^{K-1} \times 2^{K-1}$ symmetric matrix that represents the pairwise relatedness of the biomarker of type k for each equation. The diagonal entries are set to zero, and the non-negative off-diagonal entries represent the strength of the dependence between the binary indicator associated with the same biomarker (of type k). For example, high values of $\theta_{E_1E_2}^{(k)}$ will encourage the selection of the same set of biomarkers in groups $\boldsymbol{E}_1$ and $\boldsymbol{E}_2$. These MRF priors effectively link the $2^K - 1$ regression models and define a biomarker selec-

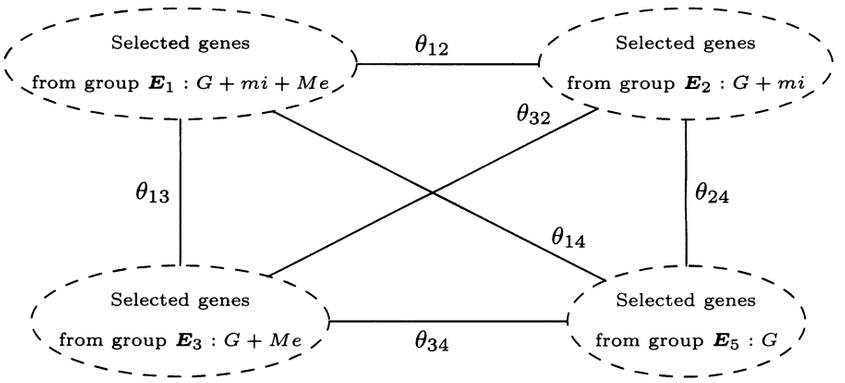

Figure 2. $K = 3$, Gene variable selection mechanism. $G + mi + Me$ stands for group of samples with gene, miRNA and methylation measurements.

tion mechanism that borrows strength across models. The normalizing constant in equation (3) is defined as

$$Z(\nu_k, \boldsymbol{\Theta_k}) = \sum_{\boldsymbol{\gamma} \in \{0,1\}^{2^{K-1}}} \exp(\nu_k \mathbf{1}_{2^{K-1}}^T \boldsymbol{\gamma} + \boldsymbol{\gamma}^T \boldsymbol{\Theta}_k \boldsymbol{\gamma}). \quad (4)$$

Although the normalizing constant (4) involves a double exponential number of terms in $K - 1$, for most settings of interest, the number of platforms K is reasonably small and the computation is straightforward. For example, if $K = 2$ or $K = 3$, there are respectively $2^{2^{2-1}} = 4$ or $2^{2^{3-1}} = 16$ possible values that $\boldsymbol{\gamma}_{k,g}$ can assume. We assume that the $\theta_{pp'}^{(k)}$'s are independent and the joint prior on the off-diagonal entries of $\boldsymbol{\Theta_k}$ is the product of gamma densities with parameters α_k and β_k,

$$p(\boldsymbol{\Theta_k}|\alpha_k, \beta_k) = \prod_{p<p'} \frac{\beta_k^{\alpha_k}}{\Gamma(\alpha_k)} (\theta_{pp'}^{(k)})^{\alpha_k - 1} \exp(-\beta_k \theta_{pp'}^{(k)}). \quad (5)$$

For example, in our application to kidney cancer, genes are involved in four distinct non-overlapping groups (see Figure 2): $\boldsymbol{E}_1$, $\boldsymbol{E}_2$, $\boldsymbol{E}_3$ and $\boldsymbol{E}_5$. Group $\boldsymbol{E}_2$, which consists of only 70 samples, borrows information from larger groups (through parameters θ_{12}, θ_{13} and θ_{14}) to encourage the selection of common markers. A detailed discussion of the effect of the MRF priors on the selection of prognostic markers is presented in Web Appendix B.

Posterior inference on the $\theta_{pp'}^{(k)}$'s will describe how strongly this modeling assumption is supported by the data.

4. Posterior Inference

Our statistical goal is to conduct inference on the selection of genomic markers and on the relatedness parameter of our MRF priors. We focus on the posterior distribution $p(\{\boldsymbol{\Gamma_k}, \boldsymbol{\Theta_k}\}_{k=1}^K | \boldsymbol{y^\star}, \boldsymbol{P}, \boldsymbol{X})$. See Web Appendix C for a full explanation. Since this posterior distribution is intractable, we construct a Markov chain Monte Carlo (MCMC) sampler. For that, we integrate out all the regression coefficients and the variance parameters by approximating the marginal likelihood using a standard Laplace approximation (Johnson and Rossell, 2012; Chekouo et al., 2015). A detailed description of our algorithm can be found in Web Appendix D.

4.1. *Prediction*

We use an M-fold cross-validation where the samples are partitioned into M subsets for each group of samples s. We set $M = 10$ for our application, described in Section 5. For each sample $n = 1, .., N_s$ in group s, we denote with $\kappa(n)$ the partition that contains n (validation set). Let $\boldsymbol{y}_{\kappa(n)}$ be the vector of the survival time in the same partition $\kappa(n)$, and $\boldsymbol{y}_{-\kappa(n)}$ be the observations from the remaining partitions (training set). Let $\hat{\boldsymbol{y}}$ be the augmented survival times for all subjects, where the censored times are replaced by the mean of the values sampled in the MCMC algorithm. We similarly define $\boldsymbol{y}_{\kappa(n)}^{(s)}$, $\boldsymbol{y}_{-\kappa(n)}^{(s)}$, $\hat{\boldsymbol{y}}_{\kappa(n)}^{(s)}$, and $\hat{\boldsymbol{y}}_{-\kappa(n)}^{(s)}$ for each group s. The cross-validation density for the n-th individual in group s is estimated via model averaging; following Gelfand and Dey (1994) and Lamnisos et al. (2012) we use $p(\boldsymbol{\Gamma}^{(s)}|\hat{\boldsymbol{y}})$ as the importance density of the target distribution $p(\boldsymbol{\Gamma}^{(s)}|\hat{\boldsymbol{y}}_{-\kappa(n)})$, where $\boldsymbol{\Gamma}^{(s)}$ is the set of variable selection parameters involved in equation s. The predictive probability can then be written as

$$\begin{aligned} p(y_n^{(s)}|\hat{\boldsymbol{y}}_{-\kappa(n)}) &= \sum_{\boldsymbol{\Gamma}^{(s)}} p\big(y_n^{(s)}|\hat{\boldsymbol{y}}_{-\kappa(n)}^{(s)}, \boldsymbol{\Gamma}^{(s)}\big) p\big(\boldsymbol{\Gamma}^{(s)}|\hat{\boldsymbol{y}}_{-\kappa(n)}\big) \\ &\approx \sum_{l=1}^{S} w_l p\big(y_n^{(s)}\big|\hat{\boldsymbol{y}}_{-\kappa(n)}^{(s)}, \boldsymbol{\Gamma}^{(s)}(l)\big) \Big/ \sum_{l=1}^{S} w_l, \quad (6) \end{aligned}$$

where $w_l = \dfrac{p(\boldsymbol{\Gamma}^{(s)}(l)|\hat{\boldsymbol{y}}_{-\kappa(n)})}{p(\boldsymbol{\Gamma}^{(s)}(l)|\hat{\boldsymbol{y}})} \propto [p(\hat{\boldsymbol{y}}_{\kappa(n)}^{(s)}|\hat{\boldsymbol{y}}_{-\kappa(n)}^{(s)}, \boldsymbol{\Gamma}^{(s)}(l))]^{-1}$, and S is the number of models $\boldsymbol{\Gamma}^{(s)}(l)$'s of equation s obtained from the MCMC samples. We approximate the densities $p(y_n^{(s)}|\boldsymbol{y}_{-\kappa(n)}^{(s)}, \boldsymbol{\Gamma}^{(s)}(l))$ and $p(\hat{\boldsymbol{y}}_{\kappa(n)}^{(s)}|\hat{\boldsymbol{y}}_{-\kappa(n)}^{(s)}, \boldsymbol{\Gamma}^{(s)}(l))$ by the respective student distributions $p(y_n|\boldsymbol{y}_{-\kappa(n)}^{(s)}, \boldsymbol{\beta}_l^\star, \boldsymbol{\Gamma}^{(s)}(l))$ and $p(\hat{\boldsymbol{y}}_{\kappa(n)}^{(s)}|\boldsymbol{y}_{-\kappa(n)}^{(s)}, \boldsymbol{\beta}_l^\star, \boldsymbol{\Gamma}^{(s)}(l))$, where $\boldsymbol{\beta}_l^\star$ is the maximum likelihood estimator of $p(\boldsymbol{y}_{-\kappa(n)}^{(s)}|\boldsymbol{\beta}_l^\star, \boldsymbol{\Gamma}^{(s)}(l))$ (Raftery et al., 1996). Thus, the survival time for subject $n' \in \kappa(n)$ in group $\kappa(n)$ can be

estimated as

$$y_{n'}^{\text{pred}} = \sum_{l=1}^{S} \left[p\left(\hat{\boldsymbol{y}}_{\kappa(n)}^{(s)} \middle| \hat{\boldsymbol{y}}_{-\kappa(n)}^{(s)}, \boldsymbol{\Gamma}^{(s)}(l) \right) \right]^{-1} \boldsymbol{Z}_{n'(l)} \boldsymbol{\beta}_l^{\star} \Big/ \sum_{l=1}^{S} \left[p\left(\hat{\boldsymbol{y}}_{\kappa(n)}^{(s)} \middle| \hat{\boldsymbol{y}}_{-\kappa(n)}^{(s)}, \boldsymbol{\Gamma}^{(s)}(l) \right) \right]^{-1}, \quad (7)$$

where $\boldsymbol{Z}_{n'(l)}$ are the observed prognostic factors for subject n', and $\boldsymbol{\Gamma}^{(s)}(l)$ indexes the selected biomarkers for model $\boldsymbol{\Gamma}^{(s)}(l)$, intercept included.

5. Application

We first test our model through extensive simulation studies and then highlight the applicability of our proposed method by analyzing the KIRC data described in Section 2.

5.1. *Hyperparameter Setting*

For both the simulations and KIRC data analysis, we set the hyperparameters as follows. A vague prior is assigned to the intercept parameters $\beta_0^{(s)}$ by specifying the hyperparameter τ_0 as a large value tending to ∞. Specifically, we set $\tau_0 = 10^4$. The hyperparameters of the prior distributions τ_k, $k = 1, ..., K$ of the selected coefficients are set to $\tau_k = 0.087$, which correspond to the prior probability of 0.1 of belonging to the interval $(-0.1, 0.1)$. We set $\nu_1 = \nu_3 = -4$ and $\nu_2 = -3$, which correspond to a prior probability of inclusion of 1.8%, and 4.7%, respectively, for genes, methylation probe and miRNA selection, when all neighbors in the MRF prior have 0 value. We set $\alpha = 0.001$ and $\psi = 0.001$ to have a vague gamma prior on σ_s^{-2}. Last, we set $\alpha_k = 40$ and $\beta_k = 10$ on the $\boldsymbol{\Theta}_k$ prior to encourage the selection of the same biomarkers through the different regression models.

5.2. *Simulated Data Analysis*

We investigate the performance of our method using simulated survival times that mimic the characteristics of the observed data. The aim of this analysis is three-fold: (i) identify the set of significant biomarkers, (ii) study the performance of our models with respect to the degree of overlapping sets of biomarkers, and (iii) study the prediction performance of the proposed approach. Aligning the analyses with the three aims, we assess the results on a group basis, whereby a group is defined as the unique samples modeled using either $\boldsymbol{E_1}$, $\boldsymbol{E_2}$, $\boldsymbol{E_3}$ or $\boldsymbol{E_5}$ models. We randomly choose ten genes, six miRNAS and ten methylation probes to be associated with survival time for each of the four groups from three scenarios: (i) the selected biomarkers are the same in each equation where they are involved; (ii) the selected biomarkers from each platform have a 50% overlap across groups, and (iii) no overlap exists between the sets of selected biomarkers. The survival times are simulated from AFT models, equation (1), with 50% censored observations and without prognostic factors; the non-zero regression coefficients are set to either 1 or -1, and the intercept terms to $\beta_0^{(s)} = 2$. To ensure that the realistic pattern of correlation across all biomarkers is preserved, we use the matrices of observed expression of mRNA, miRNA, and methylation probes as covariates.

To investigate the effect of the MRF prior on both selection and prediction, we compared our method (IMR) with a modified version, Bayesian multi-steps (BMS), which does not include the MRF interaction parameters $\boldsymbol{\Theta_k}$. This method is simply defined by setting $\theta_{jj'}^{(k)} = 0$ for every j, j'. The MCMC sampler of our IMR and BMS methods was run for 400,000 iterations, with 50,000 discarded as burn-in. We also compared the performance of these two methods with those of the L_1 penalized Cox proportional hazard model (L1-CPH) of Simon et al. (2011) by independently fitting a regression model for each group. The optimal tuning parameter was chosen from a 10-fold cross-validation and implemented with the *glmnet* R package. Finally, we compared our approach, in terms of variable selection, to a univariable analysis where we fit a CPH regression for each biomarker (Uni-CPH); p-values were adjusted following the Hommel procedure (Hommel, 1988). This simple approach has been implemented in several biomedical studies and has the advantage of maximally exploiting the available samples for each univariable analysis.

To assess performance in terms of variable selection, we computed the area under the receiver operating characteristic curve (AUC) for each biomarker platform g in each regression model s. The AUC was computed using the marginal posterior probabilities for each biomarker. To evaluate prediction performance, for each group of samples, we reported the concordance index (C-index), which is defined as the proportion of patient pairs in which the predictions and outcomes are concordant. The predictions are computed as in formula (7) using cross-validation. The C-indices of our full model were computed using the predicted survival time in each model, while for L1-CPH, we used the arithmetic mean weighted by the sample size in each group. Figure 3 shows the box plots of the AUCs for genes, miRNAs and methylation probes; the bottom-right panel shows the box plots of the C-indices. We observe that the IMR and BMS methods perform better than the L1-CPH and Uni-CPH in all scenarios. When the number of overlapped biomarkers selected between the equations is large, the IMR performs much better than the BMS in terms of variable selection, while remaining competitive with respect to prediction accuracy. As expected, we did not observe a significant difference between IMR and BMS in scenario 3, as the sets of relevant biomarkers do not overlap. These results reinforce the importance of using a MRF prior in the inferential process, particularly when some relevant biomarkers overlap across the regression models.

We performed additional simulations to demonstrate the robustness of our approach in the case of highly correlated prognostic markers. We studied how the extent of the correlation between selected markers and an unobserved gene impacts our results. Our approach performs well in terms of both variable selection (AUC) and prediction (C-index), even in the presence of highly correlated markers (see Web Appendix E for more details).

5.3. *Application to KIRC Data*

In this section, we summarize our integrative analysis of the TCGA KIRC dataset described in Section 2. To assess the clinical utility in terms of prediction, we fitted three models: (i) a multivariable Cox proportional hazard model, using only clinical data (prognostic factors such as age, gender,

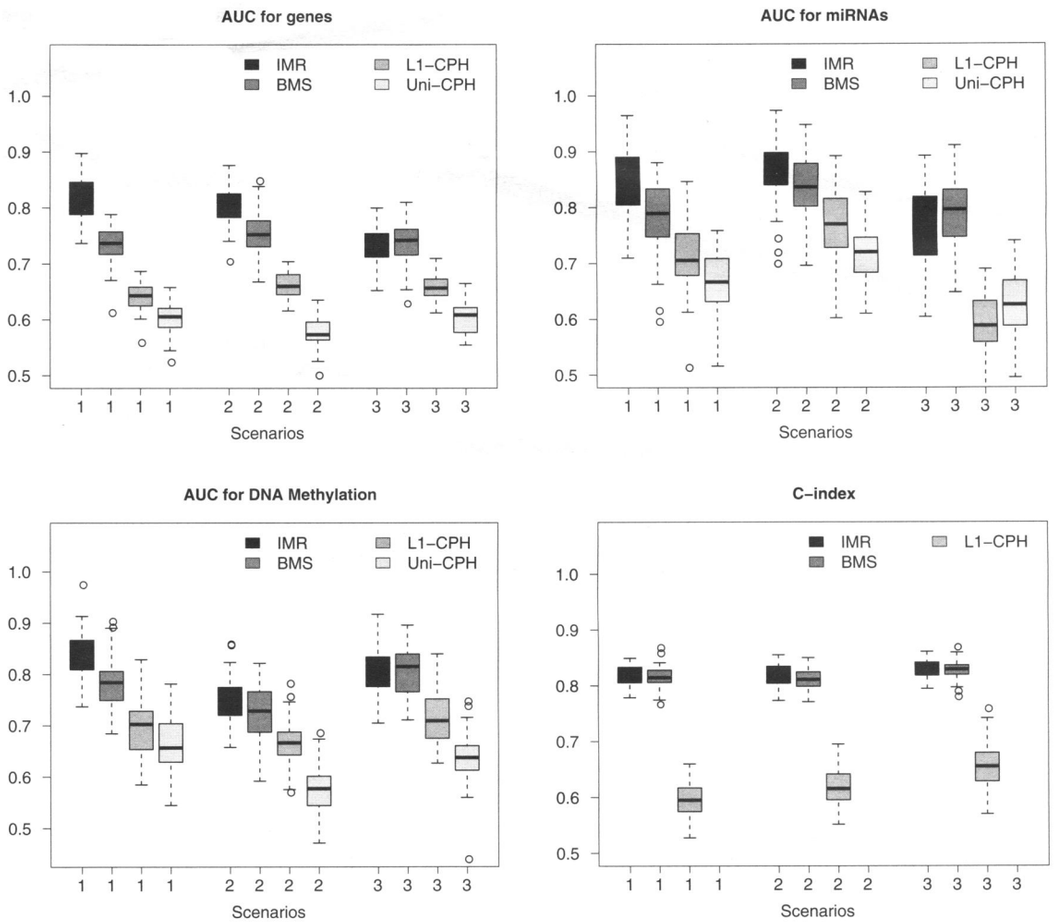

Figure 3. Simulation results: Box plots of the AUCs for variable selection and C-indices for prediction on 50 replicates.

tumor stage and grade) as covariates (CPH+C); (ii) our IMR model, using only molecular data (IMR+M), and (iii) our IMR model, using both clinical data and molecular data (IMR+C+M). To the best of our knowledge, the CPH+C model is the most commonly used for the analysis of multivariable prognostic models in biomedical studies. To assess the MCMC convergence of the IMR+M and IMR+C+M models, we fitted 8 MCMC chains with different starting points. We assessed the agreement of the results among the 8 chains by evaluating the correlation coefficients between the marginal posterior probabilities for biomarker selection. These indicated good concordance between the 8 chains, with all correlations > 0.85 for each type of biomarker in any regression model. Additional MCMC diagnostic checks can be found in Web Appendix F.

Table 1 shows the predictive performances of the aforementioned models. The results in this table were obtained by using multiple (ten times) 10-fold cross-validations (which gives 100 random validation test sets). C-indices were summarized by averaging them over the 100 validation test sets. We observe that, by including both the clinical prognostic factors and molecular data in the model, we have globally improved the predictive power. This improvement is further accentuated in groups $\boldsymbol{E_2}$ and $\boldsymbol{E_5}$, which have either a small sample size or a small number of potential biomarkers. Based on the analysis of the molecular data, our integrative method (IMR) performs better than the non-integrative method (BMS), particularly in the small group $\boldsymbol{E_2}$:G+mi. A similar trend is observed in the analysis of the clinical and molecular data, particularly in group $\boldsymbol{E_1}$:G+mi+Me. The biomarkers described in the following results were selected by using the combination of genomic and clinical information.

A list of the biomarkers with highest marginal posterior probabilities (MPP) is reported in Web Appendix G. One of the genes selected by IMR is PLAC8 (placenta-specific 8, MPP = 0.26), with low expression associated with longer survival time. PLAC8 is a known oncogene that has been studied for its role in differential apoptosis in different cell types (Mourtada-Maarabouni et al., 2013). Other genes that were found within the top ten have previously been identified as either oncogenes or genes directly related to cancer pathogenesis, including ZFP82, NUDT1 and NUDT17. Of the methylation probes assessed, we identified a probe within the well-known p53 tumor suppressor gene with a very high MPP (MPP = 0.85), closely followed by another methylation probe in the zinc finger protein 792 (ZFP792, MPP=0.67),

Table 1

Comparison of predictive performance. CPH is the multivariable Cox proportional hazard model, IMR is the proposed integrative Bayesian approach. The covariates used to fit the model follow the "+ symbol: C is the clinical covariates (prognostic factors such as age, gender, tumor stage and grade) and M is the molecular data (genes, miRNAs, and methylation probes). $\boldsymbol{E_1}$*: G+mi+Me indicates the regression model with genes, miRNAs and methylation data as potential biomarkers.* $\boldsymbol{E_2}$*: G+mi indicates the regression model built with samples having only mRNA and miRNA expression data.* $\boldsymbol{E_3}$*: G+Me indicates the regression model built with samples that have only mRNA expression and methylation data.* $\boldsymbol{E_5}$*: G indicates the regression model with only mRNA expression data. Parenthetic values are standard errors.*

Models	G+mi+Me ($\boldsymbol{E_1}$)	G+mi($\boldsymbol{E_2}$)	G+Me($\boldsymbol{E_3}$)	G ($\boldsymbol{E_5}$)	Full Model
CPH + C	0.72 (0.01)	0.76 (0.01)	0.79 (0.01)	0.71 (0.01)	0.71 (0.01)
IMR + C + M	0.89 (0.02)	0.82 (0.04)	0.89 (0.02)	0.80 (0.04)	0.86 (0.01)
BMS + C + M	0.87 (0.03)	0.82 (0.04)	0.89 (0.02)	0.79 (0.04)	0.85 (0.01)
IMR + M	0.88 (0.02)	0.71 (0.06)	0.84 (0.02)	0.72 (0.05)	0.82 (0.01)
BMS + M	0.88 (0.02)	0.67 (0.06)	0.84 (0.02)	0.72 (0.05)	0.81 (0.01)

with decreasing DNA methylation at both probe sites associated with longer survival times. Of the miRNA markers identified, meta-miRNA 9 (see Figure 2 in Web Appendix A) had the strongest MPP (MPP=0.55), including information from five miRNA markers (two of which were variants of the same miRNA). This meta-miRNA had a positive relationship, with increased miRNA expression associated with shorter survival time. At least three of these miRNA markers (mir-185 Imam et al., 2010; mir-142 Isobe et al., 2014; and mir-942 Liu et al., 2014) have been shown to be associated with poor outcomes and increased tumor growth and activity, confirming the positive association identified.

Due to the impact of the MRF priors on our models, some biomarkers were found simultaneously in more than one regression model. For example, gene NME2P1 (non-metastatic cells 2, protein (NM23B) expressed in pseudogene 1) was identified as a top biomarker by IMR in groups $\boldsymbol{E_1}$ and $\boldsymbol{E_5}$, but was not selected in group $\boldsymbol{E_1}$ by BMS. This gene, which has been identified as a nucleoside diphosphate kinase, is a candidate tumor suppressor protein and a presumed regulator of tumor metastasis (Marino et al., 2013). Identified in both $\boldsymbol{E_1}$ and $\boldsymbol{E_2}$ models, miRNA mir-335 binds many cancer-specific target genes, and has been shown to be associated with breast (Tavazoie et al., 2009), lung (Tavazoie et al., 2009), kidney (White et al., 2011), gastric (Xu et al., 2014) and colorectal (Wan et al., 2014) cancers. A second miRNA (mir-130a) found in both models has also been published widely as a potential gene regulator in colorectal (Liu et al., 2013), pancreatic (Zhao et al., 2013) and endometrial (Li et al., 2013) cancers. Both miRNA markers have been shown to be associated with disregulation of genes, leading to a worse cancer prognosis and increased tumor size, matching the positive association with shorter survival time found in this research.

To gain further biological insights into our results, we performed functional analysis using Ingenuity Pathway Analysis (IPA) software, with the top selected biomarkers from each regression model. For the selected methylation probes, we uploaded the genes co-located with each methylation probe to assess those genes regulated via DNA methylation. Figures 4 and 5 depict IPA networks for the genes identified from each of the $\boldsymbol{E_1}$ and $\boldsymbol{E_5}$ groups. Other IPA networks identified using top selected genes, miRNAs and DNA methylation probes, are shown in Web Appendix G. Interestingly, all pathways identified by this functional analysis clustered around a central hub marker ubiquitin C (UBC), a gene associated with protein degradation, DNA repair, and cell cycle regulation. This gene interconnects many other genes in the networks as a hub protein, and has been involved in kidney cancer and breast cancer, and contributes to cancer metastasis (Wu et al., 2014). Furthermore, two smaller hub genes, CTNNB and ERBB2, were seen in gene networks $\boldsymbol{E_1}$ and $\boldsymbol{E_5}$ respectively and are known widely to contribute to cancer pathology.

6. Discussion

We proposed an innovative Bayesian hierarchical model for integrating multi-regression models (IMR) to identify a small set of potential biomarkers in the context of AFT models. Motivated by the availability of multi-platform genomic data, our model aimed to integrate the maximal number of samples across mRNA, miRNA and methylation platforms to i) find clinically relevant combinations of biomarkers associated with survival time and ii) combine genomic markers with clinical information to improve predictive performances. We defined MRF priors to encourage the selection of the same biomarkers across equations, and showed through simulation studies that our model outperforms competing approaches that do not borrow strength across equations. The application of this methodology to TCGA KIRC data reinforces (i) the importance of our modeling approach, which encourages the selection of the same relevant biomarkers across groups, and (ii) that optimally combining clinical and genomic information for all samples leads to a better prediction performance. An important finding from the application of the novel IMR method to the KIRC data stems from the incorporation of the MRF priors into the models. As a direct effect of the specification of the MRF priors, we were able to detect the gene NME2P1 and the miRNA markers mir-335 and mir-130a across two groups of samples, both of which have been previously associated with cancer progression. Further, for these and other markers identified, investigation into the biology unraveled plausible links between expression levels (mRNA, miRNA) and methylation patterns with survival times. Since cancer is a complex multifactorial disease, it is interesting

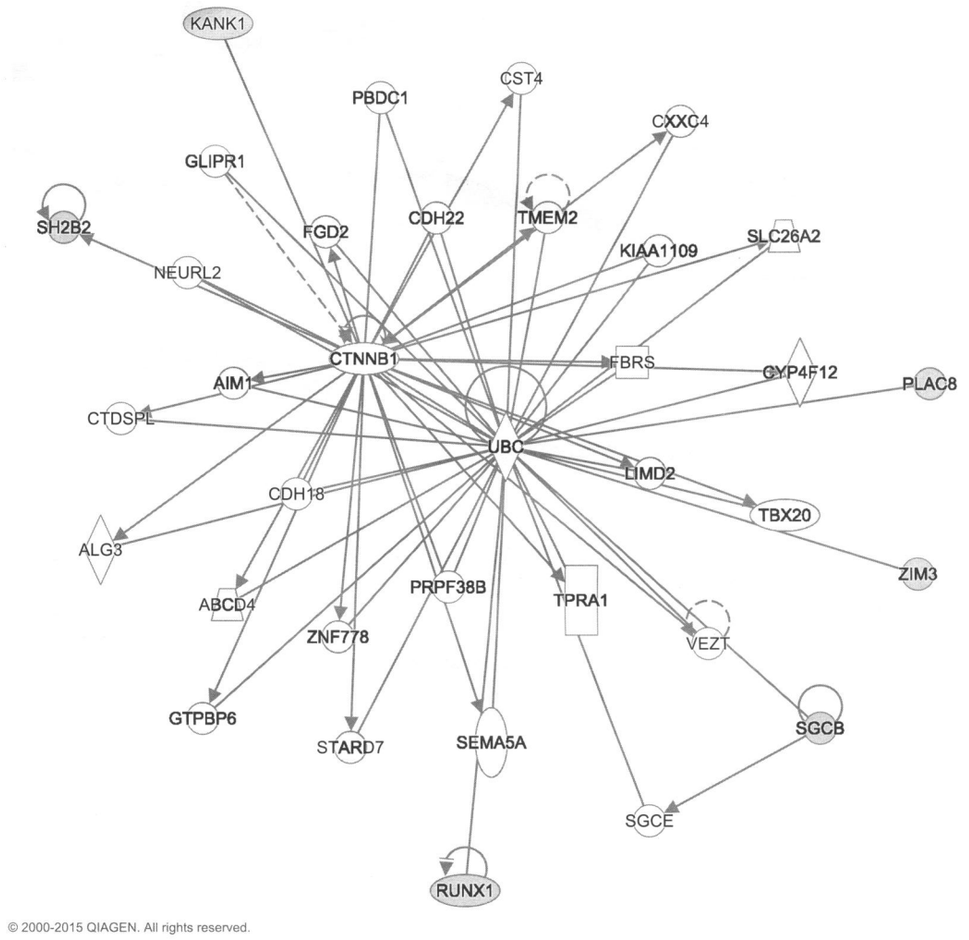

Figure 4. Gene network from group $\boldsymbol{E_1}$ (i.e., consisting of samples from all 3 platforms: mRNA, miRNA and DNA methylation): Network of cancer, cellular development, organismal injury and abnormalities.

that our IMR model identified many genes linked to the UBC gene, which is involved in multiple cancer-specific pathways, including cell death, cell cycle regulation, and DNA repair.

An alternative approach to IMR would consist of implementing a data augmentation scheme, hence simplifying the problem into a vanilla Bayesian variable selection problem. However, given the large amount of missing data and the type of missing data patterns, we decided that a data augmentation approach is not feasible in this context (see Web Appendix H for a detailed discussion). Our model could be extended to integrate group-specific prior information by including a parameter $\nu_{k,s}$ for each group $s = 1, ..., 2^{K-1}$. This would give additional flexibility to allow groups to have different degrees of sparsity. For example, we can favor particular biomarkers in groups with fewer numbers of samples or numbers of biomarker types. This implies that information is shared across the groups by both $\boldsymbol{\Theta_k}$ and $\nu_{k,s}$. For future work, we plan to extend our framework to include novel priors that account for the biological interactions between the biomarkers. Other methods such as survival trees and survival forests offer promising algorithms that can be applied in the context of high-dimensional survival data. Survival trees, which use binary decisions to recursively split observations into groups of similar hazard rates, can handle missing data directly in their algorithm by using surrogate splits or multiple imputation techniques (Ding and Simonoff, 2010; Bou-Hamad et al., 2011). We plan to pursue a Bayesian paradigm of this approach as future work; we will explore how tree-based regression can be applied to multiplatform genomic data with many missing values across platforms.

The current study aimed to investigate the strength of the novel Bayesian multi-regression approach for the integrative analysis of genomic data. Using the maximal number of samples from different platforms, the proposed method showed promising results in both simulation studies and in the analysis of the TCGA KIRC dataset. Our application to the KIRC dataset identified several prognostic biomarkers, most of which were previously identified. Most importantly, these selected biomarkers were clearly linked with well published cancer genes such as ERRB2, CTBBN and UBC.

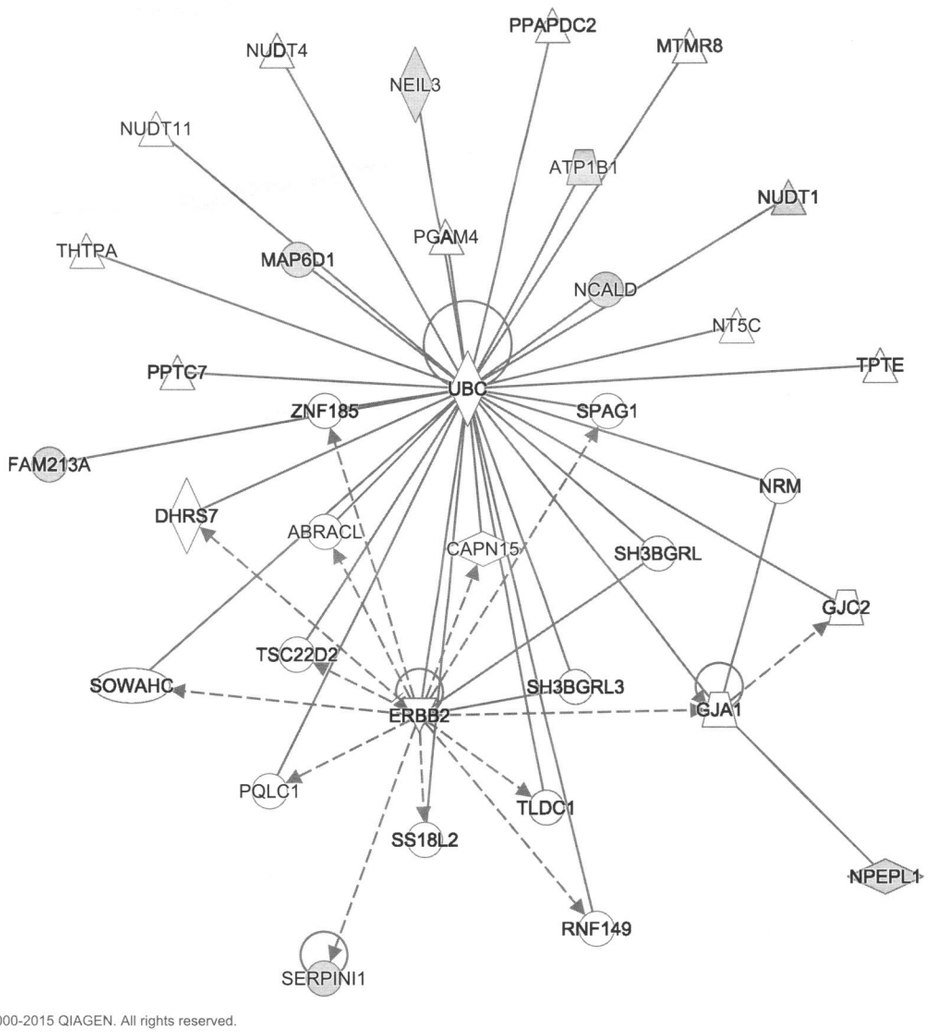

Figure 5. Gene network from group $\boldsymbol{E_5}$ (i.e., consisting of samples with mRNA expression data only): Network of DNA replication, recombination, and repair nucleic acid metabolism, small molecule biochemistry.

7. Supplementary Materials

Web Appendices A-H, referenced in Sections 2-6, the TCGA kidney cancer data analyzed in Section 5, and the accompanying computer code written in C are available with this paper at the Biometrics website through the Wiley Online Library.

Acknowledgements

F.C. Stingo and K.-A. Do are partially supported by a Cancer Center Support Grant (NCI Grant P30 CA016672). The authors thank LeeAnn Chastain for editing assistance.

References

Bou-Hamad, I., Larocque, D., and Ben-Ameur, H. (2011). A review of survival trees. *Statistics Surveys* **5**, 44–71.

Bush, W. S., Moore, J. H. (2012) Chapter 11: Genome-Wide Association Studies. PLoS Comput Biol 8(12):e1002822. doi:10.1371/journal.pcbi.1002822"

Chekouo, T., Stingo, F. C., Doecke, J. D., and Do, K.-A. (2015). miRNA-target gene regulatory networks: A Bayesian integrative approach to biomarker selection with application to kidney cancer. *Biometrics* **71**, 428–438.

Daemen, A., Gevaert, O., Bie, T. D., Debucquoy, A., Machiels, J.-P., Moor, B. D., and Haustermans, K. (2008). Integrating microarray and proteomics data to predict the response of cetuximab in patients with rectal cancer. In *Pacific Symposium on Biocomputing*, Altman R. B., A. K. Dunker, L. Hunter, T. Murray, and T. E. Klein (eds), 166–177. World Scientific.

Ding, Y. and Simonoff, J. S. (2010). An investigation of missing data methods for classification trees applied to binary response data. *Journal of Machine Learning Research* **11**, 131–170.

Gelfand, A. E. and Dey, D. K. (1994). Bayesian Model Choice: Asymptotics and Exact Calculations. *Journal of the Royal Statistical Society, Series B (Methodological)* **56**, 501–514.

Hamid, J. S., Hu, P., Roslin, N. M., Ling, V., Greenwood, C. M. T., and Beyene, J. (2009). Data integration in genetics and

genomics: methods and challenges. *Human Genomics Proteomics* **2009**.

Hastie, T., Tibshirani, R., Eisen, M., Alizadeh, A., Levy, R., Staudt, L., et al. (2000). 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biology* **1**, Genome Biology, 1(2):research0003.1–0003.21.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified bonferroni test. *Biometrika* **75**, 383–386.

Imam, J., Buddavarapu, K., Lee-Chang, J., Ganapathy, S., Camosy, C., Chen, Y., et al. (2010). MicroRNA-185 suppresses tumor growth and progression by targeting the Six1 oncogene in human cancers. *Oncogene.* **35**, 4971–4979.

Isobe, T., Hisamori, S., Hogan, D., Zabala, M., Hendrickson, D., Dalerba, P., et al. (2014). miR-142 regulates the tumorigenicity of human breast cancer stem cells through the canonical WNT signaling pathway. *Elife.* pages 4971–4979.

Johnson, V. E. and Rossell, D. (2012). Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association* **107**, 649–660.

Laird, P. W. (2003). The power and the promise of dna methylation markers. *Nature Review Cancer* **3**, 253–266.

Lamnisos, D., Griffin, J. E., and Steel, M. F. J. (2012). Cross-validation prior choice in bayesian probit regression with many covariates. *Statistics and Computing* **22**, 359–373.

Li, B., Lu, C., Lu, W., Yang, T., Qu, J., Hong, X., et al. (2013). miR-130b is an EMT-related microRNA that targets DICER1 for aggression in endometrial cancer. *Medical Oncology* **1**, 484.

Li, F. and Zhang, N. R. (2010). Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. *Journal of the American Statistical Association* **105**, 1202–1214.

Lin, S. and Gregory, R. (2015). MicroRNA biogenesis pathways in cancer. *Nature Reviews Genetics* **15**, 321–333.

Liu, L., Nie, J., Chen, L., Dong, G., Du, X., Wu, X., et al. (2013). The Oncogenic Role of microRNA-130a/301a/454 in Human Colorectal Cancer via Targeting Smad4 Expression. *PLoS ONE* **8**, e55532.

Liu, N., Zuo, C., Wang, X., Chen, T., Yang, D., Wang, J., et al. (2014). miR-942 decreases TRAIL-induced apoptosis through ISG12a downregulation and is regulated by AKT. *Oncotarget.* **13**, 4959–4971.

Lu, L. J., Xia, Y., Paccanaro, A., Yu, H., and Gerstein, M. (2005). Assessing the limits of genomic data integration for predicting protein networks. *Genome research* **15**, 945–953.

Marino, N., Marshall, J., Collins, J., Zhou, M., Qian, Y., Veenstra, T., et al. (2013). Nm23-h1 binds to gelsolin and inactivates its actin-severing capacity to promote tumor cell motility and metastasis. *Cancer Research* **19**, 5949–5962.

Mourtada-Maarabouni, M., Watson, D., Munir, M., Farzaneh, F., and Williams, G. (2013). Apoptosis suppression by candidate oncogene PLAC8 is reversed in other cell types. *Curr Cancer Drug Targets* **1**, 80–91.

Peterson, C., Stingo, F. C., and Vannucci, M. (2015). Bayesian inference of multiple gaussian graphical models. *Journal of the American Statistical Association* **110**, 159–174.

Qin, L.-X. (2008). An integrative analysis of microRNA and mRNA expression–A case study. *Cancer Informatics* **6**, 369–379.

Raftery, A. E., Madigan, D., and Volinsky, C. T. (1996). Accounting for Model Uncertainty in Survival Analysis Improves Predictive Performance (with Discussion). In *Bayesian Statistics 5.* Oxford, UK: Oxford University Press.

Simon, N., Friedman, J. H., Hastie, T., and Tibshirani, R. (2011). Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of Statistical Software* **39**, 1–13.

Srivastava, S., Wang, W., Manyam, G., Ordonez, C., and Baladandayuthapani, V. (2013). Integrating multi-platform genomic data using hierarchical bayesian relevance vector machines. *EURASIP J. Bioinformatics and Systems Biology* **2013**, 9.

Stingo, F. C., Chen, Y. A., Tadesse, M. G., and Vannucci, M. (2011). Incorporating biological information into linear models: A Bayesian approach to the selection of pathways and genes. *The Annals of Applied Statistics* **5**, 1978–2002.

Stingo, F. C., Chen, Y. A., Vannucci, M., Barrier, M., and Mirkes, P. E. (2010). A Bayesian graphical modeling approach to microRNA regulatory network inference. *Annals of Applied Statistics* **4**, 2024–2048.

Stingo, F. C., Vannucci, M., and G., D. (2012). Bayesian Wavelet-based Curve Classification via Discriminant Analysis with Markov Random Tree Priors. *Statistica Sinica* **22**, 465–488.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* **82**, 528–540.

Tavazoie, S., Alarcn, C., Oskarsson, T., Padua, D., Wang, Q., Bos, P. D., Gerald, W., and J., M. (2009). Endogenous human microRNAs that suppress breast cancer metastasis. *Nature.* **7175**, 147–152.

Wan, T., Lam, C., Ng, L., Chow, A., Wong, S., Li, H., et al. (2014). The clinicopathological significance of miR-133a in colorectal cancer. *Disease Markers* page Epub.

Wang, W., Baladandayuthapani, V., Morris, J. S., Broom, B. M., Manyam, G., and Do, K.-A. (2013). iBAG: integrative Bayesian analysis of high-dimensional multiplatform genomics data. *Bioinformatics* **29**, 149–159.

White, N. M., Bao, T. T., Grigull, J., Youssef, Y. M., Girgis, A., Diamandis, M., et al. (2011). mirna profiling for clear cell renal cell carcinoma: Biomarker discovery and identification of potential controls and consequences of mirna dysregulation. *The Journal of Urology* **186**, 1077–1083.

Wu, S., Xu, Y., Feng, Z., Yang, X., Wang, X., and Gao, X. (2012). Multiple-platform data integration method with application to combined analysis of microarray and proteomic data. *BMC Bioinformatics* **13**, 320.

Wu, X., Zhang, W., Font-Burgada, J., Palmer, T., Hamil, A. S., Biswas, S. K., et al. (2014). Ubiquitin-conjugating enzyme ubc13 controls breast cancer metastasis through a tak1-p38 map kinase cascade. *Proceedings of the National Academy of Sciences* **111**, 13870–13875.

Xu, X., Zhang, Y., Zhang, W., Li, T., Gao, H., and Wang, Y. (2014). MicroRNA-133a functions as a tumor suppressor in gastric cancer. *Journal of Biology Regulators and Homeostatics Agents* **4**, 615–624.

Zhao, G., Zhang, J., Shi, Y., Qin, Q., Liu, Y., Wang, B., et al. (2013). MiR-130b is a prognostic marker and inhibits cell proliferation and invasion in pancreatic cancer through targeting STAT3. *PLoS One* **9**, e7308.

Received October 2015. *Revised May* 2016.
Accepted August 2016.