STATISTICAL METHODS FOR INTEGRATING GENOMICS DATA

A Dissertation

by

ELIZABETH JENNINGS MCGUFFEY

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Raymond J. Carroll
Committee Members,	Jeffrey S. Morris
	Veerabhadran Baladandayuthapani
	Jay R. Walton
Head of Department,	Valen E. Johnson

May 2015

Major Subject: Statistics

ABSTRACT

This dissertation focuses on methodology to integrate multiplatform genomic data with cancer applications. Such integration facilitates the discovery of biological information crucial to the development of targeted treatments. We present iBAG (integrative Bayesian Analysis of Genomics data), a two-step hierarchical Bayesian model that uses the known biological relationships between genetic platforms to integrate an arbitrary number of platforms in a single model. This method identifies genes important to a clinical outcome, such as survival, and the integration approach also allows us to identify which platforms are modulating the important gene effects. A glioblastoma multiforme (GBM) data set publicly available from The Cancer Genome Atlas (TCGA) is analyzed with iBAG. We flag several genes as important to survival time, and we include a discussion of these genes in a biological context. We then present a nonlinear formulation of iBAG, which increases the flexibility of the model to accommodate nonlinear relationships among the data platforms. The TCGA GBM data is again analyzed, and we carefully compare the results from both the linear and nonlinear formulation. Next we present a pathway iBAG model, piBAG, which includes gene pathway membership information and utilizes hierarchical shrinkage to simultaneously select important genes and assign pathway scores. The integration of multiple genomic platforms again allows us to determine which platform is regulating each important gene, and it also provides insight as to through which platform each pathway is taking effect. We apply this method to a different subset of the TCGA GBM data. Finally, we present integrative heatmaps, a novel visualization tool for illustrating integrated data. We use a TCGA colorectal cancer data set to demonstrate the integrative heatmaps. Through

the various simulation studies and data applications in this dissertation, we conclude that the methods presented achieve their respective goals and outperform standard methods. We demonstrate that our methods provide many advantages, including increased estimation efficiency, increased power, lower false discovery rates, and deeper biological insight into the genetic mechanics of cancer development and progression.

DEDICATION

To my wonderful husband and family. Thank you, Spencer, Mom, Dad, and Anthony, for your unwavering love and support.

ACKNOWLEDGEMENTS

I am immensely grateful to my chair, Dr. Raymond Carroll, for his guidance, expertise, and encouragement. Dr. Carroll was attentive to my interests and helped me get involved in projects perfectly aligned with my passions. He was always available to offer advice and gentle guidance, and he made me feel like a priority. He has demonstrated time and time again what it means to be a world class statistician, and I could not have asked for a better adviser.

I would also like to thank Dr. Jeff Morris and Dr. Veera Baladandayuthapani for their daily mentorship at MD Anderson Cancer Center. They have contributed immeasurably to my professional growth, and I am deeply appreciative for their constant instruction, support, and encouragement.

I would also like to express my gratitude to Dr. Jay Walton for his time spent serving on my committee, as well as for his leadership of the Undergraduate Biology and Mathematics program which facilitated my first experiences with biostatistical research.

I would like to thank Dr. Gani Manyam for his biological expertise, his contributions to our publications, and his patience with my frequent requests for data extraction.

I would like to express my great appreciation to Dr. Michael Longecker for his constant support through all my years in the Statistics Department. His willingness to help with any issue and his faith in my abilities have been invaluable.

Finally I would like to thank Dr. Alan Dabney for introducing me to the field of biostatistics and patiently guiding me through my first research endeavors. I will be forever grateful that he helped me find this path.

TABLE OF CONTENTS

Page

Page

Page

LIST OF FIGURES

FIGURE Page

FIGURE Page

LIST OF TABLES

TABLE Page

TABLE Page

1. INTRODUCTION

The American Cancer Society estimates there will be over 1.6 million new cancer cases in the United States in 2015 (American Cancer Society, 2015a). Targeted therapies are currently at the forefront of research efforts to prevent death and alleviate suffering due to cancer development and growth. Whereas traditional chemotherapy treatments kill healthy cells along with diseased cells, targeted therapies are designed to affect precise molecular targets contributing to the survival and progression of cancerous cells (National Cancer Institute, 2015). However, before a targeted treatment can be developed, we must first identify the appropriate target(s). Many statistical methods have emerged that aim to find such targets. Some of them focus on one type, or platform, of genetic data (such as gene expression), and some include multiple data platforms in their analyses. Integrating multiple genomic platforms has been shown to provide many advantages, such as increased statistical power and decreased false discovery rates (Wang et al., 2013), as well as providing a clearer picture of the involved biological mechanisms. This dissertation focuses on integrative Bayesian methods with the purpose of identifying the genetic entities significantly related to cancer outcomes.

In Chapter 2 we present a brief overview of iBAG (integrative Bayesian Analysis of Genomics data), an adaptation of the method originally proposed by Wang et al. (2013). This two-step hierarchical model integrates an arbitrary number of genomic data platforms and identifies not only the genes important to a clinical response, but also the platform modulating the significant effects. We apply the method to a subset of publicly available glioblastoma multiforme (GBM) data and identify several potential prognostic markers. In Chapter 3 we provide a more in-depth presentation

of iBAG, including an expansive literature review, a thorough biological discussion of the selected GBM markers, and a straightforward algorithm for ease of method application.

We propose a nonlinear formulation of the iBAG model in Chapter 4. This formulation allows for more flexible integration of the data platforms, without creating any interpretation complications. We compare the linear and nonlinear methods and their corresponding results when applied to the GBM data set. In Chapter 5 we present piBAG, a pathway iBAG model that maintains the integration and gene selection properties of the general iBAG method, but also includes gene pathway membership information. We formulate the piBAG model to borrow strength across each pathway, which results in efficient estimation and also provides the framework to estimate pathway scores. The pathway effects on the clinical outcome can be ranked by these scores, allowing us to identify the important gene pathways as well as the important individual genes. The pathway iBAG is applied to a different subset of the GBM data, and we identify four potential prognostic markers.

In Chapter 5 we propose "integrative heatmaps," a novel visualization tool for illustrating the integration step in our iBAG models. We present three variations of the integrative heatmap (IH): additive IHs, componentwise IHs, and platform-specific IHs, and we discuss the unique objectives of each one. Finally, Chapter 6 contains a conclusion with an overarching view of the advantages offered by the methods presented throughout the dissertation.

2. HIERARCHICAL BAYESIAN METHODS FOR INTEGRATION OF VARIOUS TYPES OF GENOMICS DATA*

2.1 Introduction

The central dogma of molecular biology summarizes the steps involved in the passage of genetic information at a molecular level: DNA is transcribed to messenger RNA (mRNA), which is then translated to a protein, which carries out a specific action in an organism. In addition there are also other alterations and interferences, such as epigenetic factors, that can occur at the DNA and/or mRNA levels which affect the ultimate expression of a given gene. In this paper we consider methylation (which occurs at the DNA level and typically results in a silencing of the gene), copy number (which describes an attribute at the DNA level that affects mRNA expression), and mRNA expression (which affects protein expression); these subsequently affect a clinical phenotype (e.g. survival) (see Figure 2.1). In addition, it is believed that the mechanism of cancer development is complex and involves multiple genes (Kanu et al., 2009). It is known that genes interact and are related through certain pathways, and in this paper we focus on genes from important signaling pathways that are believed to affect cancer development processes (Memorial Sloan-Kettering Cancer Center, 2012).

Current technologies allow us to obtain data from the above-mentioned platforms (and many others) for each gene involved in the analysis. The Cancer Genome Atlas (TCGA) is a project that began in 2006 to gather comprehensive genomic data using multiple platforms on over 20 types of cancer (The Cancer Genome Atlas,

*

2012). The increasing availability of such data has motivated the development of methods that seek to improve estimation and prediction regarding genomic effects on cancer outcomes by integrating data from multiple platforms in a single analysis. The incorporation of information from more than one platform has the potential to increase power and lower false discovery rates in identifying markers related to clinical outcomes for cancer patients; such improvements would deepen our understanding of how cancer develops and spreads, offering researchers valuable insight regarding the development of drugs and procedures intended to prevent or inhibit cancer development.

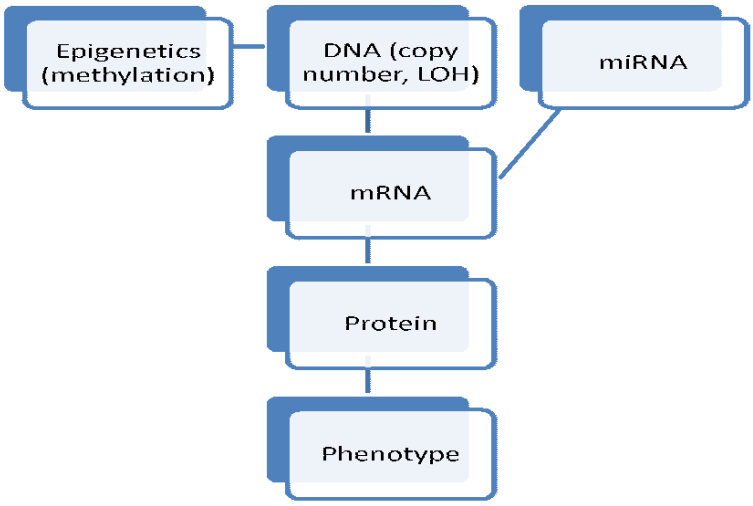


Figure 2.1: Schematic representation of the multiple molecular platforms and their biological relationships. Reprinted with permission from Jennings et al. (2012).

Methods attempting to integrate multiple genomic platforms must face the challenges of high dimensionality and complex biological relationships both within and between platforms. Tyekucheva et al. (2011) suggest a method that includes multiple platforms as predictors in a logistic regression model and show that incorporating

multiple platforms yields more power to detect differentially expressed genes than approaches that only use a single platform – but this approach does not take into account the biological relationships among platforms. Recently, Wang et al. (2013) proposed an integrative Bayesian analysis of genomics data (iBAG) framework that models the biological relationships between two platforms. This approach involves a global gene search, and uses variable selection via the Bayesian lasso-based shrinkage priors to deal with the high dimensionality of the data.

In this paper, we introduce a generalized version of iBAG that integrates data from an arbitrary (multiple) number of genomic platforms using a hierarchical model that incorporates the biological relationships among them. We focus our analysis on genes from the RTK/PI3K, P53, and RB signaling pathways and integrate mRNA, methylation, and copy number data to predict survival in Glioblastoma Multiforme (GBM) patients. In addition, we reduce dimension by regressing the clinical outcome on latent scores of the platforms (see Section 2.2.1 for details). To improve effect size estimation and to achieve sparsity, we use a Normal-Gamma (NG) prior for the effects, which increases flexibility in the estimation as compared to the Laplace prior of the Bayesian lasso (Griffin and Brown, 2010) (see Section 2.2.2 for further discussion). Section 2.3 illustrates our methodology on a synthetic example; analysis of GBM data is presented in Section 2.4; and conclusions are drawn in Section 2.5.

2.2 A Multivariate iBAG Model

Our construction of a multivariate iBAG model employs a two-component hierarchical model where the first component can be considered as the *mechanistic model*, and the second can be considered as the *clinical model*. In the first stage mechanistic model, we partition each gene's expression into the factors explained by methylation, copy number, and other (unknown/unmeasured) causes using a prin-

cipal component-based regression model. Subsequently, we include these factors as predictors in the second stage clinical model, thus finding not only those genes whose expression is directly related to clinical outcome, but also expression effects driven by methylation, copy number, or other mechanisms. We explain the construction of each of these components below.

2.2.1 Mechanistic model

Let n = number of patients, k = number of platforms being integrated, and p_j = number of genes from platform j. The mechanistic model for each gene can be expressed as:

$$\text{mRNA}_i = \text{M}_i + \text{CN}_i + \text{O}_i,$$

where each of the terms are defined as follows:

- mRNA_i is the level of gene expression for gene i (where $i = 1, ..., \max(p_j);\ j = 1, ..., k$) and is of dimension $(n \times 1)$.
- M_i is the part of gene_i expression that is attributed to methylation factors, and is of dimension $(n \times 1)$.
- CN_i is the part of gene_i expression that is attributed to changes in copy number, and is of dimension $(n \times 1)$.
- O_i represents the "other" (remaining) part of the gene expression that is unexplained, and is of dimension $(n \times 1)$.

Since the raw methylation and copy number data for any given gene can consist of multiple (up to 40) values, to estimate each of the components M_i, CN_i, and O_i, we first carry out two principal component analyses (PCA) for gene_i: one each for the methylation and copy number data. We then regress mRNA_i on the methylation

and copy number PC scores that account for $\geq 90\%$ of the variation. We use these estimated pieces and the corresponding residuals from this regression to estimate the vectors M_i, CN_i, and O_i respectively. This process is repeated for each gene independently.

2.2.2 Clinical model

The clinical component of our construction models the effect of the mechanistic parts of the genes (as estimated above) on a clinical outcome of interest (e.g. survival, in our context) and can be written as:

$$Y = \mathrm{M}\boldsymbol{\beta}_1 + \mathrm{CN}\boldsymbol{\beta}_2 + \mathrm{O}\boldsymbol{\beta}_3 + \boldsymbol{\epsilon}$$

where Y denotes the clinical outcome, $\boldsymbol{\beta}_i$ are the effects of platform i on Y, and $\boldsymbol{\epsilon}$ is the error term. The covariates in the model {M, CN, O} are the vectorized gene effects attributed to methylation, copy number, and other sources respectively, and are estimated from the mechanistic model. In essence, our clinical component jointly (additively) models the effects of all the gene expressions and their components – derived from different sources (methylation/copy number) – in a unified manner.

Our goal is to find a list of significant genes that affect the outcome via the various mechanisms; hence efficient estimation of $\boldsymbol{\beta} = \{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3\}$ is of primary interest. For estimation we could simply fit a least squares regression to estimate the parameters. However, the number of predictors is large compared to the number of samples, and, more importantly, we expect our solution to be very sparse since only a few genes will be related to clinical response; hence least squares would overfit the data and yield less accurate results as compared to approaches that induce sparsity by shrinkage/penalization. We illustrate this fact in our simulation in Section 2.3.

Therefore, to estimate the parameters in the clinical model, we specify prior distributions for each model parameter and sample from the posterior distribution

using Markov Chain Monte Carlo (MCMC). Most notably, we assign a Normal-Gamma (NG) prior distribution for $\boldsymbol{\beta}_i$. The two hyperparameters in the NG prior provide increased flexibility in the estimated shrinkage relative to the Laplace prior of the Bayesian lasso (Park and Casella, 2008) which has a single hyperparameter; thus the NG prior leads to improved estimation (Griffin and Brown, 2010). Our complete hierarchical clinical model can be written as:

$$\begin{aligned}
\mathbf{Y} &= \text{Normal}(X\boldsymbol{\beta}, \sigma^2\mathbf{I}_n), \\
\boldsymbol{\beta} &= \text{Normal}(\mathbf{0}_{\tilde{p}}, D_\psi) \text{ where} \\
&\quad D_\psi = \text{diag}(\psi_{1,1}, ..., \psi_{1,p_1}, ..., \psi_{k,1}, ...\psi_{k,p_k}), \\
\psi_{i,j} &= \text{Gamma}(\lambda_i, 1/(2\gamma_i^2)), \\
\sigma^2 &= \text{InverseGamma}(a, b), \\
\lambda_i &= \text{Exponential}(c), \\
\gamma_i^{-2} &= \text{Gamma}(\tilde{a}, \tilde{b}/(2\lambda_i)),
\end{aligned}$$

where $\tilde{p} = \sum_{i=1}^{k} p_i$ is the total number of predictors in the model. With this formulation, the complete conditionals for most parameters are available in closed form – we can use Gibbs sampling to update all parameters except λ_i, which we update using a Metropolis-Hastings random walk step.

2.2.3 Gene selection

Given the posterior samples from the MCMC, we determine which genes are significantly related to clinical outcome using a method based on the median probability model (Barbieri and Berger, 2004). First, we define a minimum effect size which is driven by practical considerations. Since we are analyzing survival data, we use accelerated failure time (AFT) models using log(survival) as the response; thus a δ-fold or larger change in survival for a unit increase in a predictor corresponds to

a $\beta_{i,j}$ outside the region (log(1-δ), log(1+δ)). Denote this region (δ_-^*, δ_+^*). (In our following analyses, we use $\delta = 0.05$ which corresponds to a 5% change in survival time.) If S is the number of MCMC samples and $\beta_{i,j}^{(s)}$ is the $\beta_{i,j}$ sample from iteration s, then $p_+(x_{i,j}) = \sum_{s=1}^{S} \mathbf{I}(\beta_{i,j}^{(s)} > \delta_+^*)/S$ is the posterior probability that $\beta_{i,j}$ is higher than the practical cutoff δ_+^*. Similarly, $p_-(x_{i,j}) = \sum_{s=1}^{S} \mathbf{I}(\beta_{i,j}^{(s)} < \delta_-^*)/S$ is the posterior probability that $\beta_{i,j}$ is lower than the practical cutoff δ_-^*. We flag a gene as "significant" if $p_+(x_{i,j}) > 0.5$ or if $p_-(x_{i,j}) > 0.5$.

2.3 Simulation

We investigate the shrinkage properties of our Bayesian penalized regression formulation of the clinical model as compared to least squares regression through a simulation. We simulate a training dataset with 90 predictors ($k = 3$ platforms with $p_1 = p_2 = p_3 = 30$ predictors from each), where 30 randomly selected $\beta_{i,j}$'s are set exactly to 0 and the other 60 are sampled from a Laplace($\mu = 0$, $b = 1/7$) distribution; this reflects the effective sparsity we expect to see in our data. The other settings for the simulated data are: $n = 100$, $\sigma^2 = 1$, each X entry is from Normal$(0, 1)$, and $\mathbf{Y} = \text{Normal}(X\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$. The test dataset used to assess performance is simulated with the same settings as the training data, but $n = 400$. We apply our method for estimating the parameters in the clinical model (using 10,000 iterations of the Gibbs sampler with 500 for a burn-in period) and compare the results to that of least squares regression in Table 2.1.

We see that our method better estimates σ^2 (recall $\sigma^2 = 1$). We also note that the least squares regression yields coverage probabilities that are too high, while the frequentist coverage probabilities of our Bayesian credible intervals are close to the nominal levels. The MSE ratio is less than 1 for the training data but much greater than 1 for the test data; this is consistent with the idea that in this high dimensional

Table 2.1: Simulation results. The estimate of σ^2 is the posterior mean for our method. "CI" is Credible Interval for our method and Confidence Interval for least squares. MSE Ratio is the Mean Squared Error from least squares divided by the MSE from the respective method. Reprinted with permission from Jennings et al. (2012).

	Our Method	**Maximum Likelihood**
$\widehat{\sigma^2}$	0.9210287	0.1180878
95% CI Coverage	0.9667	1.00
90% CI Coverage	0.8778	0.9667
MSE Ratio (train data)	0.26536	1
MSE Ratio (test data)	9.538	1

setting with expected sparsity, least squares tends to overfit the training data, while the Bayesian method performs shrinkage that leads to improved estimation on the test data and is thus more applicable to the overall population. We also see excellent shrinkage properties of our method in Figure 2.2; most least squares coefficient estimates (which are the maximum likelihood estimates) are far from the true parameter values, while the posterior means from our method shrink these estimates close to the true values. The non-linear shrinkage and flexibility provided by the NG prior facilitate more shrinkage near 0 without severe attenuation of the estimates for truly large regression coefficients.

2.4 Integrative Analysis of GBM Data

GBM is one of the most common and most malignant brain tumors. Finding prognostic biomarkers related to cancer development is an important issue, and GBM was one of first cancers to be studied in TCGA. The data currently available contains information from multiple molecular platforms (genomic/epigenomic/transcriptomic) as well clinical data on several hundred tumor samples ($\sim$500). In our integrative analysis we use 233 matched tumor samples that have been assayed by expression,

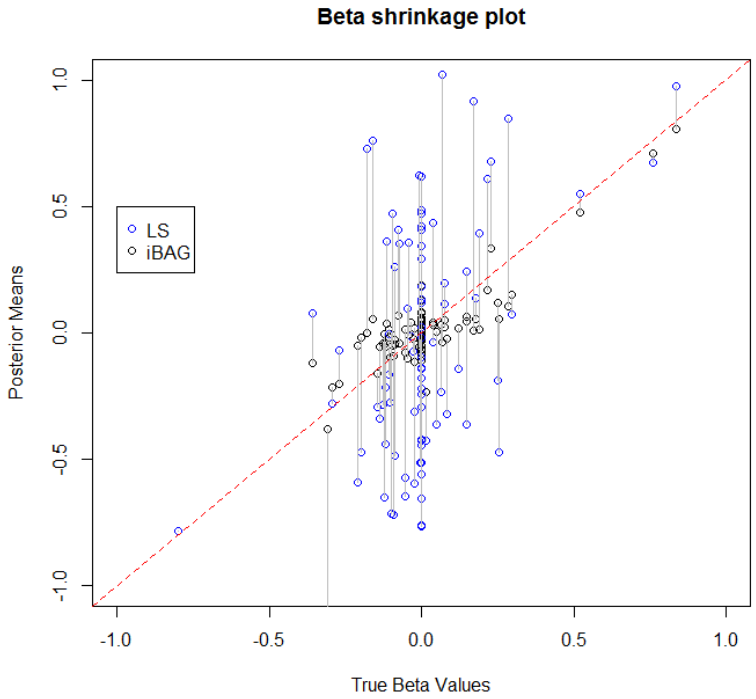


Figure 2.2: Least squares estimates and posterior means from our method are plotted against the true β values. The vertical lines show the shrinkage. Reprinted with permission from Jennings et al. (2012).

methylation and copy number platforms as described below.

2.4.1 Description of data

We focus our analysis on data corresponding to 49 genes implicated in important signaling pathways in GBM (RTK/PI3K, P53, and RB pathways (Memorial Sloan-Kettering Cancer Center, 2012)), using the following structure:

1. *OurSurvival* (233 × 1), containing days of survival after diagnosis for each patient.

2. *OurMRNA* (233×49), containing mRNA expression levels for each gene (columns) for each patient (rows).

3. *OurMeth* (233×176), containing data on the methylation markers (columns) for each patient (rows). There can be multiple (ranging from 1-21) methylation markers per gene, and the columns are ordered by gene.

4. *OurCopyNumber* (233×524), containing copy number data (columns) for each patient (rows). Again, there are multiple (ranging from 1-43) values per gene, and the columns are ordered by gene.

One gene has no methylation data, so we remove that column from the X matrix, which essentially sets that effect to be 0. Any effect that may be due to methylation for that gene would then be captured by the "other" predictor in the clinical model. After standardizing the predictors and imputing the (few) missing values, we model the data using an AFT model with log survival times as the outcome and apply our method of estimating the parameters of the iBAG model.

2.4.2 Results using iBAG model

After applying our method to the GBM data, we then use the method discussed in Section 2.2.3 to determine the significant markers using $\delta = 0.05$ (corresponding to a 5% change in survival time). Figures 2.3 and 2.4 show the posterior probabilities of the effect ($\beta_{i,j}$) being greater than δ_+^* and less than δ_-^*, respectively. We find 12 markers to be significant, 7 with positive effects on survival (more expression attributed to that platform, better prognosis) and 5 with negative effects (more expression attributed to that platform, poorer prognosis). The 7 positive markers include IRS1, RAF1, CCND1, MDM2, SRC, PDGFRB, and ERBB2. The genes IRS1 and RAF1 were determined to be related to clinical outcome through methylation effects, while expression of CCND1 and MDM2 were related to clinical outcome through copy number. For the other 3, gene expression was related to clinical outcome through some other unspecified mechanism. The 5 negative genes

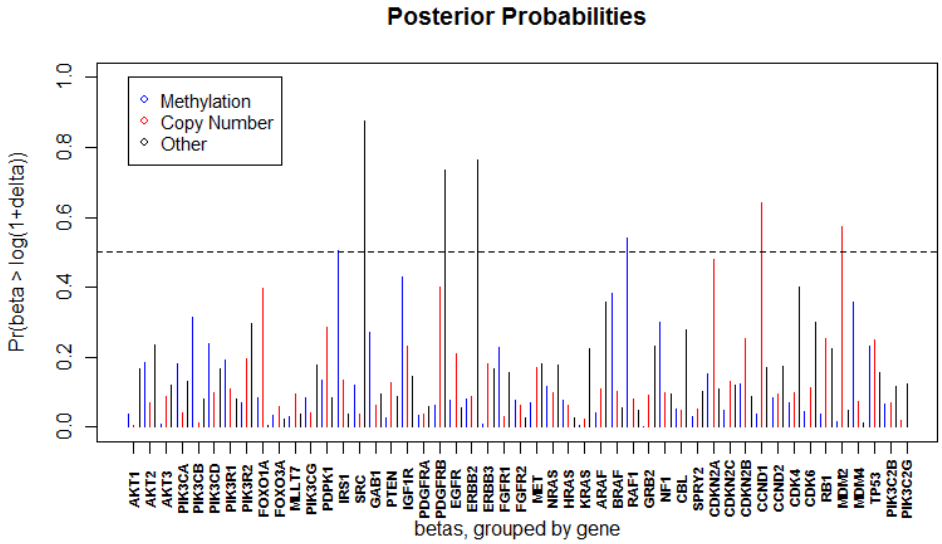


Figure 2.3: The posterior probability that $\beta_{i,j} > \delta_+^*$ is plotted. We consider the marker i, j to be significant if this probability is greater than 0.5. Reprinted with permission from Jennings et al. (2012).

were ERBB3, KRAS, GRB2, MDM2, and FOXO1A. The first four were related to clinical response through methylation, while the latter was through some mechanism other than methylation or copy number. Note that one gene (MDM2) is found to be significant on two different platforms. We have not only identified 11 genes as having a significant effect on survival, but we have also determined which platform(s) of those genes is (are) modulating the effect.

The principles behind our method suggest that we should have increased power due to incorporating additional information from the integrated platforms. To investigate this conjecture, we apply our estimation technique to the clinical model using only the 49 mRNA expression values (without methylation or copy number) as predictors. In this case only three genes are found to be significant; the expression of genes SRC, PDGFRB, and ERBB2 are found to have positive effects on patient

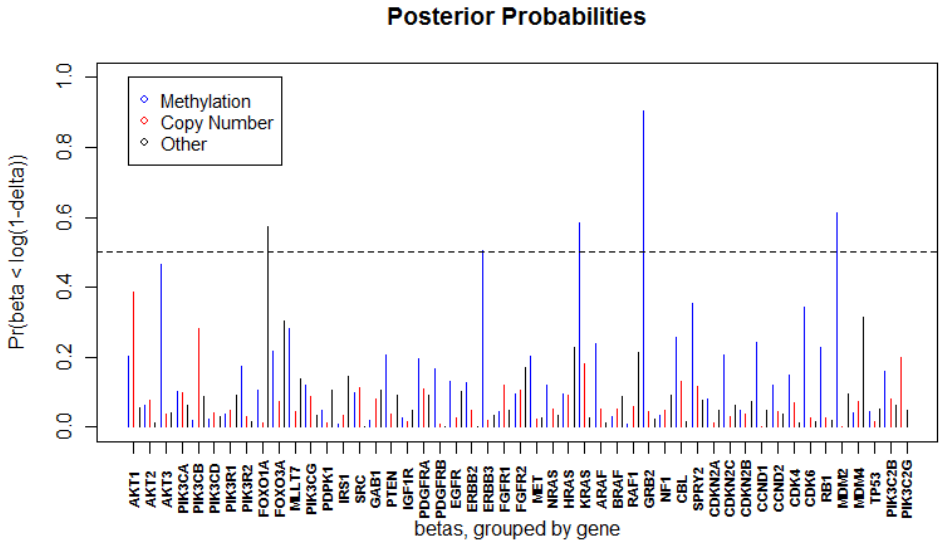


Figure 2.4: The posterior probability that $\beta_{i,j} < \delta^*_{-}$ is plotted. We consider the marker i, j to be significant if this probability is greater than 0.5. Reprinted with permission from Jennings et al. (2012).

survival. These three genes are a subset of those identified when we applied our complete model, which speaks to the consistency of our method and also suggests, qualitatively, the idea that our method may have increased power in identifying these important genes and their corresponding mechanisms. Of course, these genes and principles need to be further validated in future studies.

2.5 Conclusion

In this article, we present a hierarchical Bayesian model that integrates data from multiple genomic platforms, incorporating information about the platforms' biological relationships in order to better identify genes that are critical to patient survival and to additionally provide mechanistic information on the manner of their effect. In summary, the key advantages of our method include: (1) multiple platforms are integrated in a single model; (2) the biological relationships between platforms

are taken into account by the model; (3) high dimensional data can be handled easily, with shrinkage priors; (4) The NG prior on the predictors allows for flexible shrinkage of the parameter estimates; (5) the model can be extended to incorporate more platforms, as long as the underlying biological relationships are well-understood; (6) we see increased power in identifying biomarkers; and (7) we have the ability to not only identify genes significant to patient survival, but also gain mechanistic information on the manner by which the gene expression is related to outcome.

Applying our methodology to a GBM dataset from TCGA, our method identified several genes with effects that have a significant impact on survival time. In addition we identified whether each gene was related to clinical outcome through methylation, copy number, or some other mechanism. This is especially advantageous in investigating the biological mechanisms of cancer development and progression, and in subsequent development of novel therapeutic strategies.

Although beyond the scope of this paper, two areas of future investigation might include: (1) relaxing the parametric assumptions by using generalized additive models instead of linear models, or substituting specified parametric nonlinear models if they are justified by the science; and (2) dynamic modeling, which would require different types of data and further modeling assumptions to capture complex patterns of feedback loops both within and between platforms.

3. BAYESIAN METHODS FOR EXPRESSION-BASED INTEGRATION OF VARIOUS TYPES OF GENOMICS DATA*

3.1 Introduction

The central dogma of molecular biology summarizes the steps involved in the passage of genetic information at a molecular level: DNA is transcribed to messenger RNA (mRNA), which is then translated to a protein, which carries out a specific action in an organism. In addition, there are also other alterations and interferences, such as epigenetic factors, that can occur at the DNA and/or mRNA levels which affect the ultimate expression of a given gene. In this paper, we consider methylation (which occurs at the DNA level and typically results in a silencing of the gene), copy number (which describes an attribute at the DNA level that affects mRNA expression), and mRNA expression (which affects protein expression); these subsequently affect a clinical phenotype (e.g., survival) (see Figure 3.1). In addition, it is believed that the mechanism of cancer development is complex and involves multiple genes (Kanu et al., 2009). It is known that genes interact and are related through certain pathways, and in this paper, we focus on genes from important signaling pathways that influence cancer progression and development (Memorial Sloan-Kettering Cancer Center, 2012).

Current technologies allow us to obtain data from the above-mentioned platforms (and many others) for each gene involved in the investigations. The Cancer Genome Atlas (TCGA) is a project that began in 2006 to gather comprehensive genomic data using multiple platforms on over 20 types of cancer (The Cancer Genome Atlas,

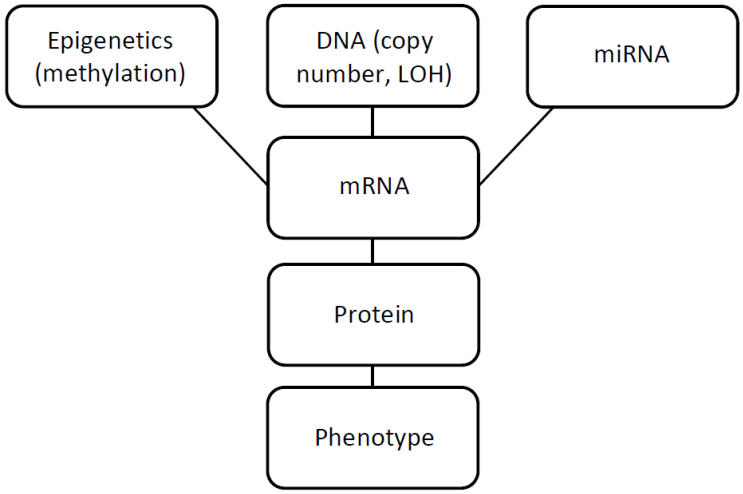


Figure 3.1: Platform relationships. Schematic representation of the multiple molecular platforms and their biological relationships. Reprinted with permission from Jennings et al. (2013).

2012). The increasing availability of such data has motivated the development of methods that seek to improve estimation and prediction regarding genomic effects on cancer outcomes by integrating data from multiple platforms in a single analysis. The incorporation of information from more than one platform has the potential to increase power and lower false discovery rates in identifying markers related to clinical outcomes for cancer patients (Wang et al., 2013); such improvements would deepen our understanding of how cancer develops and spreads, offering researchers valuable insight regarding the development of drugs and procedures intended to prevent or inhibit cancer development.

Some integration techniques consider different platforms sequentially and then draw conclusions from the combination of results. For example, the TCGA Research Network performed a large-scale study of ovarian cancer data, including specific platforms such as gene mutations, copy number, mRNA expression, miRNA expression, and DNA methylation. Within each platform, they compared normal and tumor

cells to identify significant genes and combined the information obtained from different platforms to understand the deeper biology behind the cancer mechanisms, including gene interactions. Using the prevalence of significant genes, they also identified influential pathways, including the RB1 and PI3K/RAS pathways (Bell et al., 2011). TCGA Research Network conducted a similar style study on Glioblastoma Multiforme (GBM) data and, among other things, discovered a previously unknown link between MGMT methylation and the mutation spectra of mismatch repair genes through the integration of mutation, methylation, and clinical treatment data (McLendon et al., 2008). These methods provide insight into the roles and interactions of genes as related to the development and outcome of the disease.

Another type of integrative method proposes incorporating multiple platforms in a single model. Such approaches must face the challenges of high dimensionality and complex biological relationships both within and between platforms. One such approach is iCluster, proposed by Shen et al. (2009), which is a joint latent variable model-based clustering method that integrates data from multiple genomic platforms to cluster samples into subtypes. iCluster achieves reduced dimension of the data, and it is shown to identify potentially novel subtypes of breast cancer and lung cancer (Shen et al., 2009). However, this method does not directly model the biological relationships among platforms; in addition, it is an unsupervised method, while our approach is supervised. Tyekucheva et al. (2011) suggest a method that includes multiple platforms as predictors in a logistic regression model (with phenotype as the response), and they show that incorporating multiple platforms yields more power to detect differentially expressed genes than approaches that only use a single platform (Tyekucheva et al., 2011). As with iCluster, this approach accounts for dependence between platforms, but it does not directly take into account their biological relationships.

Another method, proposed by Lanckriet et al. (2004), first represents data from each platform (such as primary protein sequence, protein-protein interaction, and mRNA expression) via a kernel function and then combines the kernels in a classification model (predicting, for example, protein type). It is shown that this method outperforms methods based on a single kernel from any one data platform (Lanckriet et al., 2004). However, this method does not directly model the relationships among the platforms, and kernel representations of the marker effects on the clinical outcomes are not directly interpretable. Liu et al. (2007) suggest another approach that integrates clinical covariates and multiple gene expressions (from a common pathway) to predict a continuous outcome through a semiparametric model; the covariates are modeled parametrically, and the pathway effect is modeled through least squares kernel machines (LSKM) (either parametrically or not). The covariate as well as pathway effects can be estimated, and the pathway effect can be tested for significance. The nonparametric LSKM regression allows for complicated interactions between genes (Liu et al., 2007), but this method only incorporates a single genomic platform (and accounts for its internal biological relationships). Recently, Wang et al. (2013) proposed an integrative Bayesian analysis of genomics data (iBAG) framework that models the biological relationships between two platforms. This approach involves a global gene search and uses variable selection via the Bayesian lasso-based shrinkage priors to deal with the high dimensionality of the data.

In this paper, we introduce a generalized version of iBAG that integrates data from an arbitrary (multiple) number of genomic platforms using a hierarchical model that incorporates the biological relationships among them. We focus our analysis on genes from several important cancer signaling pathways and integrate mRNA, methylation, and copy number data to predict survival in GBM patients. In addition, we reduce dimension by regressing the clinical outcome on latent scores of the platforms

(see Section 3.2.1 for details). To improve effect size estimation and to achieve sparsity, we use a Normal-Gamma (NG) prior for the effects, which increases flexibility in the estimation as compared to the Laplace prior of the Bayesian lasso (Griffin and Brown, 2010) (see Section 3.2.2 for further discussion). Section 3.3 illustrates our methodology on a synthetic example; analysis of GBM data is presented in Section 3.4; and conclusions are drawn in Section 3.5.

3.2 A Multivariate iBAG Model

Our construction of a multivariate iBAG model employs a two-component hierarchical model where the first component can be considered as the *mechanistic model* and the second can be considered as the *clinical model*. In the first stage mechanistic model, we partition each gene's expression into the factors explained by methylation, copy number, and other (unknown/unmeasured) causes using a principal component-based regression model. Subsequently, we include these factors as predictors in the second stage clinical model, thus finding not only those genes whose expression is directly related to clinical outcome, but also expression effects driven by methylation, copy number, or other mechanisms. We explain the construction of each of these components below.

3.2.1 Mechanistic model

Let n = number of patients, J = number of platforms being integrated, and p_j = number of genes from platform j. The mechanistic model for each gene can be expressed as:

$$\text{mRNA}_i \;=\; \text{M}_i + \text{CN}_i + \text{O}_i,$$

where each of the terms are defined as follows:

- mRNA_i is the level of gene expression for gene i (where $i = 1, ..., \max(p_j); j = 1, ..., J$) and is of dimension $(n \times 1)$.

- M_i is the part of gene$_i$ expression that is attributed to methylation, and is of dimension $(n \times 1)$. Specifically, M_i is the product of some methylation predictor and a fitted coefficient. Details are below.

- CN_i is the part of gene$_i$ expression that is attributed to changes in copy number, and is of dimension $(n \times 1)$. Specific calculation is similar to M_i – see below.

- O_i represents the 'other' (remaining) part of the gene expression that is explained by something other than methylation or copy number, and is of dimension $(n \times 1)$.

Since the raw methylation and copy number data for any given gene can contain multiple (up to 40 in our data) values from different markers within that gene, to estimate each of the components M_i, CN_i, and O_i, we first carry out two principal component analyses (PCA) for gene$_i$: one each for the methylation and copy number data, and in each case, we keep the number of principal components that retain $\geq 90\%$ of the total variation. We then regress mRNA_i on the methylation and copy number PC scores. We use the estimated pieces and the corresponding residuals from this regression to estimate the vectors $\mathrm{M}_i = \sum_{k=1}^{K} X_{i,k}^{M} B_k^{M}$ (where $X_{i,k}^{M}$ is the methylation value for gene i with $K = 1$ if there is only one methylation marker for that gene, or the methylation score for principal component k for gene i if there are multiple methylation markers for gene i, and B_k^{M} is the vector of regression coefficients), $\mathrm{CN}_i = \sum_{r=1}^{R} X_{i,r}^{CN} B_r^{CN}$ (where $X_{i,r}^{CN}$ is the copy number value for gene i with $R = 1$ if there is only one copy number marker for that gene, or the copy number score for principal component r for gene i if there are multiple copy number markers for gene i, and B_r^{CN} is the vector of regression coefficients), and O_i = residuals. This process is repeated for each gene independently.

3.2.2 Clinical model

The clinical model component of our construction relates the effect of the mechanistic parts of the genes (as estimated above) to a clinical outcome of interest (e.g., survival, in our context) and can be written as:

$$Y = \mathrm{M}\boldsymbol{\beta}_1 + \mathrm{CN}\boldsymbol{\beta}_2 + \mathrm{O}\boldsymbol{\beta}_3 + \boldsymbol{\epsilon},$$

where Y denotes the clinical outcome, $\boldsymbol{\beta}_j$ are the effects of platform j on Y, and $\boldsymbol{\epsilon}$ is the error term. The covariates in the model {M, CN, O} are the vectorized gene expression effects attributed to methylation, copy number, and other sources, respectively, and are estimated from the mechanistic model. In essence, our clinical component jointly (additively) models the effects of all the gene expressions and their components - derived from different sources (methylation/copy number) - in a unified manner. When the clinical response is survival, we use an accelerated failure time (AFT) model, taking Y to be log(survival) (Wei, 1992).

Our goal is to find a list of significant genes that affect the outcome via the various mechanisms; hence, efficient estimation of $\boldsymbol{\beta} = \{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3\}$ is of primary interest. One route would be to simply fit a least squares regression to estimate the parameters. However, the number of predictors is large compared to the number of samples, and, more importantly, we expect our solution to be very sparse since only a few genes will be related to clinical response; hence, least squares would overfit the data and yield less accurate results as compared to approaches that induce sparsity by shrinkage/penalization. We illustrate this fact in our simulation in Section 3.3.

To induce shrinkage/penalization, we follow a Bayesian approach and specify particular prior distributions for each model parameter in the clinical model and sample from the posterior distribution using Markov Chain Monte Carlo (MCMC). There are several priors known to achieve sparsity and facilitate Bayesian variable selection,

which we will discuss briefly. One option is to simply put vague Normal$(0, \infty)$ priors on each regression coefficient. This is equivalent to doing least squares regression and is impossible in cases where there are more variables than data points, because singular solutions arise. A natural extension is to place proper mean-zero Normal priors on the coefficients, which is equivalent to ridge regression. Although accommodating more predictors than data points and facilitating shrinkage, the type of shrinkage is linear which is not desirable in the current settings. This linear shrinkage leads to more shrinkage and thus greater bias for larger coefficients, while in this setting, we desire the opposite: less shrinkage for large (significant) coefficients and greater shrinkage for smaller (non-significant) ones. This type of non-linear shrinkage can be accomplished by various priors. One is the 'spike and slab' prior consisting of a mixture of a point mass at zero (the spike) and a Normal (the slab). Although this can accommodate a large number of predictors and avoids linear shrinkage, the shrinkage asymptotes to a constant which still results in attenuation of the truly large effects, something we want to avoid. In addition, computational complications and difficulties accompany the use of spike and slab priors. As we show below, all but one of our complete conditional distributions are in closed form, so we can avoid the computational difficulties associated with the spike and slab method, as well as the attenuation of large effects, by utilizing continuous shrinkage priors.

A widely known method that places a continuous sparsity prior on the regression coefficients is the Bayesian lasso (Park and Casella, 2008), which is incorporated by assigning a double exponential (i.e., Laplace) prior to $\boldsymbol{\beta}$. When posterior modes are used as the coefficient estimates, this process yields the same solutions as Tibshirani's lasso (Tibshirani, 1996). The Bayesian lasso has proven to perform well in conducting adaptive shrinkage-induced sparsity, but the single hyperparameter formulation does not allow for enough flexibility to estimate the true size of potentially

large, non-zero effects. Instead, these effect estimates are shrunk toward zero along with the smaller effects (Griffin and Brown, 2010). An alternate class of priors we use and discuss is the Normal-Gamma (NG) prior distribution for $\boldsymbol{\beta}$. Incorporating this continuous prior not only provides shrinkage of the coefficients but the extra hyperparameter in the NG prior construction facilitates more adaptability in the estimated shrinkage relative to the Bayesian lasso (Park and Casella, 2008) - with the NG, the larger effects are shrunk less than the smaller effects (Griffin and Brown, 2012), thus leading to improved estimation (Griffin and Brown, 2010). In summary, the NG prior is extremely advantageous in our situation, since it delivers the sparsity we need, while leaving larger effects mostly unshrunk, thus aiding our estimation of the important effects.

For our method, we assign a Normal-Gamma (NG) prior distribution for each $\boldsymbol{\beta}_j$. Our complete hierarchical clinical model can be written as:

$$\begin{aligned}
\mathbf{Y} &= \text{Normal}(X\boldsymbol{\beta}, \sigma^2\mathbf{I}_n), \\
\boldsymbol{\beta} &= \text{Normal}(\mathbf{0}_{\tilde{p}}, D_\psi) \text{where } D_\psi = \text{diag}(\psi_{1,1}, ..., \psi_{1,p_1}, ..., \psi_{J,1}, ...\psi_{J,p_J}), \\
\psi_{j,i} &= \text{Gamma}(\lambda_j, 1/(2\gamma_j^2)), \\
\sigma^2 &= \text{InverseGamma}(a, b), \\
\lambda_j &= \text{Exponential}(c), \\
\gamma_j^{-2} &= \text{Gamma}(\tilde{a}, \tilde{b}/(2\lambda_j)),
\end{aligned}$$

where $\tilde{p} = \sum_{j=1}^{J} p_j$ is the total number of predictors in the model. (Note that the double exponential prior of the Bayesian lasso would be constructed by assigning $\beta_{j,i}|\psi_{j,i} \sim \text{Normal}(0, \psi_{j,i})$ and $\psi_{j,i} \sim \text{Exponential}(\lambda_j)$. The single parameter in the exponential prior (λ_j) is the reason such a construction has limited flexibility as compared to the NG prior which is parameterized by both λ_j and γ_j.) With the NG formulation as given above, the complete conditionals for most parameters are

available in closed form - we can use Gibbs sampling to update all parameters except λ_j, which we update using a Metropolis-Hastings random walk step. More details for drawing MCMC samples are available in Appendix A.

3.2.3 Gene selection

Given the posterior samples from the MCMC, we determine which genes are significantly related to clinical outcome using a method based on the median probability model (Barbieri and Berger, 2004). First, we define a minimum effect size which is driven by practical considerations. Since we are analyzing survival data, we use AFT models using log(survival) as the response; thus, a δ-fold or larger change in survival for a unit increase in a predictor corresponds to a $\beta_{j,i}$ outside the region (log(1-δ), log(1+δ)), where $\beta_{j,i}$ is the regression coefficient for platform j of gene i. Denote this region (δ^*_-, δ^*_+). (In our following analyses, we use $\delta = 0.05$ which corresponds to a 5% change in survival time.) If S is the number of MCMC samples and $\beta^{(s)}_{j,i}$ is the $\beta_{j,i}$ sample from iteration s, then $p_+(x_{j,i}) = \sum_{s=1}^{S} \mathbf{I}(\beta^{(s)}_{j,i} > \delta^*_+)/S$ is the posterior probability that $\beta_{j,i}$ is higher than the practical cutoff δ^*_+. Similarly, $p_-(x_{j,i}) = \sum_{s=1}^{S} \mathbf{I}(\beta^{(s)}_{j,i} < \delta^*_-)/S$ is the posterior probability that $\beta_{j,i}$ is lower than the practical cutoff δ^*_-. We flag a gene as 'significant' if $p_+(x_{j,i}) > 0.5$ or if $p_-(x_{j,i}) > 0.5$.

Algorithm 1 provides a concise summary of implementing the multivariate iBAG model and conducting gene selection.

Algorithm 1: Method implementation.

Input: Raw data matrices, one for outcome (survival) and one for each platform (mRNA, methylation, copy number) (Rows are patients, and columns are markers arranged by gene.), number of patients n, number of platforms J, number of genes in platform j p_j, number of MCMC samples S, number of MCMC samples to use as burn-in B, and practical effect size δ.

Output: Prognostic markers with high posterior probability of having prespecified practical effect size.

Prepare data:

- Impute missing data (see Appendix A).
- For methylation and copy number platforms:
  - For each gene i:
    - Perform principal component analysis (PCA) on platform j. Keep the number of components that account for $\geq 90\%$ of the variation.
    - Get PC scores associated with retained components. Call matrix of scores M^* for methylation and CN^* for copy number, where the number of columns is the number of score vectors.
- Repeat for any other platforms available upstream of mRNA.

Fit mechanistic model:

- For each gene i:
  - Use least squares to regress response platform (mRNA) on M^* and CN^*. (Note that the modeled relationship should reflect the biological relationships between platforms.)
  - Let M be the linear combination of predicted coefficients and M^*, CN be the linear combination of predicted coefficients and CN^*, and O be the residuals.

Standardize M_i's, CN_i's, and O_i's. There should be $\sum_{j=1}^{J} p_j$ of these predictors. Log-transform survival responses and mean-center.

Fit clinical model:

- Draw S MCMC samples from the complete conditionals (see Appendix A), using the first B samples as burn-in, to fit the AFT model and obtain $S - B$ posterior samples of regression coefficients $\beta_{j,i}$.

Marker selection:

- Given practical threshold δ, compute $\delta_-^* = \log(1-\delta)$ and $\delta_+^* = \log(1+\delta)$.
- For each marker:
    - Calculate $\Pr(\beta_{j,i} > \delta_+^*)$ and $\Pr(\beta_{j,i} < \delta_-^*)$ using posterior samples.
    - Flag marker if either calculated probability is greater than 0.5.

Return: Identified markers.

3.3 Simulation

We investigate the shrinkage properties of our Bayesian penalized regression formulation of the clinical model as compared to least squares regression, Bayesan lasso, frequentist lasso, and frequentist elastic net through a simulation. We simulate a training dataset with 90 predictors ($J = 3$ platforms with $p_1 = p_2 = p_3 = 30$ predictors from each), where 30 randomly selected $\beta_{j,i}$'s are set exactly to 0 and the other 60 are sampled from a Laplace($\mu = 0$, $b = 1/7$) distribution; this reflects the effective sparsity we expect to see in our data. The other settings for the simulated data are $n = 100$, $\sigma^2 = 1$, each X entry is from Normal$(0, 1)$, and $\mathbf{Y} = \text{Normal}(X\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$. The test dataset used to assess performance is simulated with the same settings as the training data, but $n = 400$. We applied our method for estimating the parameters in the clinical model, using 10,000 iterations of the Gibbs sampler with 500 for a burn-in period. For both the frequentist lasso and elastic net, we ran the simulation with two standard choices for the penalty parameter λ: (1) '1 SE' where we used the largest λ with cross validation error within one standard error of the minimum cross validation error and (2) 'min' where we used the λ with minimum error (from cross validation). For elastic net, we set the mixing parameter (that controls the mixture of penalties) to 0.5. The results of our method are compared to those of the other methods in Table 3.1.

Table 3.1: Simulation results. Freq. EN means freqentist elastic net, which was run with mixing parameter (for penalty mixture) 0.5. The estimate of σ^2 is the posterior mean for our method and the Bayesian lasso. For the others, it is the mean sum of squared error. 'CI' is credible interval for Bayesian methods and confidence interval for frequentist methods. Note that for the frequentist lasso and elastic net, it is not possible to obtain standard errors for the coefficients set to 0, and therefore, we cannot construct the CI's. The penalty choice of '1 SE' means we used the largest parameter with error within one standard error of the minimum error, while 'min' means we used the parameter with minimum error (from cross validation). MSE ratio is the mean squared error from least squares divided by the MSE from the respective method. NA indicates not applicable. Reprinted with permission from Jennings et al. (2013).

	$\widehat{\sigma^2}$	**95% CI coverage**	**90% CI coverage**	**MSE ratio (train data)**	**MSE ratio (test data)**
Our method	0.9073	0.9778	0.8889	0.2827	9.4630
Maximum likelihood	0.1181	1.00	0.9667	1	1
Bayesian lasso	0.6407	0.9667	0.9111	0.3727	8.858
Freq. lasso (1 SE)	1.2020	NA	NA	0.0983	8.1163
Freq. lasso (min)	0.6379	NA	NA	0.1851	8.8374
Freq. EN (1 SE)	0.9278	NA	NA	0.1273	8.4439
Freq. EN (min)	0.7012	NA	NA	0.1684	8.7154

We see that our method gives a good estimate of σ^2 (recall $\sigma^2 = 1$). We also note that the least squares regression yields coverage probabilities that are too high, while the frequentist coverage probabilities of the Bayesian credible intervals are close to the nominal levels. (Note that for the frequentist lasso and elastic net, it is not possible to obtain standard errors for the coefficients set to 0, and therefore, we cannot construct the CI's.) For all methods (other than least squares), the MSE ratio is less than 1 for the training data but much greater than 1 for the test data; this is consistent with the idea that in this high dimensional setting with expected sparsity, least squares tends to overfit the training data, while methods that perform shrinkage lead to improved estimation on the test data and thus yield results more applicable to the overall population. Considering that the MSE ratio is the mean

squared error from least squares divided by the MSE from the respective method, we see that our method has the best (largest) MSE ratio on test data, which for our purposes is the most relevant comparison criterion.

We also see excellent shrinkage properties of our method in Figure 3.2; most least squares coefficient estimates (which are the maximum likelihood estimates) are far from the true parameter values, while the posterior means from our method shrink these estimates closer to the true values. The non-linear shrinkage and flexibility provided by the NG prior facilitate more shrinkage near 0 without severe attenuation of the estimates for truly large regression coefficients.

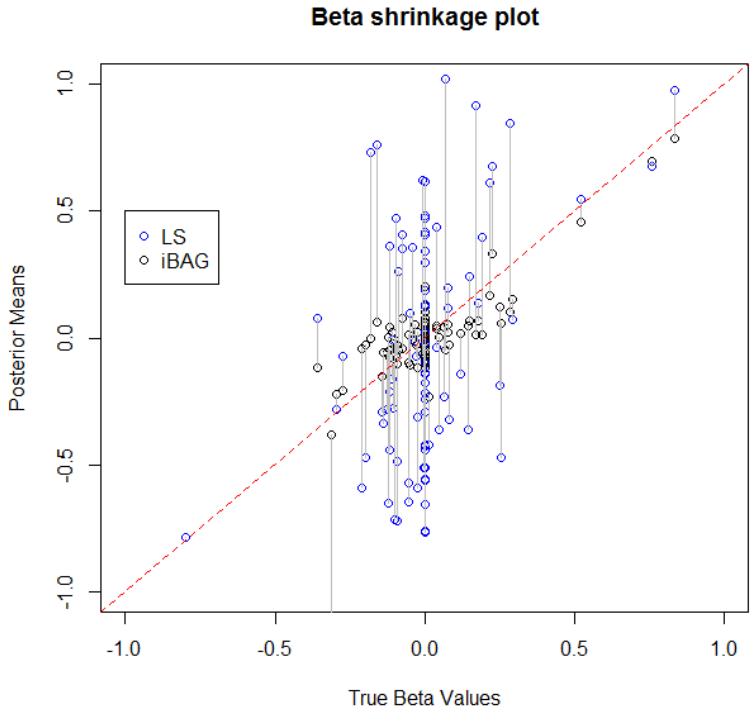


Figure 3.2: Simulation results. Least squares estimates and posterior means from our method are plotted against the true β values. The vertical lines denote the difference between the estimates from each method thus indicating the shrinkage properties of the NG prior. Reprinted with permission from Jennings et al. (2013).

3.4 Integrative Analysis of GBM Data

GBM is one of the most common and most malignant brain tumors. The American Cancer Society estimates that in the year 2013, there will be 23,130 new cases of brain and other nervous system cancers in the USA and that 14,080 Americans will die from such cancers (American Cancer Society, 2013). GBM tumors make up 17% of all primary brain tumors (American Brain Tumor Association, 2013), and prognosis is typically very poor; a study with 7,259 patients, each diagnosed with GBM from 2005 to 2008, found a median survival time of 14.6 months for patients who received tumor-directed surgery and radiation therapy and a median survival time of 2.9 months for patients who did not receive any radiation treatment (Johnson and O'Neill, 2012). Treatment options include surgery, radiation, and/or chemotherapy, but even for a patient receiving more than one of these treatments, the outlook is dismal at best. Finding prognostic biomarkers related to cancer development and patient survival is an important issue, and GBM was one of first cancers to be studied in TCGA. The data currently available contains information from multiple molecular platforms (genomic/epigenomic/transcriptomic) as well as clinical data on several hundred tumor samples (approximately 500).

The availability of such extensive genomic data has prompted several studies using the TCGA GBM data, and fortunately, there continue to be discoveries of biomarkers that aid in predicting survival and identifying subtypes of GBM. One such study conducted by Verhaak et al. (2010) combined gene expression data from multiple types of microarray assays to classify tumors into four distinct subtypes (each responding differently to therapy) and to discover which gene expression levels had a significant impact on the classification. Other platforms were also used, such as copy number and mutations, in separate analyses to test for associations with

subtype (Verhaak et al., 2010). Another study by Noushmehr et al. (2010) used the available GBM DNA methylation data to identify a subgroup of GBM tumors associated with a significantly longer survival time. In our integrative analysis, we use 163 matched tumor samples that have been assayed by expression, methylation, and copy number platforms as described below. Each of these samples has an uncensored survival time (in days), and our aim is to identify prognostic biomarkers.

3.4.1 Description of data

Our copy number data is level 2 data from the HG_CGH_244A platform; it is the normalized signal for copy number alterations of aggregated regions per probe. Our methylation data is level 3 data from the HumanMethylation27K arrays; it is the methylated sites along a gene (probe level data). Our expression data is level 3 data (summarized per gene) from the Affymetrix profiled HT_HG_U133A platform (The Cancer Genome Atlas Data Portal, 2013).

We focus our analysis on data corresponding to 49 genes implicated in important signaling pathways in GBM (RTK/PI3K, P53, and RB pathways (Memorial Sloan-Kettering Cancer Center, 2012)), using the following structure:

1. *OurSurvival* (163×1), containing days of survival after diagnosis for each patient.

2. *OurMRNA* (163×49), containing mRNA expression levels for each gene (columns) for each patient (rows).

3. *OurMeth* (163×176), containing data on the methylation markers (columns) for each patient (rows). There can be multiple (ranging from 1 to 21) methylation markers per gene, and the columns are ordered by gene.

4. *OurCopyNumber* (163×524), containing copy number data (columns) for each

patient (rows). Again, there are multiple (ranging from 1 to 43) values per gene, and the columns are ordered by gene.

One gene has no methylation data, so we remove that column from the X matrix, which essentially sets M_i to be 0 for that gene. Any effect that may be due to methylation for that gene would then be captured by the 'other' predictor in the clinical model. After standardizing the predictors and imputing the (few) missing values, we model the data using an AFT model with log survival times as the outcome and apply our method of estimating the parameters of the iBAG model.

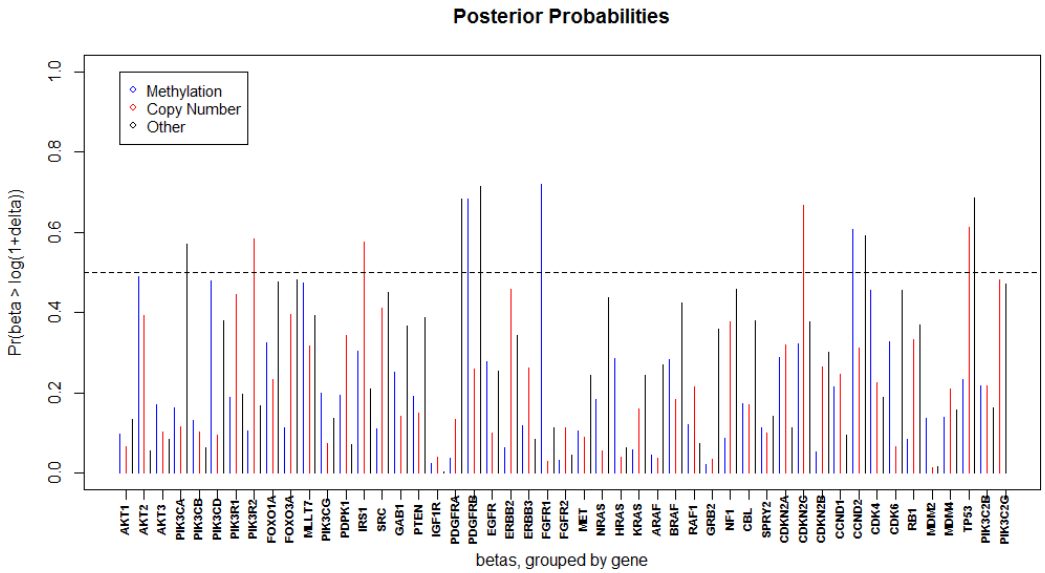


Figure 3.3: GBM data results. The posterior probabilities (based on MCMC samples) that $\beta_{j,i} > \delta_+^*$ is plotted, where $\beta_{j,i}$ is the clinical model regression coefficient for the marker associated with platform j of gene i, and $\delta_+^* = \log(1+\delta)$ is the transformed upper practical cutoff. For our analysis, we use $\delta = 0.05$, which corresponds to a 5% change in survival time, so the posterior probability shown here indicates the probability that a one unit increase in the marker results in at least a 5% increase in survival time. We consider the marker j, i to be significant if this probability is greater than 0.5. Reprinted with permission from Jennings et al. (2013).

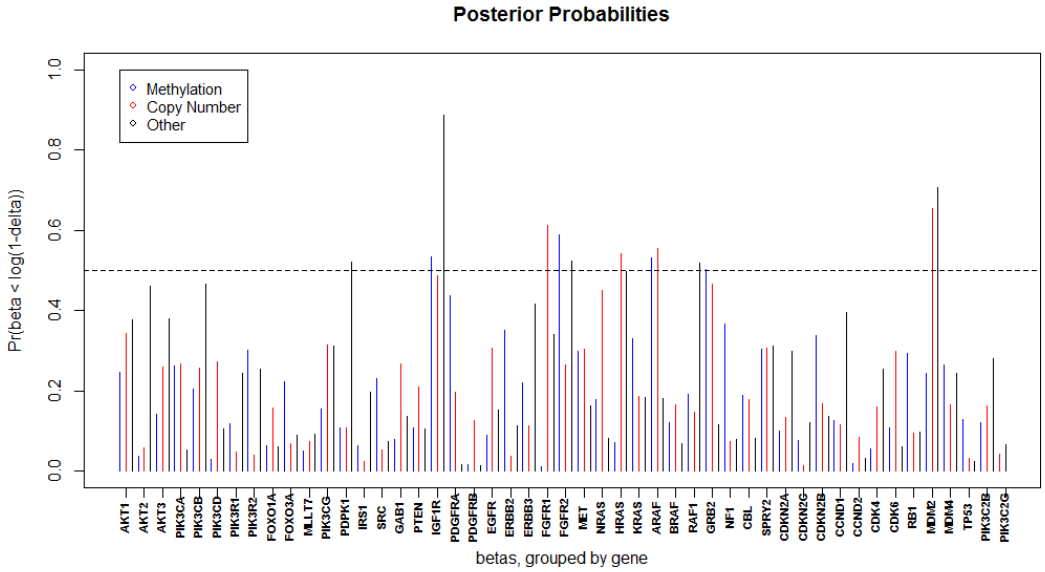


Figure 3.4: GBM data results. The posterior probabilities (based on MCMC samples) that $\beta_{j,i} < \delta^*_-$ is plotted, where $\beta_{j,i}$ is the clinical model regression coefficient for the marker associated with platform j of gene i, and $\delta^*_- = \log(1-\delta)$ is the transformed lower practical cutoff. For our analysis, we use $\delta = 0.05$, which corresponds to a 5% change in survival time, so the posterior probability shown here indicates the probability that a one unit increase in the marker results in at least a 5% decrease in survival time. We consider the marker j, i to be significant if this probability is greater than 0.5. Reprinted with permission from Jennings et al. (2013).

3.4.2 Results using iBAG model

After applying our method to the GBM data, we then use the method discussed in Section 3.2.3 to determine the significant markers using $\delta = 0.05$ (corresponding to a 5% change in survival time). Figures 3.3 and 3.4 show the posterior probabilities of the effect ($\beta_{j,i}$) being greater than δ^*_+ and less than δ^*_-, respectively. Figure 3.5 depicts the posterior means of the $\beta_{j,i}$'s and also indicates which were flagged as significant. We find 25 markers to be significant, 12 with positive effects on survival (more expression attributed to that platform, better prognosis) and 13 with negative effects (more expression attributed to that platform, poorer prognosis). The genes

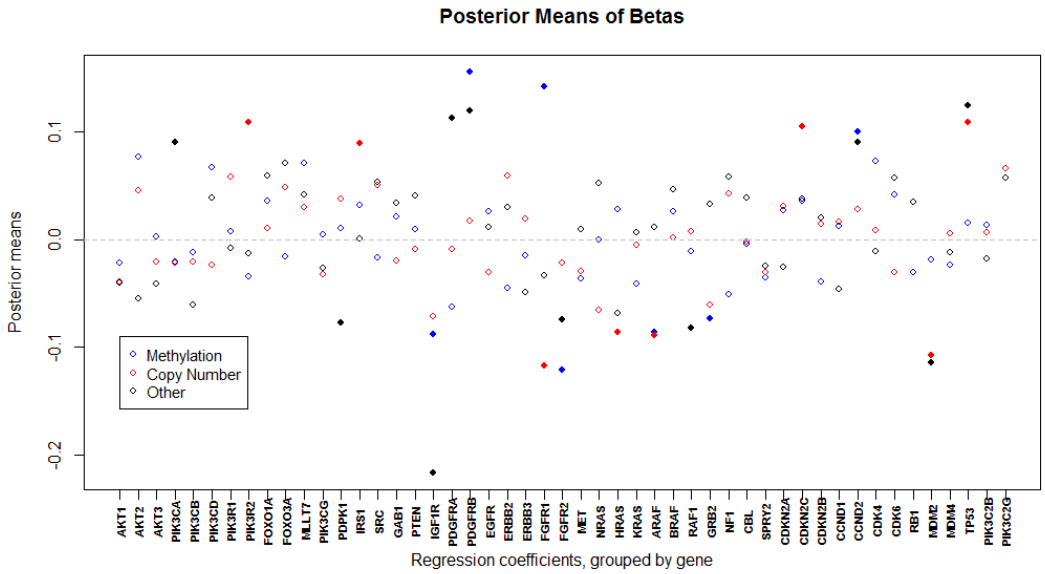


Figure 3.5: Regression coefficient posterior means. The estimates of the regression coefficients in the clinical model ($\beta_{j,i}$'s) are shown, where $\beta_{j,i}$ is the coefficient for the marker associated with platform j of gene i; the estimates are computed as the posterior means from our MCMC samples. The multiple platforms for each gene are labeled by color, and solid plot markers indicate that the effect was found to be significant, meaning that the posterior probability that a one unit increase in the marker results in at least a 5% change in survival time is at least 0.5. Reprinted with permission from Jennings et al. (2013).

with the 12 positive markers were PDGFRB, FGFR1, CCND2, PIK3R2, IRS1, CDKN2C, TP53, PIK3CA, and PDGFRA. The genes PDGFRB, FGFR1, and CCND2 were determined to be related to clinical outcome through methylation effects, while expressions of PIK3R2, IRS1, CDKN2C, and TP53 were related to clinical outcome through copy number. For PIK3CA, PDGFRA, PDGFRB, CCND2, and TP53, gene expression was related to clinical outcome through some other unspecified mechanism. The genes with the 13 negative markers were IGF1R, FGFR2, ARAF, GRB2, FGFR1, HRAS, MDM2, PDPK1, and RAF1. The first four were related to clinical response through methylation, while FGFR1, HRAS, ARAF, and

MDM2 were related through copy number, and PDPK1, IGF1R, FGFR2, RAF1, and MDM2 were related through some mechanism other than methylation or copy number. Note that eight genes (IGF1R, PDGFRB, FGFR1, FGFR2, ARAF, CCND2, MDM2, and TP53) are found to be significant on two or more different platforms. We have not only identified 17 genes as having a significant effect on survival (Table 3.2), but we have also determined which platform(s) of those genes is (are) modulating the effect.

Table 3.2: Gene results. All 49 genes appearing in the data are listed. Italic genes were identified by our method to have at least one significant marker. Reprinted with permission from Jennings et al. (2013).

Gene names				
AKT1	MLLT7	EGFR	BRAF	*CCND2*
AKT2	PIK3CG	ERBB2	*RAF1*	CDK4
AKT3	*PDPK1*	ERBB3	*GRB2*	CDK6
PIK3CA	*IRS1*	*FGFR1*	NF1	RB1
PIK3CB	SRC	*FGFR2*	CBL	*MDM2*
PIK3CD	GAB1	MET	SPRY2	MDM4
PIK3R1	PTEN	NRAS	CDKN2A	*TP53*
PIK3R2	*IGF1R*	*HRAS*	*CDKN2C*	PIK3C2B
FOXO1A	*PDGFRA*	KRAS	CDKN2B	PIK3C2G
FOXO3A	*PDGFRB*	*ARAF*	CCND1	

3.4.3 Biological interpretation

There are a total of 17 genes found to affect the expression of glioblastoma tumors significantly. Of these, nine genes are negatively affecting the survival and nine genes are affecting the survival positively. The positive and negative prognostic markers are reviewed within the context of glioblastoma biology in this section.

Negative prognostic markers: Fibroblast growth factor pathway signaling is associated with significant tumor enhancement in glioblastoma (Loilome et al., 2009). Fibroblast growth factor receptors FGFR1 and FGFR2 play an oncogenic role in various tumor types and can be targeted by multiple small molecules in cancer therapy (Katoh and Nakagama, 2013). FGFR1 expression can be regulated by methylation level of the upstream CpG island (Goldstein et al., 2007). Hyper-methylation of FGFR1 would provide positive effects by reducing the expression level of FGFR1 and thus appear to be affecting the survival in both ways. Insulin-like growth factor receptor 1 (IGF1R) is a well-known target to treat GBM and has been found to be associated with astrocytoma and meningioma as well (Carapancea et al., 2009). It is also associated with anti-EGFR resistance in GBM and is a pan-cancer biomarker connected with many different tumor types (Chakravarti et al., 2002; Hewish et al., 2009). MDM2 is a well-known oncogene and inhibitor of the tumor suppressor TP53. Previous studies in glioblastoma using expression and copy number platforms indicated the abnormal over-expression and amplification of MDM2 (Ruano et al., 2006; Yin et al., 2009). ARAF is a serine/threonine protein kinase of RAF family, known to stabilize the hetero-dimerization of RAF proteins, BRAF and CRAF (Rebocho and Marais, 2013). Its role and over-expression are observed in other tumors but are not explored in the context of glioblastoma (Craig et al., 2013). Growth factor receptor-bound protein 2 (GRB2) is involved in RAS signaling pathway and known to be associated with EGFR (Lowenstein et al., 1992). GRB2 is an interacting partner of EGFRvIII, a common mutated variant of EGFR in the molecular signaling of EGFR-driven glioblastoma (Prigent et al., 1996; Kapoor and O'Rourke, 2010).

Positive prognostic markers: The tumor suppressor gene TP53 is a positive prognostic marker as expected. The Cyclin-dependent kinase inhibitor CDKN2C, a known tumor suppressor of glioblastoma, is also identified as a positive marker

(Solomon et al., 2008). Platelet-derived growth factors (PDGF) receptors PDGFRA and PDGFRB show positive survival effects, whose oncogenic role is well established in the context of glioma (Suzuki et al., 2010; Nazarenko et al., 2012). These PDGF receptors are the representative genes of the pro-neural subtype of glioblastoma (Verhaak et al., 2010; Jiang et al., 2011). Interestingly, the pro-neural subtype of glioblastoma is enriched in oligodendroglioma and has higher survival rates compared to other subtypes of glioblastoma (Cooper et al., 2010). The insulin receptor substrate gene IRS1 is shown to be one of the representative candidates for mesenchymal subtype of GBM with poor survival (Brennan et al., 2009). The role of IRS1 is not clear, given that we found it to be a positive marker in our analysis. Overall, the positive markers are generally enriched in the pro-neural subtype of glioblastoma, which was found to have prolonged survival (Verhaak et al., 2010).

3.5 Conclusions

In this article, we present a hierarchical Bayesian model that integrates data from multiple genomic platforms, incorporating information about the platforms' biological relationships in order to better identify genes that are critical to patient survival and to additionally provide mechanistic information on the manner of their effect. In summary, the key advantages of our method include (1) multiple platforms are integrated in a single model; (2) the biological relationships between platforms are taken into account by the model; (3) high dimensional data can be handled easily, with shrinkage priors; (4) the NG prior on the predictors allows for flexible shrinkage of the parameter estimates; (5) the model can be extended to incorporate more platforms, as long as the underlying biological relationships are well understood; and (6) we have the ability to not only identify genes significant to patient survival but also gain mechanistic information on the manner by which the gene expression is

related to outcome.

Applying our methodology to a GBM dataset from TCGA, our method identified several genes with effects that have a significant impact on survival time. In addition, we identified whether each gene was related to clinical outcome through methylation, copy number, or some other mechanism. This is especially advantageous in investigating the biological mechanisms of cancer development and progression, and in subsequent development of novel therapeutic strategies.

Although beyond the scope of this paper, two areas of future investigation might include (1) relaxing the parametric assumptions by using generalized additive models instead of linear models or substituting specified parametric non-linear models if they are justified by the science, and (2) dynamic modeling, which would require different types of data and further modeling assumptions to capture complex patterns of feedback loops both within and between platforms.

4. BAYESIAN MODELS FOR FLEXIBLE INTEGRATIVE ANALYSIS OF MULTIPLATFORM GENOMICS DATA*

4.1 Introduction

Traditional cancer treatments include surgery, chemotherapy, and radiation. Although these treatments can be lifesaving, after a tumor is removed it is well known that the cancer can relapse (Ikeda et al., 1993; Martini et al., 1995; Khuri et al., 2001), and both chemotherapy and radiation treatments have terrible, sometimes permanent, side effects, including nausea and vomiting, hair loss, mouth sores, nerve damage, peeling skin, tinnitus, infertility, organ damage, and secondary cancer (Cancer.net, 2012; Michaelson and Oh, 2013). The allure of a treatment with lower chances of relapse (thus increasing patient survival) and with reduced side effects (thus improving patient quality of life) has motivated researchers to investigate the development of therapeutic strategies that target the specific genetic causes of a particular type of cancer. In particular, the availability of such therapies facilitates the practice of personalized medicine; a patent's genetic profile can be assayed, and the patient's treatment and dosages can be chosen to address the genetic abnormalities specific to the observed profile. This approach offers the potential to increase treatment efficacy and decrease incidence of negative side effects on the level of the individual patient.

To develop such a targeted treatment, it is first necessary to understand the mechanics of cancer development and progression on a molecular level. In general, within a cell, DNA is transcribed to messenger RNA (mRNA); mRNA is then trans-

*Reprinted, with permission, from "Bayesian models for flexible integrative analysis of multiplatform genomics data," McGuffey, E. J., Morris, J. S., Manyam, G. C., Carroll, R. J., and Baladandayuthapani, V., in *Integrating Omics Data*, eds. Tseng, G. C., Ghosh, D., and Zhou, X. J.

lated into a protein, and a specific action is carried out by the protein. The segments of mRNA that code for different proteins are known as genes, and cancer is believed to involve a complex interaction of these genes. In addition to the direct DNA to mRNA to protein process, there are many different genetic alterations and interferences, such as methylation, copy number, and loss of heterozygosity (LOH), that have the potential to affect gene expression (the abundance of mRNA) and thus eventually impact clinical outcomes that manifest as symptoms of disease development (see Figure 4.1). The discovery of which genes are significantly associated with patient-specific outcomes and understanding the biological mechanisms associated with such genes' expression are critical to developing targeted therapies.

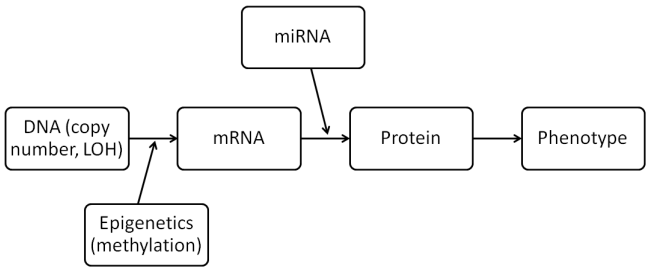


Figure 4.1: Schematic of the multiple molecular platforms and their biological relationships. Reprinted with permission from McGuffey et al. (2015).

Owing to rapid technological advances combined with decreasing costs, multiplatform data sets are becoming increasingly available on matched samples. Such data sets provide measurements from multiple platforms (mRNA, protein, methylation, copy number, genotype, etc.) on each patient involved in the study. (We note that when we use the term *platform* in this chapter, we are referring to a biological entity or the molecular characteristic such as methylation, copy number, or expression.) Although previous analyses generally studied the effect of a single platform on a

clinical outcome, these comprehensive data sets have motivated the development of statistical models that integrate data from several platforms, facilitating a deeper and more complete understanding of the genetic causes of cancer development and progression and offering the potential to increase power and lower false discovery rates (Wang et al., 2013). Statistical methods attempting to achieve this goal face several analytic challenges: high dimensionality, complex correlation structures, and unclear (biological) interpretations.

1. *High dimensionality* arises when the number of predictors or variables (usually on the order of thousands) is greater than the number of patients or samples (usually on the order of a few hundred); this is common when modeling genomic data. It is well known that only a few of these variables play an active role in disease modulation. Thus, effective strategies need to be developed to address these challenges, especially by inducing shrinkage and sparsity.

2. The issue of *complex correlation structures* refers to the correlations between genes (within a platform) as well as the correlations between platforms that arise due to their complex biological relationships. A model that accounts for these biological relationships should produce more accurate and efficient effect estimates.

3. Last, *interpretations* can become complex, especially depending on the technique employed to handle high dimensionality. Even if the method provides clear interpretations, it is important that there is a straightforward scientific translation. If the results are to be useful in developing targeted therapies, estimates from an integrative model should offer direct insight into the mechanics and the genes related to the clinical outcome.

Many models have been developed recently that face these challenges and integrate data from multiple platforms. The Cancer Genome Atlas (TCGA) has performed large-scale studies on ovarian cancer and glioblastoma multiforme (GBM), a brain tumor, in which they analyzed data from multiple platforms such as gene mutations, microRNA expression, and mRNA expression. TCGA researchers began by analyzing each platform separately and then combined the platform results to draw integrated conclusions. The ovarian cancer study identified influential gene pathways and made new discoveries regarding gene interactions, while the GBM study discovered a previously unknown link between MGMT methylation and the mutation spectra of mismatch repair genes (McLendon et al., 2008; Bell et al., 2011). By combining the information gained from multiple platforms, instead of focusing on only one type of data, researchers were able to make novel discoveries and gain important insights into the genetic factors of these cancers.

Other methods integrate multiple platforms directly by incorporating them all into a single model. Tyekucheva et al. (2011) proposed integrating data from multiple platforms as predictors in a single logistic model. Their method identifies significant gene sets, and they show that their integrative approach has more power than approaches using a single platform of data (Tyekucheva et al., 2011). Lanckriet et al.'s (2004) integrative method, in which data from each genomic platform are first represented as kernel functions and then all the kernels are included in a classification model, was also shown to have more power than an analogous single-platform approach. Another approach that integrates multiple genomic platforms into a single model was proposed by Shen et al. (2009). Their approach, iCluster, uses joint latent variable models to cluster samples into tumor subtypes. Through applications to breast and lung cancer data, iCluster identified potential novel tumor subtypes (Shen et al., 2009). Verhaak et al. (2010) performed an analysis on the GBM data

from TCGA in which they combined gene expression data from multiple types of microarray assays to classify tumors into four distinct subtypes and to identify which genes have significant influence on the classification. Because the subtypes respond differently to treatments, this classification information is valuable in making treatment decisions (Verhaak et al., 2010). These and other integrative methods improve our understanding of the genetic causes and mechanics of cancer, and they allow us to make increasingly informed prognosis and treatment decisions.

The integrative Bayesian analysis of genomics data (iBAG) model we present in this chapter was originally proposed by Wang et al. (2013) and was later generalized by Jennings et al. (2013). It is a two-step Bayesian hierarchical model that integrates data from an arbitrary number of platforms while taking into account the biological relationships between DNA characteristics (such as methylation and copy number) and RNA-level entities (such as gene expression). In Section 4.2, we present the model details (both for the linear case and with a novel nonlinear extension), and we explain how each of the challenges described is overcome. In Section 4.3, we illustrate the method on a publicly available GBM data set, and we offer a discussion in Section 4.4.

4.2 iBAG Models

The basic construction of the iBAG model consists of two components: a *mechanistic model* that attempts to capture mechanistic information by partitioning the gene expression into components explained by different upstream platforms, and a *clinical model* that subsequently incorporates these components to model the effects on a clinical outcome of interest. Through the joint estimation of these components, we not only identify the genes significantly related to the clinical outcome, but we also gain insight into the biological mechanisms modulating these effects based on

the information from the upstream platforms. (Note that throughout this chapter, when we refer to modulation by one or more platforms, we mean regulation by that platform or by the interaction across platforms that regulate at different levels. This is not to be confused with the term *modulation* as it is used regarding dynamic conditions in network data.) We present the linear construction first for ease of exposition (Section 4.2.1) and subsequently propose a more flexible non-linear extension in Section 4.2.2.

First we present the notation common to both the linear and non-linear formulations. Let n be the number of patients, J the number of platforms being integrated, and p_j the number of genes from platform j, and then define the following terms:

- Variable mRNA_g is the level of gene expression for gene g (where $g = 1, ..., \max(p_j);\ j = 1, ..., J$) and is of dimension $(n \times 1)$.

- Variable X_{jg} is the part of gene g expression that is attributed to upstream platform j and is of dimension $(n \times 1)$. Specifically, X_{jg} is the product of some platform-specific jth predictor and a fitted coefficient. Details are expounded later in the chapter.

- O_g represents the "other" (remaining) part of the gene expression that is explained by something other than the $J - 1$ upstream platforms and is of dimension $(n \times 1)$.

- The covariates X_j and O are the vectorized gene expression effects attributed to platform j $(j = 1, ..., J - 1)$ and other sources, respectively, and are estimated from the mechanistic model. Each X_j is of dimension $(n \times p_j)$, and O is of dimension $(n \times \max(p_j))$.

- Y denotes the clinical outcome and is of dimension $(n \times 1)$; it is assumed

continuous for now.

- Parameter $\boldsymbol{\beta}_j$ is the vector of effects of platform j on Y and is of dimension $(p_j \times 1)$.

- Parameter $\boldsymbol{\epsilon}$ is the error term and is of dimension $(n \times 1)$.

4.2.1 Linear case

The linear iBAG model is as follows:

$$\text{mRNA}_g = \sum_{j=1}^{J-1} \text{X}_{jg} + \text{O}_g \tag{4.1}$$

$$Y = \sum_{j=1}^{J-1} \text{X}_j \boldsymbol{\beta}_j + \text{O}\boldsymbol{\beta}_J + \boldsymbol{\epsilon} \tag{4.2}$$

where Equation 4.1 is the mechanistic model and Equation 4.2 is the clinical model.

4.2.1.1 Mechanistic model

The mechanistic model takes into account the biological relationships between platforms by modeling gene expression as it is affected by each of an arbitrary number of upstream platforms known to influence gene expression. Each of the terms in the mechanistic model is defined as follows.

In the linear case, estimation of the components X_{jg} and O_g is done via least squares regression, and a separate regression is fit for each gene. The mRNA expression is typically summarized for each gene, but the number of raw data values from the other platforms associated with each gene can vary greatly owing to multiple measurements within each gene, that is, different probes mapped to specific genomic locations within the gene. For example, in the data we use in Section 4.3, the number of methylation values per gene ranges from 1 to 8, and the number of copy number values per gene ranges from 1 to 16. Thus, to prevent complications due to high dimensionality, and to match to specific genes, for each gene, we begin by performing separate principal component analyses (PCAs) on each of the upstream platforms.

For each upstream platform, we retain the principal components that account for at least 90% of the variation, and then regress mRNA_g on the corresponding PC scores. For gene g, our estimate of X_{jg} is the linear combination of PC scores from platform j and the corresponding coefficient estimates; specifically,

$$\text{X}_{jg} = \sum_{k=1}^{K} R_{jgk} B_{jgk} \tag{4.3}$$

where R_{jgk} is the raw data value from platform j for gene g with $K = 1$ if there is only a single raw data value or the kth PC score from platform j for gene g if there are multiple markers, and where B_{jgk} is the vector of regression coefficients. O_{jg} is then estimated as the residuals from the least squares regression.

After these steps have been carried out for each gene, we have partitioned each gene into J pieces: $j - 1$ X_{jg}s that represent the part of gene g expression explained by upstream platform j and 1 O_g that represents the part of gene g expression not explained by any of the included upstream platforms. We carry forward the X_{jg}s and O_gs into the clinical model.

4.2.1.2 Clinical model

This component models the effect that each of the pieces estimated in the mechanistic model has on a (continuous) clinical outcome. Each of the terms in the clinical model (Equation 4.2) are defined as follows.

Our main goal is accurate and efficient estimation of the effects $\boldsymbol{\beta}_j$. If a β_{jg}, the coefficient associated with platform j of gene g, is flagged as "significant," then this reveals several aspects: (1) gene i is related to the clinical outcome, (2) platform j is modulating the expression, and (3) a unit increase of gene g expression attributed to platform j is associated with a β_{jg} change in the clinical outcome. To achieve the expected sparsity in the βs (and combat high dimensionality), we employ a Bayesian hierarchical setup, including a sparsity-inducing prior on the $\boldsymbol{\beta}_j$s. As stated previ-

ously, it is also important to preserve the estimates of the truly large or important βs, so we must choose a prior that allows for flexible shrinkage. Two popular choices of shrinkage priors are the spike and slab prior and the Laplace prior. The spike and slab is a mixture of a "spike" at 0 and a Gaussian "slab" distribution. It does facilitate shrinkage, but the shrinkage asymptotes to a constant, resulting in large effects being shrunk just as much as small effects. In our scenario, we want to avoid this because we believe the large effects are the ones that are truly important, whereas the effects close to zero are not of interest. The Laplace prior (or the normal-exponential prior) results in the Bayesian lasso formulation (Park and Casella, 2008). The shrinkage offered by the Bayesian lasso is more flexible than that offered by many other priors, including spike and slab; however the single hyperparameter in the Laplace prior limits its flexibility, and the estimates of large effects are shrunk toward zero along with the smaller effects (Griffin and Brown, 2010).

To induce sparsity but still allow for accurate estimation of truly large, nonzero effects, we choose to employ the normal-gamma prior, which provides more adaptive shrinkage than the lasso prior through its two hyperparameters (Griffin and Brown, 2010). Our complete hierarchy (based on a formulation by Griffin and Brown (2010)) is as follows:

$$
\begin{aligned}
\mathbf{Y} &= \text{Normal}(X\boldsymbol{\beta}, \sigma^2\mathbf{I}_n) \\
\boldsymbol{\beta} &= \text{Normal}(\mathbf{0}_{\tilde{p}}, D_\psi) \text{ where } D_\psi = \text{diag}(\psi_{11}, ..., \psi_{1p_1}, ..., \psi_{J1}, ..., \psi_{Jp_J}) \\
\psi_{jg} &= \text{Gamma}(\lambda_j, 1/(2\gamma_j^2)) \\
\sigma^2 &= \text{InverseGamma}(a, b) \\
\lambda_j &= \text{Exponential}(c) \\
\gamma_j^{-2} &= \text{Gamma}(\tilde{a}, \tilde{b}/(2\lambda_j))
\end{aligned}
$$

where $\tilde{p} = \sum_{j=1}^{J} p_j$ is the total number of predictors in the model. All of the complete

conditional distributions are in closed form, except for that of λ_j, so we estimate the parameters via Gibbs sampling, with a random walk Metropolis-Hastings update step for the λ_js. (See Appendix B for complete conditionals.)

4.2.1.3 Marker selection

After we obtain MCMC samples, we can obtain point estimates of the parameters in the clinical model (in particular, $\boldsymbol{\beta}$) by simply taking the mean of our posterior samples. We also need to identify which markers (gene-platform combinations) to flag as significantly related to the clinical outcome. We use a method based on the median probability model (Barbieri and Berger, 2004), and we flag important markers through these steps:

1. Based on practical considerations, define a minimum effect size δ that is of interest. Further define (δ_-, δ_+) as the region such that you want to flag markers whose effects are outside that region, that is, less than δ_- or greater than δ_+.

2. Given S MCMC samples and $\beta_{jg}^{(s)}$ is the β_{jg} sample from iteration s, calculate the following posterior probabilities: $p_+(x_{jg}) = \sum_{s=1}^{S} \mathbf{I}(\beta_{jg}^{(s)} > \delta_+)/S$, the posterior probability that β_{jg} is greater than the practical cutoff δ_+, and $p_-(x_{jg}) = \sum_{s=1}^{S} \mathbf{I}(\beta_{jg}^{(s)} < \delta_-)/S$, the posterior probability that β_{jg} is smaller than the practical cutoff δ_-.

3. Flag a marker as significant if either $p_+(\beta_{jg})$ or $p_-(\beta_{jg})$ is greater than 0.5.

4.2.2 Nonlinear extensions

Previous publications (Wang et al., 2013; Jennings et al., 2013) have been limited to linear iBAG models, whereas here we introduce a nonlinear version of the model. In the linear iBAG case, we assume that the predictors in the mechanistic model (Equation 4.1) are linearly related to the gene expression. Depending on the

upstream platforms included in the analysis and the available information (or lack thereof) regarding how those platforms affect gene expression, such an assumption may not be reasonable. In that case, for each gene, we still begin by performing PCA on each platform's raw values, but a more flexible modeling technique for estimating the pieces of the mechanistic model might be necessary to accommodate the additional flexibility.

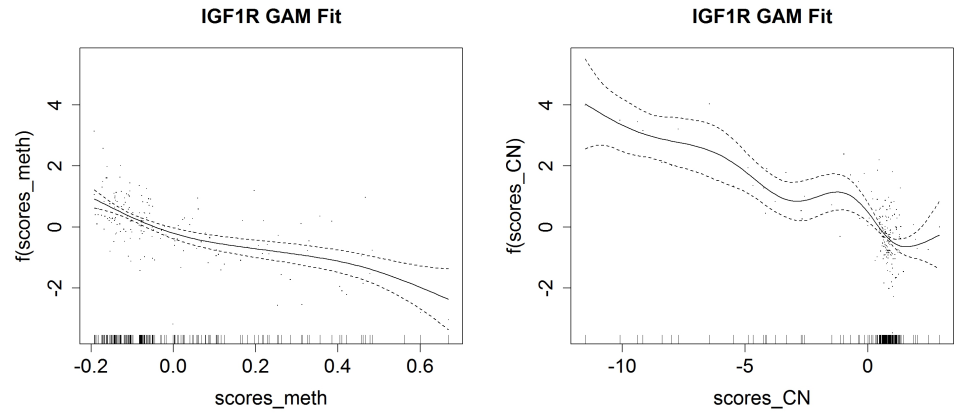
To achieve this flexibility in the mechanistic model, we propose using generalized additive models (GAM), a class of models proposed by Hastie and Tibshirani in 1986 that replaces the linear $X\beta$ in the generalized linear model formulation for exponential family responses with a sum of smooth functions (Hastie and Tibshirani, 1986). The specific formulation for our non-linear mechanistic model is as follows:

$$g\{E(\text{mRNA}_g)\} = b_0 + \sum_{j=1}^{J}\sum_{k=1}^{K_j} f_{jgk}(R_{jgk}) \tag{4.4}$$

where $g(\cdot)$ is a specified link function, $f_{jgk}(\cdot)$ is a smooth function, and R_{jgk} is the kth PC score for platform j of gene g (or the raw values for platform j of gene g if $k = 1$).

The choices we must make are what to use as the $g(\cdot)$ and $f(\cdot)$ functions. The $g(\cdot)$ function is a link function chosen based on the exponential family chosen for the outcome; in our case, we have continuous mRNA values as the response, so we model them as normal with $g(\cdot)$ as the identity funcion. As for the smooth function $f(\cdot)$, we require a good fit to the data, but one that does not *overfit*. We propose using penalized regression splines, which fit the data closely but include a penalty for too much "wiggliness" as the fit becomes closer to an interpolation. We implement this model using Wood's R package `mgcv` and take advantage of the option of automatic smoothness selection for the penalty parameter using generalized cross-validation (GCV) (Wood, 2014). Figure 4.2 illustrates the fit from GAM for *IGF1R*, one of the genes we use in our analysis in Section 4.3. This particular gene only has one

value for each of two platforms (methylation and copy number), so there were only two smooth functions needed ($f(R_{1,18,1})$ and $f(R_{2,18,1})$), both of which are shown.



(a) Methylation data. (b) Copy number data.

Figure 4.2: For gene *IGF1R* the fitted smooth curves for the methylation data ($f(R_{1,18,1})$) and for the copy number data ($f(R_{1,18,1})$) are plotted. The dots are the partial residuals, that is, the residuals that would have arisen from not including the predictor of interest (methylation for panel (a) and copy number for panel (b)) but keeping the other estimates fixed. The hash marks on the x-axis are the data values, and the error bounds extend 2 standard deviations above and below the smooth estimate. Reprinted with permission from McGuffey et al. (2015).

After fitting the nonlinear mechanistic model using GAM, we estimate each X_{jg} as $\sum_{k=1}^{K_j} \widehat{f_{jgk}}(R_{jgk})$ and the O_is as the residuals. We carry these forth into the clinical model, which is the same hierarchical Bayesian model presented in the linear case. The estimation of parameters in the clinical model progresses as previously discussed in the linear case, with the normal-gamma prior facilitating adaptive shrinkage. In general, two potential disadvantages of using nonparametric smooth functions as the $f(\cdot)$ functions are (1) a lack of parsimony and (2) possible difficulties in obtaining a straightforward scientific interpretation of the parameter estimates. For some scenar-

ios, this can be problematic, but our goal in this step is simply to partition the gene expression (as accurately as possible) into pieces explained by the different upstream platforms; we are not concerned about parsimony or precise interpretations until the clinical model (Equation 4.2). Because the estimates from the smooth functions give us the partitioned pieces we need, while the lack of parametric assumptions provides the flexibility for producing a closer fit to the data, we can use them without reservation.

4.3 Illustrations

We apply our method to a publicly available GBM data set from TCGA. TCGA is an organization that began in 2006 with the goal of compiling and analyzing comprehensive genomic data sets for different types of cancer. Thus far, they provide data on more than 20 types of cancer, with GBM being one of the first that they studied (The Cancer Genome Atlas, 2012). The American Cancer Society estimates that in 2014, there will be 23,380 new cases of brain and nervous system cancers, with 14,320 American fatalities from such cancers (American Cancer Society, 2014). GBM is one of the most common and lethal brain tumors, primarily affecting people between 45 and 70 years of age (American Association of Neurological Surgeons, 2012). Without radiation treatment, a person with GBM usually lives less than 3 months; even with radiation treatment, the typical survival time is under 15 months (Johnson and O'Neill, 2012). Our goal is to apply our method to identify prognostic biomarkers related to GBM development and patient survival, which could potentially be used in developing a targeted therapy.

4.3.1 Data description

We consider a subset of the available TCGA GBM data, consisting of mRNA expression, two upstream platforms known to affect gene expression (methylation

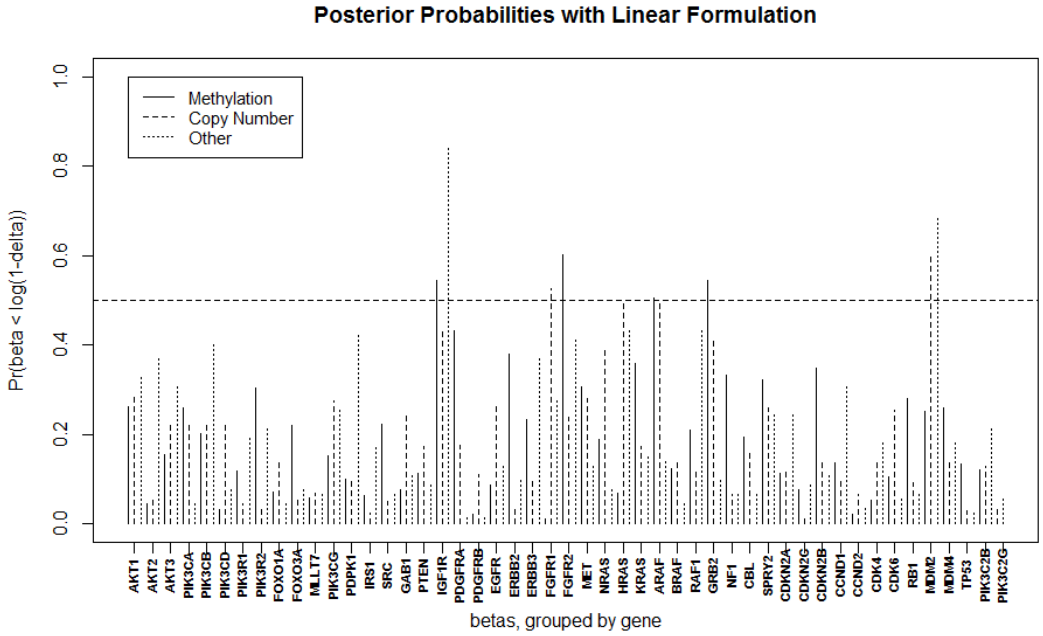
and copy number), and uncensored survival time (in days) for 163 patients. Our copy number data are level 2 data from the HG_CGH_244A platform; they are the normalized signal for copy number alterations of aggregated regions per probe. Our methylation data are level 3 data from the HumanMethylation27K arrays; they are the methylated sites along a gene (probe-level data). Our expression data are level 3 data (summarized per gene) from the Affymetrix profiled HT_HG_U133A platform (The Cancer Genome Atlas Data Portal, 2013). Because our clinical response is survival time, we use an accelerated failure time (AFT) model, taking Y (in Equation 4.2) to be log(survival) (Wei, 1992). We focus our analysis on 49 genes from three signaling pathways important to GBM: RTK/PI3K, P53, and RB (Memorial Sloan-Kettering Cancer Center, 2012). One gene has no methylation data, so we remove its corresponding column in the X matrix; any effect due to methylation would then be captured by the "other" predictor in the clinical model.

After standardizing the predictors and imputing the few missing values, we apply our method and obtain 10,000 MCMC samples, with 500 used as burn-in. In determining whether to flag a marker as "significant," we choose our practical minimum effect size to correspond to a 5% change in survival time, resulting in $(\delta_-, \delta_+) = (\log(0.95), \log(1.05))$. As discussed in Section 4.2, we flag a marker as significant if $p_+(\beta_{jg}) > 0.5$ or if $p_-(\beta_{jg}) > 0.5$.

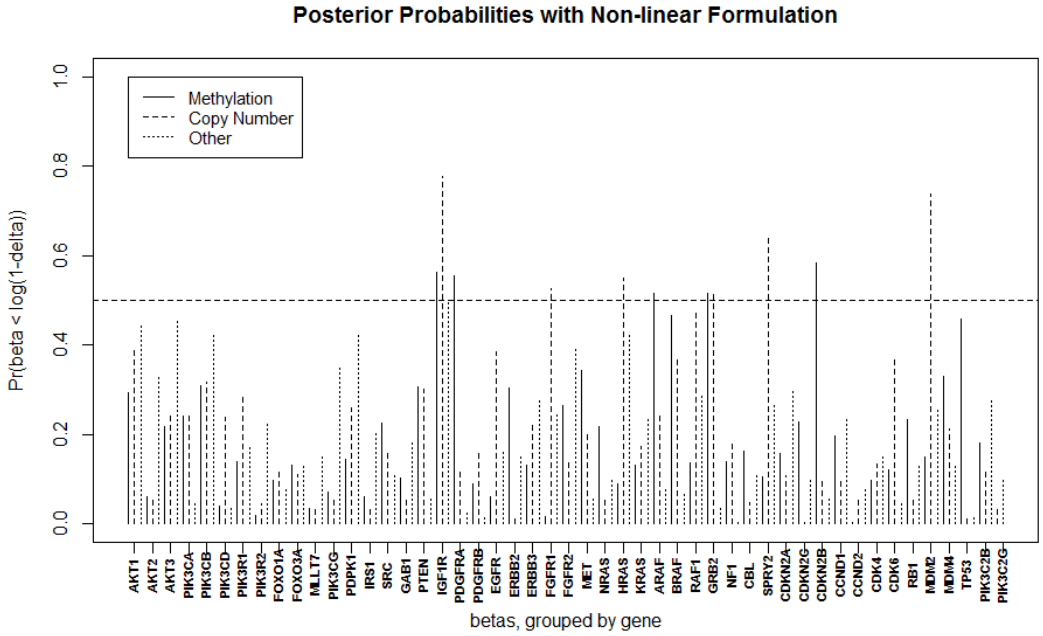
4.3.2 Results

Using the linear formulation of our method, we identify 21 significant prognostic markers, 12 of them positive markers (more gene expression explained by that platform, better prognosis) and 9 of them negative markers (more gene expression explained by that platform, worse prognosis). Figures 4.3a and 4.4a show the posterior probabilities of β_{jg} being less than δ_- and greater than δ_+, respectively. Figure

4.5a shows the posterior means of each β_{jg} as well as which markers were flagged as significant. The genes with the 12 positive markers were *PDGFRB*, *FGFR1*, *CCND2*, *PIK3R2*, *IRS1*, *CDKN2C*, *TP53*, *PIK3CA*, and *PDGFRA*. In the following results details, the percentage given after the gene name is the percentage of that gene expression explained by the flagged platform under the linear formulation (see Appendix B for calculation details). Genes *PDGFRB* (0.2%), *FGFR1* (2.0%), and *CCND2* (6.0%) were determined to be related to patient survival through methylation effects, whereas expression of *PIK3R2* (7.6%), *IRS1* (4.8%), *CDKN2C* (19.7%), and *TP53* (16.0%) was related to patient survival through copy number. For *PIK3CA* (70.0%), *PDGFRA* (44.5%), *PDGFRB* (98.5%), *CCND2* (81.2%), and *TP53* (83.1%), gene expression was related to patient survival through some other unspecified mechanism. The genes with the nine negative markers were *IGF1R*, *FGFR2*, *ARAF*, *GRB2*, *FGFR1*, and *MDM2*. *IGF1R* (0.3%), *FGFR2* (6.1%), *ARAF* (0.1%), and *GRB2* (0.4%) were related to clinical response through methylation, whereas *FGFR1* (8.8%), *ARAF* (2.8%), and *MDM2* (81.4%) were related through copy number, and *IGF1R* (84.9%) and *MDM2* (18.4%) were related through some mechanism other than methylation or copy number. Note that seven genes (*IGF1R*, *PDGFRB*, *FGFR1*, *ARAF*, *CCND2*, *MDM2*, and *TP53*) are found to be significant on two different platforms. Our method has not only identified 14 genes as having a significant effect on survival, but it has also determined which platform(s) of those genes is (are) modulating the effect. In addition, we can use the posterior means of the effects to gain more specific insight; a one-unit increase in gene g expression attributed to platform j is estimated to result in a $\{\exp(\widehat{\beta_{jg}}) - 1\} \times 100\%$ change in survival time.

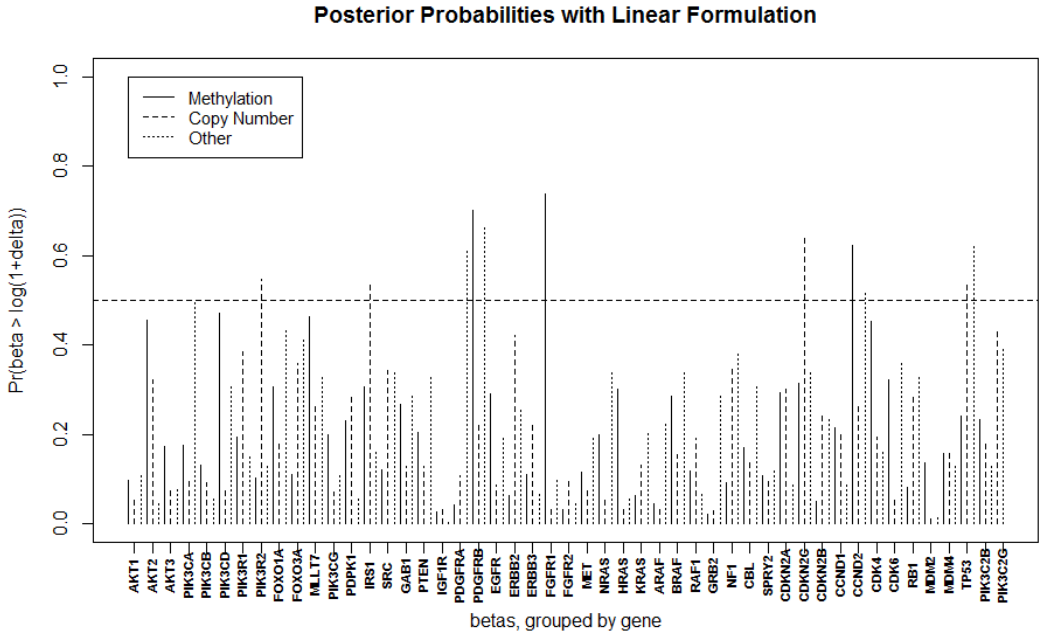


(a) Linear formulation.

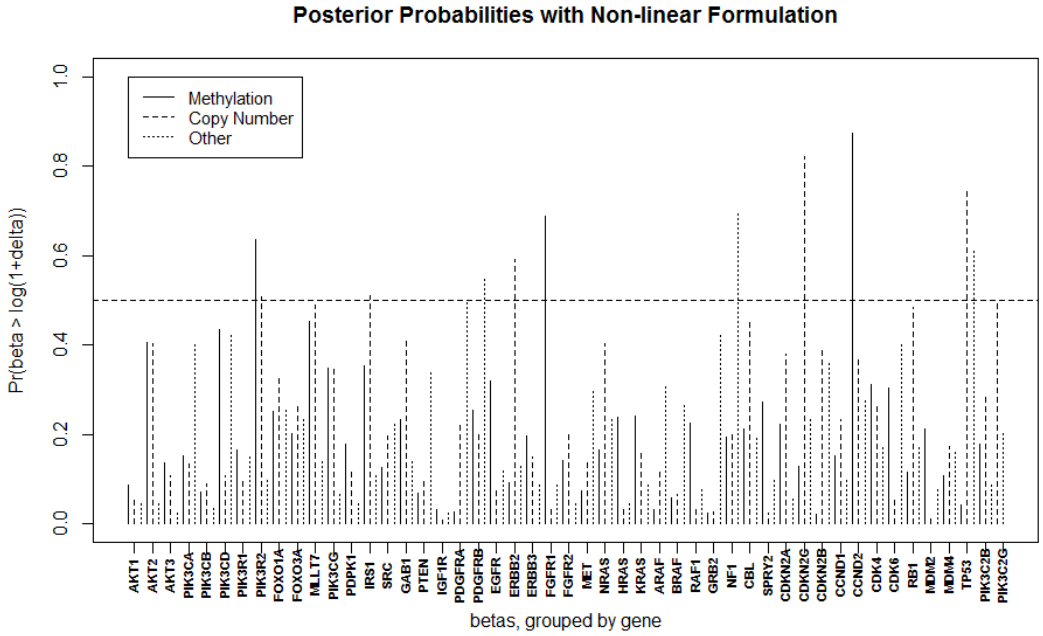


(b) Nonlinear formulation.

Figure 4.3: The posterior probabilities (based on MCMC samples) that $\beta_{jg} < \delta_{-}$. Reprinted with permission from McGuffey et al. (2015).

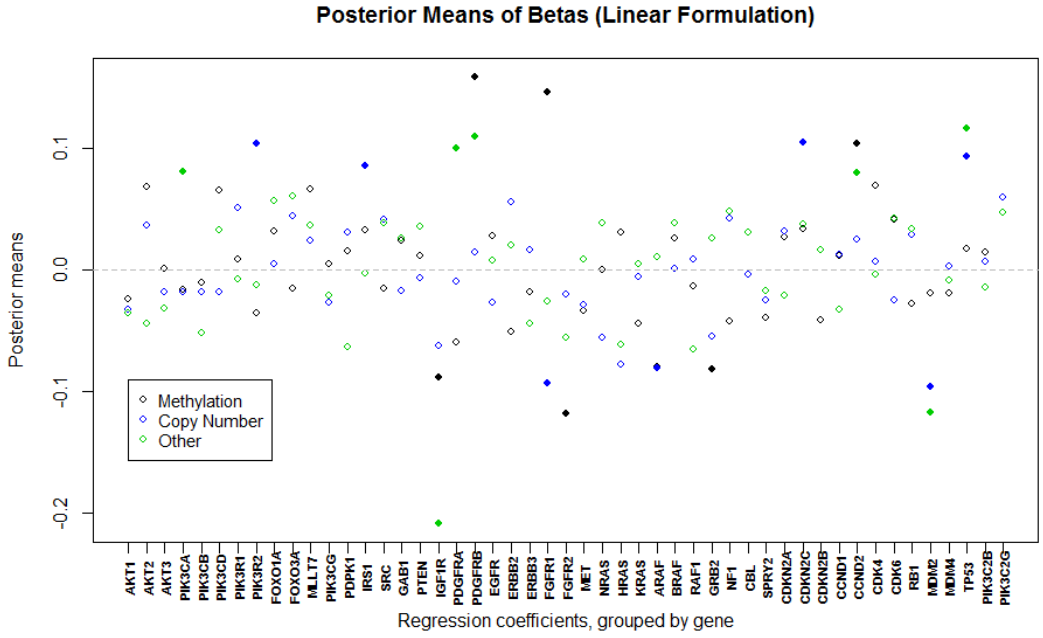


(a) Linear formulation.

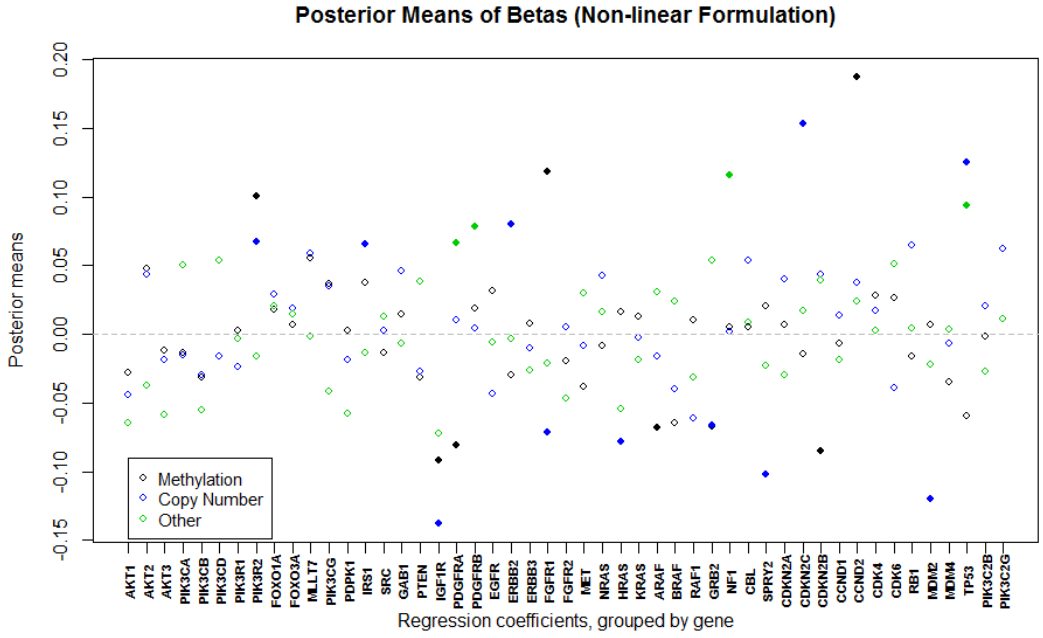


(b) Nonlinear formulation.

Figure 4.4: The posterior probabilities (based on MCMC samples) that $\beta_{jg} > \delta_+$. Reprinted with permission from McGuffey et al. (2015).



(a) Linear formulation.



(b) Nonlinear formulation.

Figure 4.5: The estimates (posterior means) of the regression coefficients in the clinical model (β_{jg}s) are shown, with the multiple platforms for each gene labeled by color. Solid plot markers indicate that the effect was found to be significant. Reprinted with permission from McGuffey et al. (2015).

Using the nonlinear formulation of our method, we see some of the same markers flagged, but there are also some differences: we identify 23 significant prognostic markers, 12 positive markers, and 11 negative markers. Figures 4.3b and 4.4b show the posterior probabilities of β_{jg} being less than δ_- and greater than δ_+, respectively. Figure 4.5b shows the posterior means of each β_{jg} as well as which markers were flagged as significant. The genes with the 12 positive markers were *PIK3R2*, *FGFR1*, *CCND2*, *IRS1*, *ERBB2*, *CDKN2C*, *TP53*, *PDGFRA*, *PDGFRB*, and *NF1*. In the following results details, the percentage given after the gene name is the percent of that gene expression explained by the flagged platform under the nonlinear formulation (see Appendix B for calculation details). Genes *PIK3R2* (2.9%), *FGFR1* (3.7%), and *CCND2* (8.9%) were determined to be related to clinical outcome through methylation effects, whereas expression of *PIK3R2* (9.8%), *IRS1* (4.8%), *ERBB2* (8.0%), *CDKN2C* (46.1%), and *TP53* (20.3%) was related to clinical outcome through copy number. For *PDGFRA* (44.5%), *PDGFRB* (96.3%), *NF1* (40.0%), and *TP53* (70.7%), gene expression was related to clinical outcome through some other unspecified mechanism. The genes with the 11 negative markers were *IGF1R*, *PDGFRA*, *ARAF*, *GRB2*, *CDKN2B*, *FGFR1*, *HRAS*, *SPRY2*, and *MDM2*. *IGF1R* (<0.1%), *PDGFRA* (19.6%), *ARAF* (0.2%), *GRB2* (0.3%), and *CDKN2B* (6.7%) were related to clinical response through methylation, whereas *IGFR1* (23.8%), *FGFR1* (19.3%), *HRAS* (13.9%), *GRB2* (23.7%), *SPRY2* (16.1%), and *MDM2* (86.9%) were related through copy number. Note that six genes (*PIK3R2*, *FGFR1*, *TP53*, *PDGFRA*, *IGR1R*, and *GRB2*) are found to be significant on two different platforms.

Fourteen markers are flagged identically (same gene, same platform, same sign) by both the linear and nonlinear model formulations. Differences between the results in moving from the linear formulation to the nonlinear formulation include the

following. First, *PIK3CA* and *FGFR2* are no longer flagged, whereas *ERBB2*, *NF1*, *CDKN2B*, *HRAS*, and *SPRY2* appear as significant. In four instances, a gene goes from being flagged on two platforms to only being flagged on one: (1) instead of being identified on methylation and "other" platforms, *PDGFRB* is only identified on other platforms; (2) instead of being flagged on methylation and other platforms, *CCND2* is only flagged on methylation effects; (3) *ARAF* goes from being flagged on both methylation and copy number to only methylation; and (4) *MDM2* goes from being flagged on both copy number and other platforms to only on copy number. In addition, three genes go from being flagged on one platform to being identified as significant on two: (1) *PIK3R2* is still flagged on copy number, but it is also flagged on methylation effects; (2) *PDGFRA* is still flagged as a positive marker modulated by other platforms, but it becomes flagged as a negative marker on methylation in addition; and (3) *GRB2* is flagged on copy number effects, in addition to maintaining significance on methylation effects. Finally, we see *IGF1R* still flagged on two platforms, but one of them changes; instead of being flagged on both methylation and other effects, *IGF1R* is flagged on methylation and copy number effects. Tables 4.1 and 4.2 display the markers flagged as important negative and positive prognostic markers, respectively, and they also provide a comparison of results from the linear versus nonlinear formulations.

Because, in the case of our data, we have no compelling reason to believe that the relationship between methylation, copy number, and gene expression is truly linear, we consider the results from the nonlinear formulation to be more accurate. That there are differences in the results at all speaks to the fact that we must be careful not to make unfounded assumptions on the specific structure of the biological relationships modeled in the mechanistic model. With a more flexible partitioning of gene expression, we see two genes no longer found to be significant, whereas five

Table 4.1: Results: Negative markers. All 49 genes appearing in the data are listed, along with the three platforms (M,CN,O). Genes are bolded if they were found to have a significant negative prognostic marker on any platform by either the linear or nonlinear formulation. Italic platforms indicate that the platform was flagged by the linear formulation, and underlined platforms indicate that the platform was flagged by the nonlinear formulation. Reprinted with permission from McGuffey et al. (2015).

AKT1 (M,CN,O)	**IGF1R** (*M*,CN,*O*)	CBL (M,CN,O)
AKT2 (M,CN,O)	**PDGFRA** (M,CN,O)	**SPRY2** (M,CN,O)
AKT3 (M,CN,O)	PDGFRB (M,CN,O)	CDKN2A (M,CN,O)
PIK3CA (M,CN,O)	EGFR (M,CN,O)	CDKN2C (M,CN,O)
PIK3CB (M,CN,O)	ERBB2 (M,CN,O)	**CDKN2B** (M,CN,O)
PIK3CD (M,CN,O)	ERBB3 (M,CN,O)	CCND1 (M,CN,O)
PIK3R1 (M,CN,O)	**FGFR1** (M,*CN*,O)	CCND2 (M,CN,O)
PIK3R2 (M,CN,O)	**FGFR2** (*M*,CN,O)	CDK4 (M,CN,O)
FOXO1A (M,CN,O)	MET (M,CN,O)	CDK6 (M,CN,O)
FOXO3A (M,CN,O)	NRAS (M,CN,O)	RB1 (M,CN,O)
MLLT7 (M,CN,O)	**HRAS** (M,CN,O)	**MDM2** (M,*CN*,*O*)
PIK3CG (M,CN,O)	KRAS (M,CN,O)	MDM4 (M,CN,O)
PDPK1 (M,CN,O)	**ARAF** (*M*,*CN*,O)	TP53 (M,CN,O)
IRS1 (M,CN,O)	BRAF (M,CN,O)	PIK3C2B (M,CN,O)
SRC (M,CN,O)	RAF1 (M,CN,O)	PIK3C2G (CN,O)
GAB1 (M,CN,O)	**GRB2** (*M*,CN,O)	
PTEN (M,CN,O)	NF1 (M,CN,O)	

previously unidentified genes were flagged. Also, we see differences in results on one platform for eight other genes; the difference in partitioning in the mechanistic model appears to have had a direct effect on which platform(s) was (were) identified as significant, reinforcing the importance of accuracy in that first step.

4.4 Discussion

We have proposed a two-step, hierarchical Bayesian model to integrate different types of genomic data in a single model with the goal of identifying genetic markers significant to a clinical outcome. In addition, we have presented two different

Table 4.2: Results: Positive markers. All 49 genes appearing in the data are listed, along with the three platforms (M,CN,O). Genes are bolded if they were found to have a significant positive prognostic marker on any platform by either the linear or nonlinear formulation. Italic platforms indicate that the platform was flagged by the linear formulation, and underlined platforms indicate that the platform was flagged by the nonlinear formulation. Reprinted with permission from McGuffey et al. (2015).

AKT1 (M,CN,O)	IGF1R (M,CN,O)	CBL (M,CN,O)
AKT2 (M,CN,O)	**PDGFRA** (M,CN,*O*)	SPRY2 (M,CN,O)
AKT3 (M,CN,O)	**PDGFRB** (*M*,CN,*O*)	CDKN2A (M,CN,O)
PIK3CA (M,CN,*O*)	EGFR (M,CN,O)	**CDKN2C** (M,*CN*,O)
PIK3CB (M,CN,O)	**ERBB2** (M,CN,O)	CDKN2B (M,CN,O)
PIK3CD (M,CN,O)	ERBB3 (M,CN,O)	CCND1 (M,CN,O)
PIK3R1 (M,CN,O)	**FGFR1** (*M*,CN,O)	**CCND2** (*M*,CN,*O*)
PIK3R2 (M,*CN*,O)	FGFR2 (M,CN,O)	CDK4 (M,CN,O)
FOXO1A (M,CN,O)	MET (M,CN,O)	CDK6 (M,CN,O)
FOXO3A (M,CN,O)	NRAS (M,CN,O)	RB1 (M,CN,O)
MLLT7 (M,CN,O)	HRAS (M,CN,O)	MDM2 (M,CN,O)
PIK3CG (M,CN,O)	KRAS (M,CN,O)	MDM4 (M,CN,O)
PDPK1 (M,CN,O)	ARAF (M,CN,O)	**TP53** (M,*CN*,*O*)
IRS1 (M,*CN*,O)	BRAF (M,CN,O)	PIK3C2B (M,CN,O)
SRC (M,CN,O)	RAF1 (M,CN,O)	PIK3C2G (CN,O)
GAB1 (M,CN,O)	GRB2 (M,CN,O)	
PTEN (M,CN,O)	**NF1** (M,CN,O)	

formulations of the first step of the model (the mechanistic model) that facilitate more or less flexibility of gene expression partitioning, depending on the amount and nature of prior knowledge regarding the structure of biological relationships among data platforms. After applying our method to a TCGA brain tumor data set, we identify 21 significant prognostic markers on 14 genes using the linear formulation and 23 significant prognostic markers on 17 genes using the nonlinear formulation.

Advantages of our method are numerous. (1) Our model flexibly incorporates multiple types of genetic data in a single model. (2) We utilize principal components

and a normal-gamma prior on the effects to effectively induce flexible shrinkage and combat high dimensionality. (3) Our model accounts for the known biological relationships between DNA characteristics and RNA-level entities, which allows us not only to identify which genes are significant to the clinical outcome but also to obtain the mechanistic information of which platform(s) is (are) modulating the effect. This information is critical to developing targeted cancer therapies. (4) Our effect estimates have a direct interpretation. (5) Our method can be applied to gain insight regarding any type of cancer, as long as an appropriate data set is available and the biological relationships among the platforms are understood.

Note that we currently use two steps (mechanistic and clinical models), but additional layers could be constructed to incorporate other platforms, such as protein expression data. For example, protein expression would be regressed on mRNA expression, with an additional step of mRNA regressed on platforms such as copy number, miRNA, mutation status, and so on. When the data are available, incorporating this additional step would give the potential of obtaining even more insight into the mechanics of what exactly is driving the clinical phenotype expression.

5. PIBAG: HIERARCHICAL PATHWAY SHRINKAGE IN INTEGRATIVE GENOMICS

5.1 Introduction

Targeted cancer therapies are a class of cancer treatments that target the specific molecular cause(s) of a type of cancer, with the general goal of preventing cancer growth or possibly eliminating the cancer altogether, depending on factors such as the cancer stage and metastasis level. Targeted treatments often are administered as a pill or intravenously, and they typically have reduced and less severe side effects as compared to traditional treatment options such as chemotherapy and radiation. Many targeted drugs are currently available. For example, gefitinib targets EGFR to treat advanced non-small cell lung cancer; sunitinib is a multi-targeted kinase inhibitor that treats advanced kidney cancer and some gastrointestinal stromal tumors; and trastuzumab is used to treat metastatic breast cancer in cases where the protein HER2/neu is overexpressed (American Cancer Society, 2015b; National Cancer Institute, 2015). Still other drugs are being tested in clinical trials, and research into other targets is ongoing.

Identifying effective targets, that is, the specific genes and genetic mechanisms that are involved in cancer development and progression, provides a first step in the development of such treatment options. The search for targeted cancer treatments, as well as the increased availability of high-throughput genetic data, has led to a surge in genetic analyses in recent years. Some of these genetic analyses focus on a specific genomic data platform, such as gene expression, protein expression, or DNA mutations, and look for the entities (genes, proteins, etc.) involved in cancer development and progression. For example, Welsh et al. (2001) analyzed the levels of

expression of about 9,000 genes in the context of prostate cancer. They first filtered the genes by intersample variability and then assigned each gene a score based on the difference between tumor and normal expression means, the ratio of tumor and normal expression means, and the results of a t-test comparing expression in the tumor and normal samples. This approach was chosen so that high scoring genes had large and consistent differences in expression between tumor and normal samples, making them prime candidates as potential therapy targets. Some of the top ranked genes included FASN (which codes a known tumor marker), PSA, MIC-1, and *hepsin*, and validation of differential expression for most of the top genes was included as part of the study (Welsh et al., 2001). In anther case, Bardelli et al. (2003) conducted a mutational analysis on 138 genes from three of the nine major groups of genes that code protein kinases. They identified seven recurrently mutated genes that appear to be involved in colorectal tumor growth, and they concluded that at least 30% of colorectal tumors have at least one mutation in the tyrosine kinome (Bardelli et al., 2003). Although this study is exploratory and its findings require further validation, the results are especially encouraging considering the success of the drug imatinib, a tyrosine kinase inhibitor which targets mutant kinases to treat chronic myeloid leukaemia and gastrointestinal stromal tumors (Sawyers, 2003, 2004).

Other investigative methods consider multiple genomic platforms; these are commonly called *integrative* methods and have been shown to have increased power and lower false discovery rates as compared to single platform analyses (Wang et al., 2013). For example Verhaak et al. (2010) analyzed DNA copy number, gene expression, and gene mutations in an integrative analysis of glioblastoma multiforme (GBM), a type of brain tumor. A cluster analysis on the gene expression of 1,740 genes identified four novel subtypes of GBM, which proceeded to validate on an independent data set. Signature genes associated with each subtype were identified

using ClaNC (Dabney, 2006), and the remaining data platforms were then included to find distinguishing genetic characteristics of each subype; the genes PDGFRA, IDH1, EGFR, and NF1 were found to be particularly involved. A significant difference in therapy response among the subypes indicated that patients may benefit by being treated differently depending on their classification (Verhaak et al., 2010). The genes found to be important could become potential targets for subtype-specific targeted therapies. Another type of integrative method includes the multiple data platforms in a single model. For example, the iBAG (integrative Bayesain Analysis of Genomics data) method originally proposed by Wang et al. (2013) and later adapted by Jennings et al. (2013) and McGuffey et al. (2015) integrates data platforms by modeling their known biological relationships. In a typical analysis, gene expression is partitioned into components explained by upstream platforms, and then those components are related to a clinical outcome. Important genes are flagged, but the integration technique also provides additional mechanistic information as to which data platforms are regulating the effects. The iBAG approach was also applied to a GBM dataset and identified several potential genetic targets.

Although finding individual genes that are involved in cancer outcomes is hugely beneficial, there is also value in identifying gene *pathways* involved in cancer development and progression, as well as the corresponding mechanistic information. The discovery of important pathways can facilitate the design of pathway-level medications. Such drugs have the potential to more easily treat a larger portion of cancer patients, considering that, even within the same cancer type, the specific genes involved can differ, and would in turn require different gene-targeted medications, while the differing genes are often members of the same pathway and have similar overal pathway effects (Vogelstein and Kinzler, 2004; National Human Genome Research Institute, 2015). A common way to identify significant pathways is to first

construct a list of genes found to be important to a cancer outcome and then conduct a separate pathway analysis to determine which pathway(s) are overactive. One popular pathway analysis is Ingenuity Pathway Analysis (IPA). IPA takes a gene list (obtained by the researcher) as input and assesses the association of the input gene list with certain pathways by considering the number of overlapping genes. Specifically, for each pathway of interest, IPA computes a one-sided Fisher's exact test of membership in the input gene list versus membership in the pathway and reports a p-value (or Benjamini-Hochberg corrected p-value) indicating whether the pathway is overexpressed in your gene list (Ingenuity Systems, 2015). One major factor in the p-value calculations is the reference gene set, which is the total number of genes considered for the Fisher's exact test. Ideally, one would use the initial list of genes included in his or her analysis. This can be manually input to IPA, or one can choose the reference set as the Ingenuity knowledge base or one of a few gene lists comprising common expression arrays. Regardless, the IPA manual warns against strict interpretations of the provided p-values (Ingenuity Systems, 2015). Thus, instead of using the p-values to select statistically significantly overexpressed pathways, they are typically used to rank the association of the input gene list with various pathways.

The method we will present in this paper provides, among other things, a pathway score for each pathway involved in the analysis, and these scores can then be used to rank the pathways by their relevance to the clinical outcome. The scores are estimated as parameters in the model, which provides two distinct advantages. First, there is no extra step to obtain pathway rankings; the gene and pathway level results are obtained simultaneously. Second, and more importantly, the pathway scores are related to the magnitudes of the effects within a pathway, as well as the number of important effects present in the pathway. Any approach that simply takes a gene

list as input is not accounting for the size of the effects. Also, our approach only considers the genes included in the model, so the scores are inherently based on the correct reference set.

In this paper, we present a pathway iBAG (piBAG) method that achieves three goals in a single model:

1. *Integration:* We integrate an arbitrary number of genomic platforms into a single model in a way based on their known biological relationships.

2. *Gene selection:* We flag the genes statistically and practically related to a clinical outcome, as well as provide mechanistic information as to which platform is regulating the effect. The mechanistic information is available because of the integration method.

3. *Pathway ranking:* We simultaneously assign each gene pathway a score for each platform, indicating the strength of the pathway's effect on the clinical outcome through that platform and providing a natural way to rank the pathways.

In Section 5.2 we present the pathway iBAG model and describe the estimation and selection procedures. We demonstrate the advantages of our method through a simulation study in Section 5.3, and we apply our method to a GBM data set from The Cancer Genome Atlas (TCGA) in Section 5.4. We conclude with a discussion of the method and its results in Section 5.5.

5.2 Methods

Our pathway iBAG model is a two-step hierarchical Bayesian model, based on the iBAG model first proposed by Wang et al. (2013) and later adapted by Jennings et al. (2013) and McGuffey et al. (2015). The first step, the *mechanistic model*, partitions gene expression into the components explained by upstream platforms by modeling

the biological relationships among the genetic platforms. The second step, the *clinical model*, then relates these partitioned components to a clinical outcome and identifies which genes are significantly involved and which platforms are modulating the effect of those genes.

5.2.1 Mechanistic model

The first step in the pathway iBAG model is the *mechanistic model*, and it is the step where we integrate the multiple platforms by regressing mRNA on upstream platforms known to affect gene expression. We use the nonlinear formulation of the original iBAG mechanistic model, proposed by McGuffey et al. (2015), which facilitates more flexible modeling of the potentially nonlinear relationship between mRNA and its upstream platforms. Evidence of such nonlinear relationships for one of the genes flagged in our data application, gene ATP2BI, is shown in Figure 5.1. The mechanistic model is fit independently for each gene, and it partitions each gene's expression into the components explained by each upstream platform and a component due to effects from other sources not included in the analysis.

Let there be $i = 1, ..., n$ samples, $p = 1, ..., P$ data platforms, $k = 1, ..., K$ gene pathways, and $g = 1, ..., G_k$ genes in pathway k, with $\mathcal{G} = \sum_{k=1}^{K} G_k$ being the total number of genes in the analysis. Note that an index of kg denotes a unique gene. (We only consider genes with a single pathway membership. Extensions to genes with multiple pathway memberships is discussed in Section 5.5.) Let mRNA_{kg} denote the $(n \times 1)$ vector of gene expression values for gene kg. Gene expression is typically summarized at the gene level, i.e., one value per gene, but other platforms may be summarized on the probe level, with multiple values per gene. If there are multiple values for gene kg on platform p, we perform a principal component analysis (PCA) and retain the J_{pkg} scores accounting for at least 90% of the variation; call them

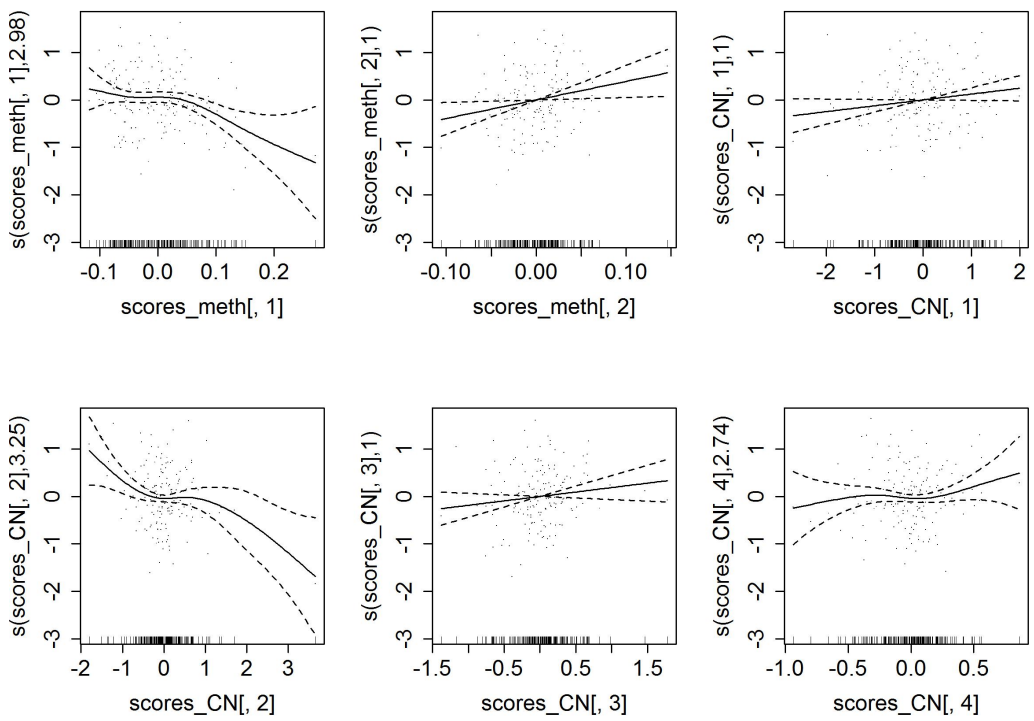


Figure 5.1: The mechanistic model fit for gene ATP2BI, which is later flagged in our data application, is plotted. There are four predictors: two PC scores for methylation and four for copy number. Raw predictor values are on the x-axes, the solid line is the predicted fit, and the dashed lines are the error bounds extending two standard deviations above and below the estimated curve. The partial residuals – the residuals that would arise by leaving out the predictor of interest and keeping the other estimates fixed – are also shown as points on each plot. Although some of the predictors appear to have a relatively linear relationship, some have a clear nonlinear pattern which is captured effectively by the penalized splines.

R_{pkg1}, R_{pkg2}, ..., $R_{pkgJ_{pkg}}$. Note that each PCA could result in a different number of scores being retained; hence we index J by pkg. If there is only one value for gene kg on platform p, denote its $(n \times 1)$ raw data vector R_{pkg1}, with $J_{pkg} = 1$. The following mechanistic model is fit independently for each gene:

$$\text{mRNA}_{kg} = \sum_{p=1}^{(p-1)} \sum_{j=1}^{J_{pkg}} f_{pkgj}(R_{pkgj}) + \phi_{kg} \tag{5.1}$$

where $f_{pkgj}(\cdot)$ is a penalized regression spline and $\phi_{kg} \sim \text{Normal}(0, \sigma^2_{\phi_{kg}})$. With

this formulation, each R vector corresponding to a particular gene is included as a predictor for that gene's expression, and the penalized regression splines allow us to obtain a close fit to the data without overfitting, by penalizing too much "wiggliness" in the fit.

After fitting the model, we use the estimates to calculate the partitioned components, specifically:

$$X_{pkg} = \sum_{j=1}^{J_{pkg}} \widehat{f_{pkgj}}(R_{pkgj}), \text{ for } p = 1, ..., P-1 \quad (5.2)$$

$$X_{Pkg} = \widehat{\psi_{kg}}. \quad (5.3)$$

This is repeated for each gene, and we obtain X_{pkg} for each platform ($p = 1, ..., P$) and each gene ($k = 1, ..., K$ and $g = 1, ..., G_k$). For $p = 1, ..., P-1$ we interpret X_{pkg} as the part of gene kg expression explained by platform p, and X_{Pkg} is the part of gene kg expression explained by something other than the platforms included in the analysis.

By integrating the data platforms in a way that accounts for the biological relationships among them, we are (1) harnessing the available biological information to de-noise the gene expression data and carry forward clearer signals, and (2) providing more precise biological interpretations of the effects significantly related to the clinical outcome. For example, if a certain gene's component explained by methylation is then flagged as important in the clinical model, we know not only that the gene has a significant effect, but that methylation is modulating that gene's effect. As such, an understanding of the biological relationships among the data platforms is necessary to appropriately construct the mechanistic model. For example, DNA methylation and DNA copy number are known to affect gene expression, so if these are the platforms in a data set we would regress mRNA on methylation and copy number. Other platforms known to affect gene expression, such as microRNAs and

mutations, could also be included as predictors of mRNA. However, if we had another platform that was not upstream of gene expression, say protein expression, the mechanistic model can still accommodate this. Gene expression can first be regressed on its upstream platforms, and then those partitioned pieces can become the predictors for protein expression. So a final partitioned piece might be interpreted as the component of protein expression explained by the part of gene X's expression that is regulated by methylation. (If such an effect were then flagged as having a negative effect on survival time, it would be potentially useful to study altering the methylation of gene X so that less of the protein would be expressed.)

After fitting the mechanistic model, we subsequently carry forward each of the partitioned pieces (the X_{pkg}s) to the second step in the piBAG model.

5.2.2 Clinical model

Our model's second step is the *clinical model*, and it is here that we relate the clinical outcome to the partitioned pieces estimated from the mechanistic model. Let Y be the $(n \times 1)$ vector of mean-centered continuous clinical outcome values, and let $X_p = \{X_{p11}, ..., X_{p1G_1}, X_{p21}, ..., X_{p2G_2}, ..., X_{pK1}, ..., X_{pKG_K}\}$ for $p = 1, ..., P$. In other words, we form a matrix for each platform whose columns are all the partitioned pieces corresponding to that platform, one column per gene, so that each X_p has dimension $(n \times \mathcal{G})$. Call $X = \{X_1, X_2, ..., X_P\}$, and then the clinical model is as follows:

$$\begin{aligned} Y &= X_1\boldsymbol{\beta}_1 + X_2\boldsymbol{\beta}_2 + ... + X_P\boldsymbol{\beta}_P + V \quad (5.4) \\ &= X\boldsymbol{\beta} + V \quad (5.5) \end{aligned}$$

where each $\boldsymbol{\beta}_p$ is the $(\mathcal{G} \times 1)$ coefficient vector, $\boldsymbol{\beta} = \{\boldsymbol{\beta}_1^{\mathrm{T}}, \boldsymbol{\beta}_2^{\mathrm{T}}, ..., \boldsymbol{\beta}_P^{\mathrm{T}}\}^{\mathrm{T}}$, and $V \sim \text{Normal}(0, \tau^2)$ is the error term.

Our primary objective is efficient estimation of the $\boldsymbol{\beta}$ effects. In this genomic

setting, it is common to encounter the situation where the number of predictors is greater than the number of samples, and we also believe that the solution will be sparse, with most of the genes not strongly related to the clinical outcome. Thus, an ordinary least squares estimate would not be appropriate: if there are more predictors than samples, the solution would not be unique, and if there are more samples than predictors, least squares would overfit at best. A popular estimation choice for scenarios similar to ours is Tibshirani's lasso (Tibshirani, 1996), a penalized regression approach that sets some coefficients to exactly zero and provides estimates of the non-zero effects. The frequentist lasso has been shown to perform well is many situations, but it does have two drawbacks worth noting. First, the lasso can select at most n non-zero predictors. This is not a major concern for our particular setting, since we already expect a sparse solution. Second, and more importantly, there is not a way to obtain standard error estimates, and thus confidence statements, for the effects set to zero (Kyung et al., 2010).

A natural solution is to implement the Bayesian lasso (Park and Casella, 2008) which provides standard errors for all estimates through Monte Carlo averages of posterior samples. In fact, the Bayesian lasso was used in Wang et al.'s original proposal of iBAG model (Wang et al., 2013). While the shrinkage induced by the Bayesian lasso does promote the desired sparsity, putting a normal-gamma (NG) prior on the $\boldsymbol{\beta}$ effects, as opposed to the Laplace prior of the Bayesain lasso, provides increased flexibility in the shrinkage and improved efficiency in estimation (Griffin and Brown, 2010). Ideally, the truly smaller effects would be shrunk even closer to zero, while the truly larger effects would be shrunk less, or not at all, and remain essentially intact. The two hyperparameters of the NG prior, as opposed to the single hyperparameter of the Laplace prior, provide the increased flexibility necessary to achieve that goal. The NG prior has been used in recent adaptations of iBAG

(Jennings et al., 2013; McGuffey et al., 2015), and we will use it in the pathway iBAG model as well.

The complete hierarchy for the pathway iBAG clinical model is shown below:

$$Y \sim \text{Normal}(X\boldsymbol{\beta}, \tau^2\mathbf{I}_n) \tag{5.6}$$

$$\tau^2 \sim \text{InvGamma}(a, b) \tag{5.7}$$

$$\beta \sim \text{Normal}(\mathbf{0}, D_{\sigma^2}) \text{ where } D_{\sigma^2} = \tag{5.8}$$

$$\text{diag}(\sigma^2_{111}, ..., \sigma^2_{11G_1}, \sigma^2_{121}, ..., \sigma^2_{12G_2}, ..., \sigma^2_{1K1}, ..., \sigma^2_{1KG_K}, ..., \sigma^2_{P11}, ..., \sigma^2_{PKG_K})$$

$$\text{(i.e., } \beta_{pkg} \sim \text{Normal}(0, \sigma^2_{pkg})) \tag{5.9}$$

$$\sigma^2_{pkg} \sim \text{Gamma}(\alpha, 1/(2\xi^2_{pk})) \tag{5.10}$$

$$\xi^{-2}_{pk} \sim \text{Gamma}(\tilde{a}, \tilde{b}/(2\lambda)) \tag{5.11}$$

$$\alpha \sim \text{Exp}(\tilde{c}) \tag{5.12}$$

$$\lambda \sim \text{Exp}(\tilde{d}). \tag{5.13}$$

In the normal component of the NG priors assigned to the βs, each β_{pkg} has its own variance parameter (see Equation 5.9). The larger the σ^2_{pkg} is estimated to be, the larger the β_{pkg} is "allowed" to be, so generally larger σ^2 values correspond to a larger β magnitude. The ξ^2 parameter is indexed by platform and pathway (pk) only, so the σ^2 values associated with the effects of pathway k through platform p are shrunk to a common mean (see Equation 5.10). This translates to the corresponding effect estimates being shrunk toward a similar magnitude, i.e., a β will be shrunk more if many other βs in the same pathway (acting through the same platform) are also small. Vice versa, a β will be shrunk less if its fellow pathway βs are also large. Hence, we are borrowing strength within the pathways, which leads to improved efficiency of estimation. There is a unique ξ^2 parameter for each pathway/platform combination, and they are all shrunk toward a single common mean (see Equation 5.11). Simlarly to the relationship between the βs and σ^2s, a larger ξ^2 value allows

larger σ^2 values for that pk index, which in turn allows larger β effect sizes for that pk index. This facilitates the use of the ξ_{pk}^2 estimate (or equivalently, ξ_{pk}^{-2}, as written above) as a score for the relative size of the effect that pathway k has on the clinical outcome through platform p. If the clinical outcome is survival time, this can even be interpreted as a prognostic pathway score.

All the complete conditional distributions are in closed form except for that of λ, so a Gibbs sampler can be implemented with a single Metropolis-Hastings update step. Note that it is *very* important to standardize each X column before doing estimation. Otherwise, the priors on the βs will not be on the correct scale. The complete conditional distributions and our choices for hyperparameters and initial values can be found in Appendix C.

5.2.3 Selection and summary

To select important genes, any appropriate selection procedure based on posterior samples can be implemented, but here we present two options: one based on the median probability model and one based on controlling the false discovery rate (FDR). The choice of the selection method can be determined by the user's objective.

The first option, based on the median probability model (Barbieri and Berger, 2004), proceeds as follows.

1. Choose a minimum effect size that is of practical interest, and define (δ_-, δ_+) as the region of β effect sizes you do not want to flag as important. Depending on the nature of the clinical response, this region may not be symmetric.

2. Let S be the number of MCMC samples and $\beta_{pkg}^{(s)}$ be the posterior sample of β_{pkg} from iteration s. For each β_{pkg}, calculate the posterior probability that β_{pkg} is greater than δ_+ and the posterior probability that β_{pkg} is less than δ_-. In particular, calculate $p_+(\beta_{pkg}) = \sum_{s=1}^{S} \mathbf{I}(\beta_{pkg}^{(s)} > \delta_+)/S$ and $p_-(\beta_{pkg}) =$

$\sum_{s=1}^{S} \mathbf{I}(\beta_{pkg}^{(s)} < \delta_-)/S$.

3. Flag gene kg as significant through platform p if $p_+(\beta_{pkg}) > 0.5$ or $p_-(\beta_{pkg}) > 0.5$.

The second option is a Bayesian adaptation of Benjamini and Hochberg's procedure (Benjamini and Hochberg, 1995) presented by Muller et al. (2006). It controls the average local FDR and is implemented as follows.

1. Define α_{FDR} as the upper bound for the average local FDR.
2. Sort the posterior probabilities $p_+(\cdot)$ in descending order.
3. Compute the cumulative average of these sorted probabilities.
4. Flag the effects with cumulative average greater than $1 - \alpha_{FDR}$.
5. Repeat steps 2-4 for posterior probabilities $p_-(\cdot)$.

To identify the pathways related to the clinical response, we focus on the ξ_{pk}^2 parameters in the clinical model. After much investigation, we believe that the posterior estimates of these parameters are most informative when considered as a pathway score and used to rank the pathways by prognostic relevance, as opposed to making a binary selection decision. Also, ξ_{pk}^2 summarizes pathway k's effect *through* platform p, so a separate ranking of pathway importance can be obtained within each platform. As seen in the simulation and data application, these rankings do differ across platforms; this information is helpful in understanding which platforms are the mechanisms for each pathway effect.

5.3 Simulation

To assess the performance of the clinical model, we simulate a data set with realistic settings, described in Section 5.3.1, apply several models, and compare the results. We will see that the pathway iBAG has the best performance.

5.3.1 Settings

We begin by simulating each value in the predictor matrix X from a standard normal distribution. We use $P = 2$ platforms, $K = 6$ pathways, $G_1 = 5$ genes in pathway 1, $G_2 = 10$ genes in pathway 2, $G_3 = 20$ genes in pathway 3, $G_4 = 50$ genes in pathway 4, $G_5 = 85$ genes in pathway 5, and $G_6 = 130$ genes in pathway 6. This results in $\mathcal{G} = 300$ genes in total, so we have $\mathcal{G} \times P = 600$ predictors in the clinical model. We restrict the number of samples simulated to $n = 500$, so that we are assessing performance in the setting of more predictors than samples. Also, we choose the number of genes in each pathway to vary dramatically because that reflects what can be seen in the biological pathways. We will show that the estimates of pathway importance are not solely based on pathway size, but instead are based on the strength of their gene effects.

After simulating the (500×600) predictor matrix, we set the values for the true β_{pkg} coefficients. There are $P \times K = 12$ platform/pathway combinations, and 6 are chosen to be important while 6 are chosen as unimportant. In particular, pathway 6 is assigned as important through platform 2 but unimportant through pathway 1. To set the true β values, we first sample $G_6 = 130$ values from a Laplace$(\mu = 0, b = 2)$ distribution, and another 130 values from a mixture of 50 zeros and 80 values from a Laplace$(\mu = 0, b = 1/5)$ distribution. The Laplace distribution with scale parameter 2 is much less peaked at its mean (0), and it has much thicker tails, resulting in larger values being sampled overall, but still some smaller effects present. The βs for the

important platform/pathway combinations are then chosen as subsets of these 130 values. On the other hand, the Laplace distribution with scale parameter 1/5 is very peaked at 0 and has thin tails, resulting in primarily small, but not identically 0, values being sampled from it. We assign the βs in the unimportant platform/pathway combinations as subsets of the mixture distribution, so that we simulate the existence of some genes not related to the clinical outcome at all as well as (most) genes being only slightly related. The βs are nested to allow for direct comparison of shrinkage properties in pathways of varying size.

Once obtaining the true β values, we set $\tau^2 = 4$ and simulate the clinical response vector $Y = X\boldsymbol{\beta} + \text{Normal}(0, \tau^2)$. We then supply the simulated X matrix and Y vector to the pathway iBAG model, and perform 5500 iterations of the Gibbs sampler, 500 of which are used as burn-in. Note that the only parameters we set are the β effects and overall variance τ^2, and we do not simulate directly from our hierarchy of priors. Thus, we do not intend to give our method undue advantage through the simulation settings; instead, we hope to simulate under settings likely to be encountered in real data sets.

In addition to applying the pathway iBAG model presented in this paper, we also apply 3 other variations. First, we consider the pathway iBAG model *without* any pathway information, that is, we set the number of pathways to one and consider no pathway distinctions among genes. Second, we include pathway membership information, but do not do integration of multiple data platforms. To achieve this, we use the previously simulated X matrix and sum the two platform components to recover the unpartitioned mRNA value for each gene. Then we apply our method using each gene's single recovered value as the predictors, as opposed to the partitioned components. For this setting, we do not know the true effect values, but we do know which genes should be flagged – we base that on whether either of the gene's two

partitioned components show significance through that platform. The last variation we apply also uses the recovered mRNA values (no integration) as predictors and in addition does not include gene pathway membership information.

5.3.2 Results

Table 5.1 shows performance assessment summaries for each of the four model variants applied to the simulated data set. Overall, we see the pathway iBAG model (with integration and using pathway information) is the superior model, indicated by higher credible interval coverage, tighter credible intervals, lower mean squared error, higher sensitivity and specificity, and lower false discovery rates and false negative rates (FNRs). We discuss the results in more detail below.

For this simulation study, we call the effect "important" if $|\beta_{pkg}| > 0.5$. This designation is used when assessing CI widths for important and unimportant βs and also for the selection procedures. Recall that the non-integrative methods use the recovered mRNA values as predictors, and we do not set those corresponding β values explicitly. Thus, we cannot compare some of the assessments directly for the integrative versus non-integrative methods. This is why the coverage probabilities and the CI widths when separated by β importance have values of "NA" for the non-integrative methods – we do not know those true β values. As can be seen in Table 5.1, the integrative pathway model has the highest CI coverage and the lowest average CI widths, with the integrative non-pathway model having the second smallest CI widths, and the non-integrative methods coming in last with very similar widths. As expected, borrowing strength within the platform/pathway combinations leads to tighter bands for the estimates. Also, as an expected effect of the induced shrinkage, we see the CI widths for the unimportant βs are smaller than for the important βs in the integrative methods. The mean squared error (MSE) is the predictive error based

Table 5.1: Pathway iBAG simulation results. "Avg." abbreviates average, and "imp." abbreviates important. The β CIs are pointwise 90% credible intervals, and MSE is predictive mean squared error. FDR is false discovery rate, and FNR is false negative rate. Selection option 1 is based on the median probability model, and selection option 2 controls average local FDR. Both are described in Section 5.2.3.

Method	**PI**	**I**	**P**	**neither**
Integration	Yes	Yes	No	No
Pathway Information	Yes	No	Yes	No
β CI coverage	0.9333	0.8900	NA	NA
Avg. β CI width	0.7198	0.8446	4.7695	4.8035
Avg. imp. β CI width	0.8259	0.9569	NA	NA
Avg. unimp. β CI width	0.6375	0.7576	NA	NA
MSE	30.18	50.21	2138.76	2154.70
Selection option 1:				
Gene/platform sensitivity	0.939	0.901	NA	NA
Gene/platform specificity	0.973	0.92	NA	NA
Gene/platform FDR	0.035	0.103	NA	NA
Gene/platform FNR	0.061	0.099	NA	NA
Gene sensitivity	0.976	0.943	0.633	0.649
Gene specificity	0.885	0.891	0.582	0.600
Gene FDR	0.032	0.025	0.129	0.122
Gene FNR	0.024	0.057	0.367	0.351
Selection option 2:				
Gene/platform sensitivity	0.962	0.924	NA	NA
Gene/platform specificity	0.902	0.867	NA	NA
Gene/platform FDR	0.116	0.157	NA	NA
Gene/platform FNR	0.038	0.076	NA	NA
Gene sensitivity	0.996	0.963	0.200	0.208
Gene specificity	0.655	0.745	0.982	0.982
Gene FDR	0.072	0.056	0.020	0.019
Gene FNR	0.004	0.037	0.800	0.792

on a test set generated with the same fixed settings as discussed in Section 5.3.1. Our primary objective for the pathway iBAG is not prediction; rather, we aim to select genes, assign pathway scores, and provide the mechanistic information regarding the platforms modulating the effects. Thus we include MSE as another way to assess

performance, but it does not carry as much weight as the other results. In any case, we do see that the integrative pathway method provides the best prediction, with the integrative non-pathway method coming in second with an approximate two thirds increase in MSE. Again, the non-integrative methods come in last, and again their results are quite similar to each other. From the assessments discussed so far, we see evidence that the integrative pathway model facilitates the most efficient estimation.

Selection options 1 and 2 are described in Section 5.2.3. We set our practical region of importance as $(\delta_-, \delta_+) = (-0.5, 0.5)$, and for option 2 we set $\alpha_{FDR} = 0.1$. For the integrative methods, we know the true β_{pkg} values, and we can consider whether each effect (one per gene/platform combination) should be flagged and whether it was flagged. For both selection options, the integrative pathway model exhibits increased sensitivity and specificity and decreased FDR and FNR, as compared to the integrative non-pathway method. To calculate these quantities at the gene level and allow a comparison to the non-integrative methods, we consider the true β_{pkg} values used in the integrative methods, and each gene is designated as "should be flagged" if either of its corresponding gene/platform effects is important. Looking at the gene-level assessment, we see the integrative methods performing similarly, with the pathway model getting slightly better sensitivity and FNR scores and the non-pathway model getting slightly better specificity and FDR scores. The non-integrative methods also perform similarly to each other, with a very slight advantage going to the non-pathway version. From the gene level comparisons we can conclude that the integrative versions provide a clear advantage as far as power, considering the dismal sensitivity rates and false negative rates of the non-integrative methods. Then looking at the gene/platform level comparisons, we see that the integrative pathway model outperforms the non-pathway version for both selection options.

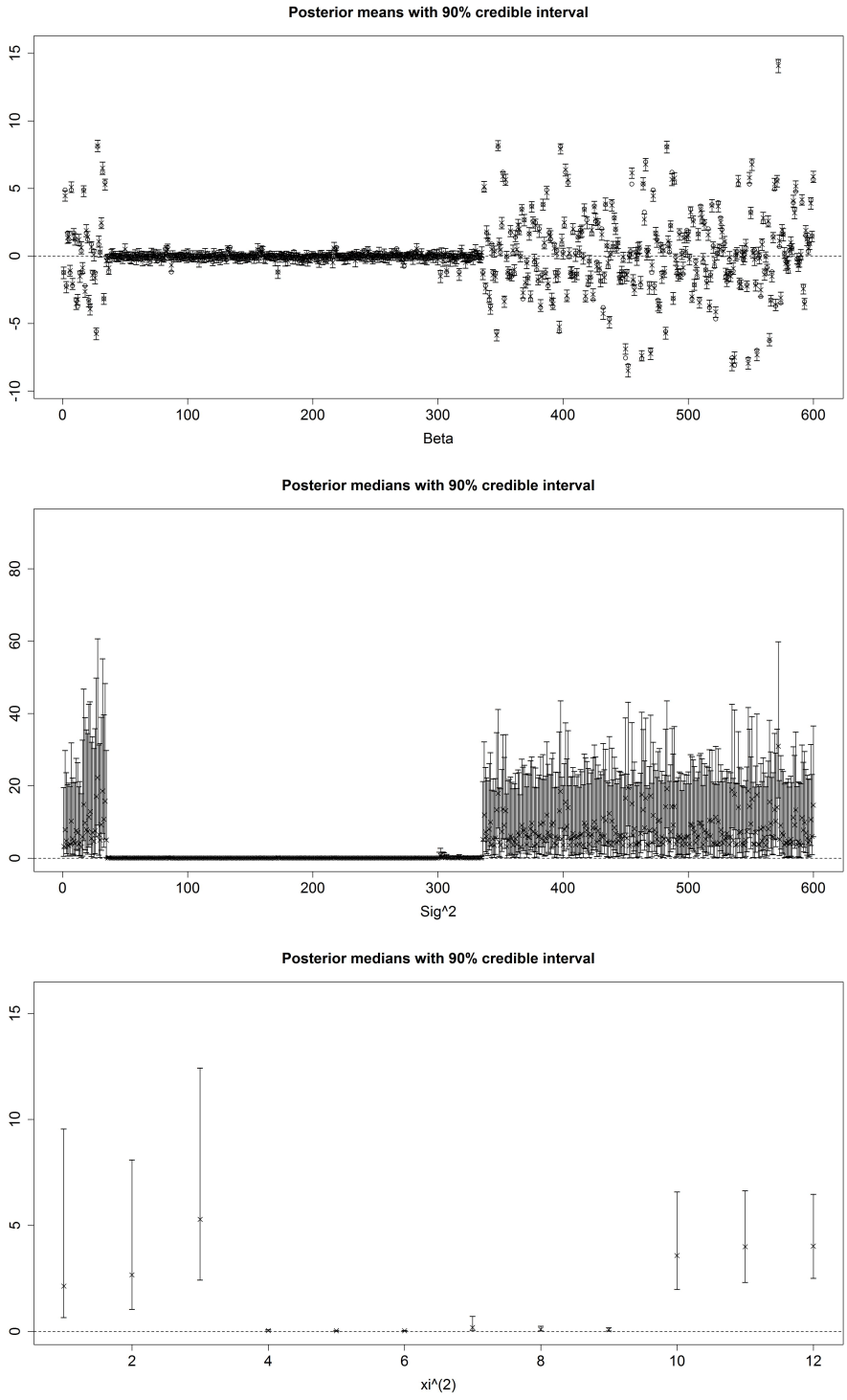


Figure 5.2: Pointwise credible intervals for β, σ^2, and ξ^2 parameters in the piBAG model. The 90% credible bands are shown, and in the β plot, "x" indicates the posterior mean and "o" marks the true value.

Figure 5.2 plots the 90% pointwise credible intervals constructed from posterior samples of the piBAG model (i.e., with integration and pathway information) for the β, σ^2, and ξ^2 parameters. These plots illustrate the important property discussed in Section 5.2.2: a larger β_{pkg} magnitude corresponds to a larger σ_{pkg} value, which in turn corresponds to a larger ξ^2_{pk} value.

The pathway score estimates are the posterior medians of the ξ^2 parameters. We can rank these to find which pathways are most related to the clinical outcome and through which platform the pathway is taking effect. The setting where such a comparison is most informative is the integrative pathway model, and Figure 5.3 plots the scores with the important platform/pathway combinations in blue. (The x and y axes are identical, that is, the ξ^2_{pk} estimate is both the x and y coordinate.) As the plot illustrates, piBAG correctly distinguishes between the important and unimportant platform/pathway combinations. We can also sort these scores within platform; for example, it is clear that through platform 1, pathway 3 is the most important, followed by pathway 2 and then pathway 1. This is consistent with our simulation settings, considering that all three of those pathways were simulated as important through platform 1, and pathway 3 had significantly more genes that pathways 1 and 2. In the integrative version without pathway information, the ξ^2 scores cannot say anything about pathways, but they can still rank the platforms. In our simulation, the scores from the integrative version without pathway information rank platform 2 ($\widehat{\xi^2_{21}} = 16.64$) as more important than platform 1 ($\widehat{\xi^2_{11}} = 0.93$). Although we did not explicitly set platform importance, this seems reasonable considering platform 2 had the largest three pathways as important, and platform 1 had the smallest three as important. Finally, we can consider the ξ^2 scores from the non-integrative version that still included pathway membership information. This method provides a single score per pathway, which should indicate which pathways contribute more

to the clinical outcome. Again, since we did not explicitly assign each pathway an importance measure, we cannot say if the estimated ranking is precisely correct. However, the highest score is for pathway 3, which is found by the integrative pathway method to be the most related pathway (through platform 1). The consistency goes further: although we do not mean to suggest the scores have any formal additivity property, if we sum each pathway's two scores obtained by the integrative pathway method and then rank the pathways based on those, we obtain the exact same ranking as produced by the non-integrative pathway method.

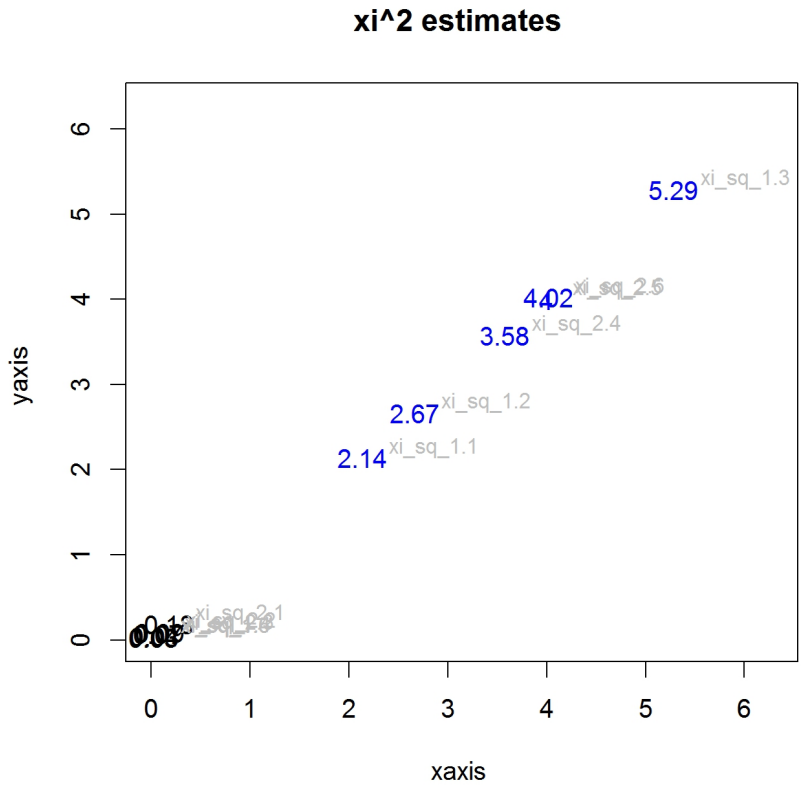


Figure 5.3: Plotted estimates of ξ^2 parameters. A large ξ^2_{pk} is interpreted as pathway k having an important clinical effect through platform p. Truly important platform/pathway combinations are plotted in blue, and each value is labeled in grey. The x-axis and y-axis coordinates are both the estimated ξ^2 value.

Through the simulation study, we demonstrate the advantages of the full piBAG model, particularly the increased efficiency in estimation and power in selection. The pathway scores also provide valuable mechanistic information, specifically a ranking of which pathways have the strongest effect on the clinical outcome and through which platform(s) those effects are modulated. It is also worth noting that including pathway information can be beneficial even in an analysis with a single predictive data platform; from Table 5.1 we saw that the gene selection is quite similar for the non-integrative methods, regardless of including pathway information, but by including the pathway memberships, we are able to procure additional mechanistic information as to which pathways are most clinically relevant.

5.4 Data Application

We apply our pathway iBAG method to a glioblastoma multiforme (GBM) data set made publicly available by The Cancer Genome Atlas (TCGA). TCGA is a project started in 2006 with the goal of compiling comprehensive multiplatform data sets for many different cancer types (The Cancer Genome Atlas, 2012). GBM is a very deadly brain tumor, and it was one of the first types of cancer to be studied by the TCGA. We focus our analysis on a subset of the available data, as described below.

5.4.1 Data description

The genomic data platforms we include in our analysis are mRNA expression, DNA methylation, and DNA copy number. The mRNA data is level 3 data from the Affymetrix profiled HT_HG_U133A platform and is summarized at the gene level. Our methylation data is level 3 data from Human Methylation27K arrays and is summarized at the probe level; it quantifies DNA methylation at multiple sites per gene. The copy number data is level 2 data from the HG_CGH_244A platform and

is also on the probe level; it is the normalized signal for copy number alterations of aggregated regions per probe, and we use the $\log_2$ ratios of matched normal-tumor samples to quantify copy number changes (The Cancer Genome Atlas Data Portal, 2013). The clinical response we consider is uncensored survival time, measured in days from diagnosis, for 163 patients. Since we are modeling survival times, we implement an accelerated failure time model by using log(survival) as the Y vector in the clinical model (Equation 5.4) (Wei, 1992). We include 157 genes from 10 signaling pathways, each with a unique pathway membership, and the number of genes per pathway ranges from 9 to 37.

We impute the few missing values in the methylation and copy number data subsets and then apply piBAG using 10,500 MCMC iterations, 500 of which are used as a burn-in period. Our primary interest is the full piBAG model, but we also run the method without gene pathway information (achieved by setting $K = 1$ pathway, as in the simulation study) for comparison.

5.4.2 Results

After fitting the model, we have two types of results: gene selection and pathway scores. For flagging important genes, we choose to apply selection option 1, the method based on the median probability model. We set our practical importance threshold to correspond to a change in survival time of at least 5%, which translates to $(\delta_-, \delta_+) = (\log(0.95), \log(1.05))$, and apply our selection procedure. Posterior probabilities that $\beta_{pkg} > \delta_+$ and that $\beta_{pkg} < \delta_-$ are plotted in Figures 5.4 and 5.5, respectively. The gene/platform combinations with either probability greater than 0.5 are flagged as important.

Four genes are selected as clinically important. The genes FABP2 and CCNG1 are found to have positive effects on survival time through a platform other than

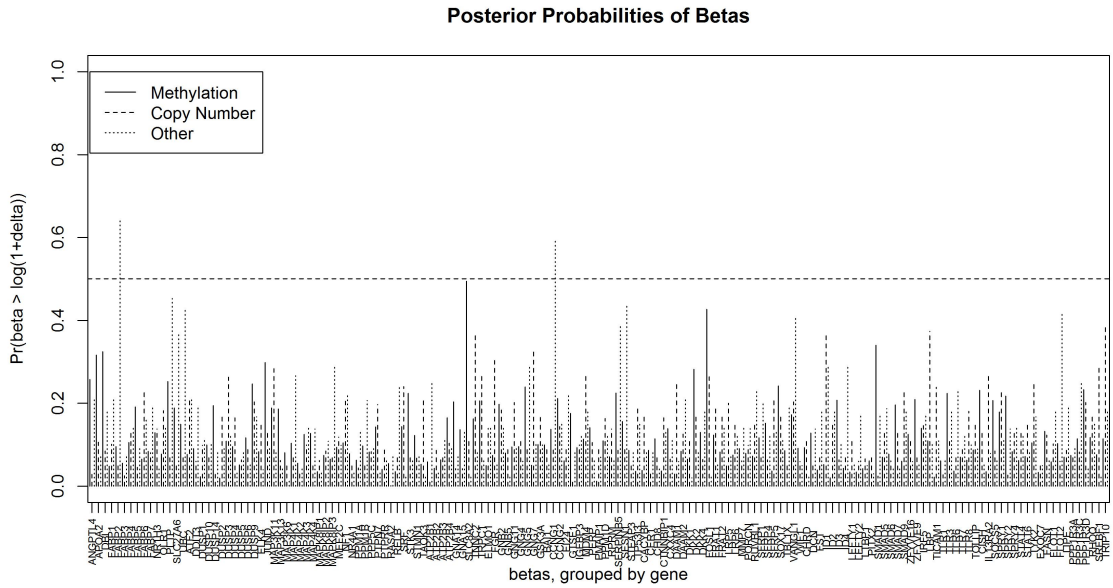


Figure 5.4: GBM application results: posterior probabilites from piBAG method that $\beta_{pkg} > \delta_+$. The dashed line is at probability 0.5.

methylation or copy number, meaning more expression explained by "other" effects is associated with longer survival times. Genes RASA2 and ATP2B1 are found to have negative effects on survival time, with both genes' effects regulated by copy number changes; more gene expression explained by copy number is related to a worse prognosis. Although survival data can be quite noisy, our piBAG model has still found several potential prognostic genes, and based on the low error rates in the simulation study, we can be confident that these discoveries are real signal. The posterior means for all the β_{pkg}s are shown in Figure 5.6, and the means for the flagged effects are designated by a solid dot.

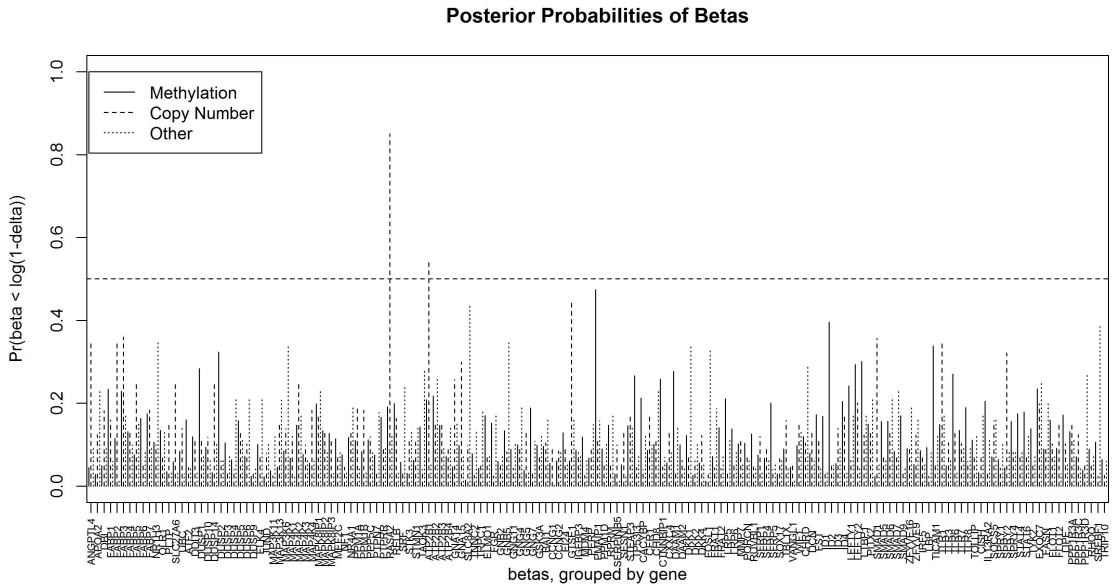


Figure 5.5: GBM application results: posterior probabilites from piBAG method that $\beta_{pkg} < \delta_{-}$. The dashed line is at probability 0.5.

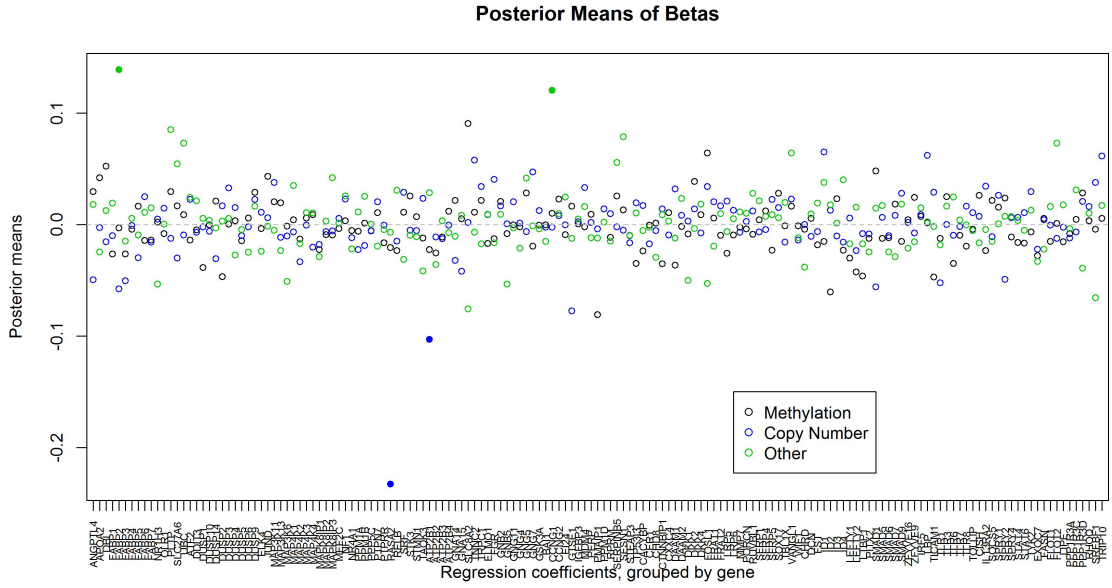


Figure 5.6: GBM application results: β_{pkg} posterior means from piBAG method. The solid dots represent the effects found to have a significant effect on survival time.

To rank the pathways, we summarize the pathway scores in Tables 5.2, 5.3, and 5.4. Each table summarizes the pathway scores within a different regulating platform - methylation, copy number, or something other than methylation and copy number, respectively. The PPAR signaling pathway is found to be most clinically important on all platforms, but other than that the rankings change from platform to platform. As far as the genes flagged as important, FABP2 is in the PPAR signaling pathway, CCNG1 is in the P53 signaling pathway, RASA2 is in the MAPK signaling pathway, and ATP2B1 is in the calcium signaling pathway. Three of these pathways (PPAR, P53, and calcium) are ranked as highly important through the platform corresponding to their flagged gene, but the MAPK signaling pathway has the second lowest ranking through copy number, the platform on which its gene is flagged. This is further evidence that although it is harder, it is certainly not impossible for a gene in an unimportant (or less important) pathway to be flagged on its own strong effect.

Table 5.2: Pathway rankings, from high to low, within the methylation platform for the GBM data application. The pathway score is the ξ_{pk}^2 estimate, that is, the posterior median. A larger score indicates a stronger effect from that pathway on the clinical outcome.

	Signaling Pathway	Score	Number of Genes
ξ_{11}^2	PPAR	0.0365	15
ξ_{13}^2	Calcium	0.0259	9
ξ_{19}^2	JAK-STAT	0.0237	9
ξ_{18}^2	Toll-like receptor	0.0235	9
ξ_{14}^2	Chemokine	0.0218	9
ξ_{17}^2	TGF beta	0.0215	18
ξ_{15}^2	P53	0.0204	15
$\xi_{1,10}^2$	Insulin	0.0200	11
ξ_{16}^2	WNT	0.0165	25
ξ_{12}^2	MAPK	0.0138	37

Table 5.3: Pathway rankings, from high to low, within the copy number platform for the GBM data application. The pathway score is the ξ^2_{pk} estimate, that is, the posterior median. A larger score indicates a stronger effect from that pathway on the clinical outcome. The bold pathways contain a flagged effect.

	Signaling Pathway	Score	Number of Genes
ξ^2_{21}	PPAR	0.0418	15
ξ^2_{23}	**Calcium**	0.0295	9
ξ^2_{28}	Toll-like receptor	0.0251	9
ξ^2_{29}	JAK-STAT	0.0243	9
ξ^2_{24}	Chemokine	0.0234	9
$\xi^2_{2,10}$	Insulin	0.0225	11
ξ^2_{25}	P53	0.0191	15
ξ^2_{27}	TGF beta	0.0185	18
ξ^2_{22}	**MAPK**	0.0174	37
ξ^2_{26}	WNT	0.0152	25

Table 5.4: Pathway rankings, from high to low, for a regulating platform other than methylation or copy number for the GBM data application. The pathway score is the ξ^2_{pk} estimate, that is, the posterior median. A larger score indicates a stronger effect from that pathway on the clinical outcome. The bold pathways contain a flagged effect.

	Signaling Pathway	Score	Number of Genes
ξ^2_{31}	**PPAR**	0.0589	15
$\xi^2_{3,10}$	Insulin	0.0270	11
ξ^2_{35}	**P53**	0.0251	15
ξ^2_{33}	Calcium	0.0250	9
ξ^2_{34}	Chemokine	0.0240	9
ξ^2_{38}	Toll-like receptor	0.0218	9
ξ^2_{39}	JAK-STAT	0.0216	9
ξ^2_{37}	TGF beta	0.0194	18
ξ^2_{36}	WNT	0.0173	25
ξ^2_{32}	MAPK	0.0149	37

When we apply the method without incorporating pathway information, only two

genes are flagged; increased CCNG1 expression, explained by something other than methylation or copy number, is found to have a negative effect on survival time, and increased RASA2 expression, explained by copy number, is found to have positive effects on survival time. Both of these genes were flagged through these platforms in the full pathway iBAG, but the full piBAG also flagged two other genes, as discussed above. The higher power of the full pathway iBAG is consistent with the tighter credible interval bands of the full pathway model.

5.5 Discussion

We have presented the pathway iBAG model, or piBAG, a hierarchical two-step Baysian model that integrates multiple genomic data platforms by modeling their biological relationships and identifies clinically important genes and pathways. The flagged genes are not only found to be important, but we also provide the mechanistic information of which data platform is driving the important effects. The pathways are ranked by clinical relevance within each platform via a pathway score, estimated as a parameter in the model. The biological insight provided by piBAG is critical to the development and improvement of targeted cancer treatments.

The model design facilitates highly flexible shrinkage and the borrowing of strength among platform/pathway combinations, both of which contribute to efficient parameter estimation and an increase in power as compared to non-integrative and non-pathway models. We have demonstrated these advantages through a simulation study. The simulation results present a strong case for data integration, but even without multiplatform data, we showed that incorporating pathway membership information leads to a gain of mechanistic information without a substantial loss of selection power. We also analyzed a TCGA GBM data set and identified four potential prognostic biomarkers, two with a positive effect on survival and two with

a negative effect. The pathway rankings obtained from the GBM application also provide insight as to the more heavily involved gene pathways. In particular, we found the PPAR signaling pathway to be the most important, as related to survival time, through every platform. Additional evidence of the improved power offered by piBAG was provided through a comparison of the GBM application without using pathway information; this method only found two of the four genes flagged by the full pathway iBAG model.

There are two natural extensions to the piBAG model presented here. First, the inclusion of other genomic data platforms is certainly of interest. We discussed how to include protein expression data in section 5.2.1. Another popular platform is microRNA (miRNA). MicroRNAs are known to affect gene expression, and with a list of the miRNAs known to target a particular gene, miRNA data can be included as another predictor, or as multiple predictors, in the mechanistic model. Then, depending on the researcher's purpose, the miRNA effect could be carried forward as a single partitioned piece, or as multiple pieces. The second extension is the inclusion of genes with multiple pathway memberships. We have not implemented this yet, but we propose introducing the parameter ψ_{kg}, which takes an integer value to indicate which pathway membership should be considered for each MCMC iteration. There should only be one ψ parameter per gene, so if, for example, a certain gene could be gene 5 in pathway 1 or gene 8 in pathway 3, we would require $\psi_{15} = \psi_{38}$. Then we could replace Equations 5.8 and 5.9 with

$$\beta_{pkg}|\psi_{kg} = k \sim \text{Normal}(0, \sigma^2_{pkg}) \tag{5.14}$$

$$\psi_{kg} \sim \text{Multinomial}(\pi_{kg}) \tag{5.15}$$

$$\pi_{kg} \sim \text{Dirichlet}(\mathbf{a}_{kg}) \tag{5.16}$$

where the length of $\mathbf{a}_{kg}$ is the number of potential pathways for that gene, and

its entries are either all ones or weighted by the number of genes in the pathway. This formulation would allow a gene to have a unique membership for each MCMC iteration, but jump between its possible pathways based on the sampled parameter.

6. INTEGRATIVE HEATMAPS

6.1 Introduction

Data sets with measurements on multiple genomic platforms, such as messenger RNA (mRNA) expression and DNA methylation, for each patient are becoming more widely available. It has been shown that integrating these multiple data platforms into a single analysis provides the advantages of increased power, lower false discovery rates, and more in-depth biological understanding (Wang et al., 2013; McGuffey et al., 2015). One such way of integrating these genomic data platforms is to regress mRNA expression on upstream platforms that are known to affect mRNA expression, essentially partitioning the mRNA expression into the components explained by the various upstream platforms (McGuffey et al., 2015).

A concise, effective way to illustrate the integration of the platforms is necessary to provide a better grasp of the integration results and to understand the underlying structure of the platform components. To achieve these goals, we present integrative heatmaps (IHs), a novel visualization tool for integrated genomic data.

6.2 Methods

First we briefly present the integration method referenced in Section 6.1, in the context of the data platforms to be used in our application. Say we have raw data values, summarized at the gene level, for mRNA expression, DNA methylation, and DNA copy number. Let $mRNA_g$, $R_g^{(M)}$, and $R_g^{(C)}$ denote the vectors of mRNA, methylation, and copy number values for gene g, respectively. Then to integrate these platforms, we fit the following model independently for each gene:

$$mRNA_g = b_0 + f_g^{(M)}(R_g^{(M)}) + f_g^{(C)}(R_g^{(C)}) + \epsilon \tag{6.1}$$

where b_0 is an intercept term, $\epsilon \sim \text{Normal}(0, \sigma^2)$, and $f_g^{(\cdot)}(\cdot)$ is a penalized regression

spline. We then calculate the fitted pieces as $M_g = \widehat{f_g^{(M)}}(R_g^{(M)})$, $C_g = \widehat{f_g^{(C)}}(R_g^{(C)})$, and $O_g =$ the residuals. We interpret M_g as the part of gene g expression explained by methylation, C_g as the part of gene g expression explained by copy number, and O_g as the part of gene g expression explained by something "other" than methylation or copy number. (Although not discussed in this note, this partitioning is useful for model building, and can also be extended to incorporate an arbitrary number of platforms, including protein expression, as long as the underlying biological relationships among the platforms are understood (Jennings et al., 2013; McGuffey et al., 2015).)

Traditional heatmaps depict a matrix of data through color, with different colors and intensities representing a certain range of data values. Rows and columns may be clustered, depending on the user's objective. Our integrative heatmaps also illustrate the data through color intensities and have clustering options, but the matrix of data to be depicted is grouped by platform. Specifically, we will work with the matrices $mRNA = \{mRNA_1, ..., mRNA_G\}$, $M = \{M_1, ..., M_G\}$, $C = \{C_1, ..., C_G\}$, and $O = \{O_1, ..., O_G\}$ where G is the total number of genes in the analysis.

We present three variations of the integrative heatmap (IH), each achieving a distinctly different purpose. For our application, we use a subset of publicly available colorectal cancer (CRC) data from The Cancer Genome Atlas (TCGA). Our heatmaps are especially useful for comparing structures between sample groups; here we use four CRC subtypes based on the classification of Guinney et al. (2015). If a user did not have sample classes of interest, the IHs could accommodate this by simply considering all samples to be of the same class, and the results would still be useful to compare structures accross platforms. We also restrict ourselves to 423 genes that have been previously filtered and determined to be useful in characterizing the CRC subtypes. The user supplies the raw data matrices and thus may filter

genes (or not) however he or she chooses.

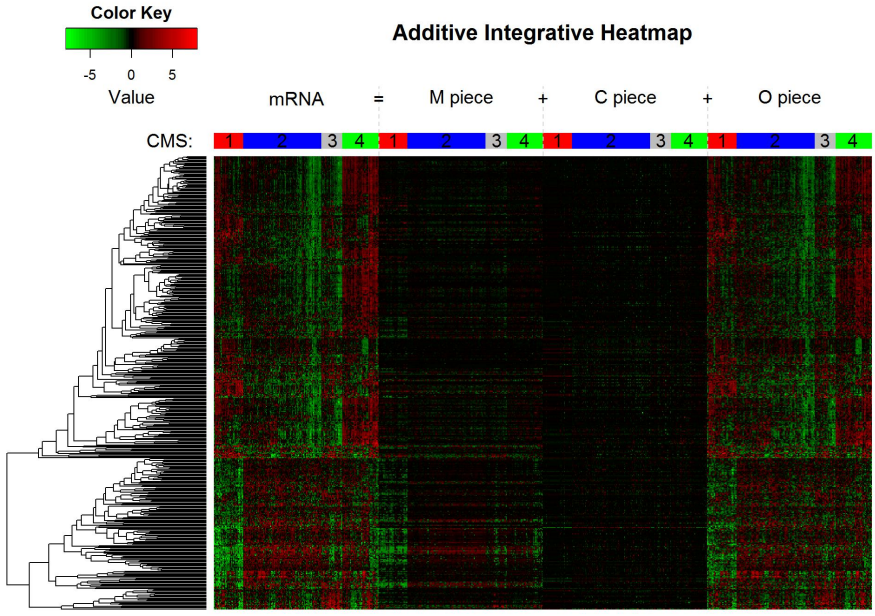
6.2.1 Additive integrative heatmaps

In the additive IH, the priorities are (1) to visualize the additivity of the partitioned components and (2) to compare the structure within each sample class for each platform on a common scale. The matrix of data values that is plotted is $\{mRNA, M, C, O\}$, with clear platform distinctions. The rows are clustered using the entire matrix. The columns are ordered based on clustering of the matrix $mRNA$; the columns of $mRNA$ are clustered *within sample class*, and that order is repeated for the columns of M, C, and O to maintain additivity.

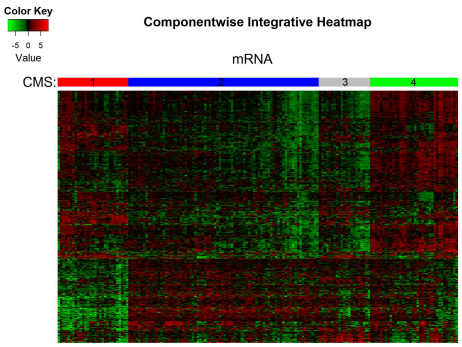
The common color scale for all platforms and the clustering approach allows us to quickly see how the mRNA expression values are (additively) partitioned into the different components. Including all the platforms also facilitates easy identification of which platform(s) stand(s) out as a driving force for which sample class. The additive heatmap for our CRC example is shown in Figure 6.1a.

6.2.2 Componentwise heatmaps

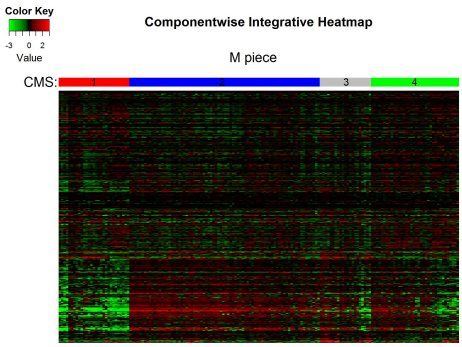
The goal for the componentwise IH is to "zoom in" on a component of the additive IH ($mRNA$, M, C, or O) and investigate the differences in structure between the sample classes for that particular component. To achieve this goal, we plot the chosen component matrix, but allow a new color scheme suited to the range of values observed for that component. This makes the underlying structure more evident, especially for the components with a smaller range of values. We also calculate new column clustering within each sample class, whereas the row ordering remains the same as in the additive IH to facilitate the comparison of important genes across componentwise IHs. The componentwise IH for each component in our example can be seen in Figure 6.1b-6.1e.



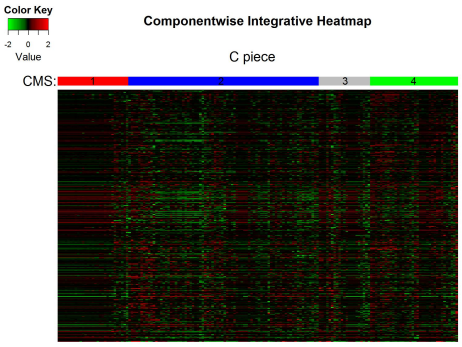
(a) Additive integrative heatmap.



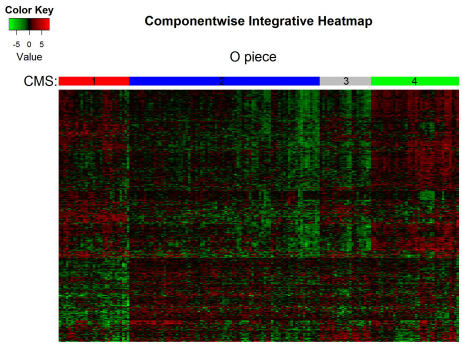
(b) Componentwise IH for $mRNA$.



(c) Componentwise IH for M.



(d) Componentwise IH for C.



(e) Componentwise IH for O.

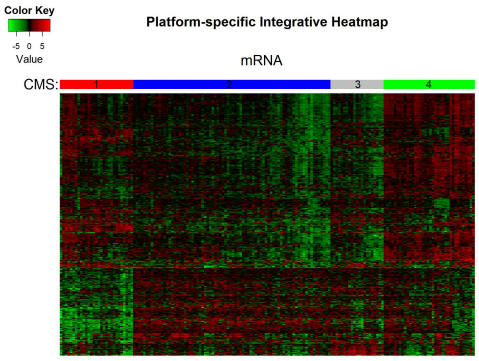
Figure 6.1: Integrative heatmaps. (a) Additive IH for TCCGA CRC data. (b)-(e) Componentwise IHs for mRNA, methylation, copy number, and other platforms, respectively. The row (gene) order is held constant for all heatmaps. "CMS" is the label for sample subtype.

6.2.3 Platform-specific integrative heatmaps

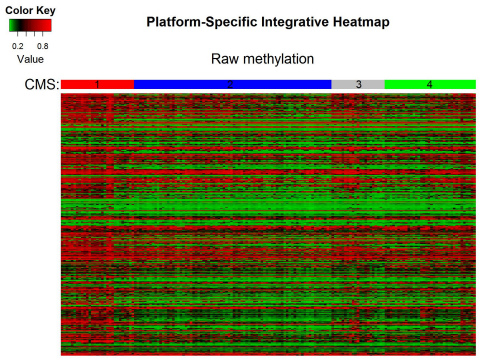
The platform-specific IHs are constructed in the same manner as the componentwise IHs (new color scheme; same row ordering as additive heatmap; new column clustering within sample classes), but they plot the raw data values, as opposed to the partitioned components. This facilitates comparisons of the information present in each raw data platform to the structure seen in the corresponding partitioned piece. For example, the platform-specific IH for methylation depicts the varying levels of methylation, whereas the componentwise IH illustrates the varying levels of gene expression *explained by* methylation. The platform-specific IHs for the TCGA CRC data are shown in Figure 6.2.

6.3 Results

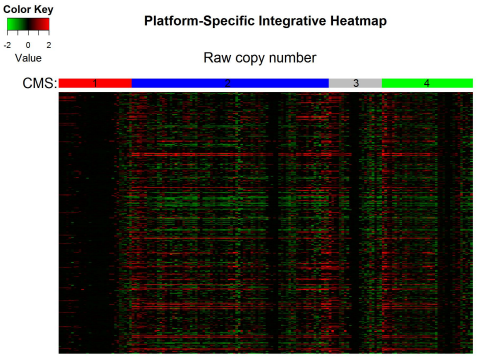
In the additive IH (Figure 6.1a) we see clear structural differences between sample classes in the $mRNA$ matrix. For example, we see that class 1 is distinguished from the other classes by low expression of the bottom third genes. It is also evident that the low levels of expression of these genes is explained by methyation and something other than methylation or copy number (the M and O pieces, respectively). The additivity property is demonstrated nicely by this same group of genes; the bright green and red intensities in the $mRNA$ matrix are decomposed into the slightly less intense colors of the M and O pieces and the even darker C piece. The varying color intensities depict the amount of each gene's expression that is explained by each platform, with duller intensities indicating that less expression can be attributed to that platform. Although the common color scheme makes the additivity clear, it makes some of the structure in the M and C pieces appear only faintly. However, we can utilize the componentwise IHs to further investigate and compare the differences between sample classes.



(a) Platform-specific IH for $mRNA$.



(b) Platform-specific IH for methylation.



(c) Platform-specific IH for copy number.

Figure 6.2: Platform-specific integrative heatmaps for TCGA CRC data. (a) Platform-specific IH for mRNA. This is identical to the componentwise IH for mRNA because both use the raw data. (b)-(c) Platform-specific IHs for (raw) methylation and (raw) copy number, respectively. The row (gene) order is held constant for all heatmaps. "CMS" is the label for sample subtype.

The M componentwise IH (Figure 6.1c) provides a sharper illustration of the structure in the M piece, the part of gene expression explained by methylation. The low levels of expression explained by methylation characterizing class 1 in the bottom third of the genes that we observed in the additive IH is now even brighter, and we also see structures that were not clear in the additive IH. For example, higher levels of expression explained by methylation in the bottom third genes (shown as red) are now seen to characterize class 2, and similarly appear to characterize class 4 through the top two thirds of the genes. Not as much information appears in the componentwise heatmap for the C piece (Figure 6.1d), but it does become clear that a lack of gene expression explaination by copy number, shown in black, characterizes many of the class 1 samples.

It is logical to wonder whether the same structures would be present in the raw platform data. The partitioning step can be thought of as a de-noising of the raw platform values, and as such it is common for the structure seen in the componentwise IHs to be different, and often clearer, than that of the platform-specific IHs. This can be seen in the platform-specific IHs shown in Figure 6.2. Specifically, in the platform-specific IH of the raw methylation data, we do not see clear class distinctions from the groups of genes discussed in the componentwise IH summary – the information in the raw data values is different than that in the partitioned pieces. It is worth noting, however, that in class 1 most of the methylation values for the bottom third of the genes are high (red). Because increased methylation typically results in lower gene expression, this is consistent with the previous finding that low levels of gene expression explained by methylation in this group of genes characterizes class 1. If the rows (genes) in the platform-specific IH were to be re-clustered, some structural differences between classes may become apparent, but they would be driven by different groups of genes than seen earlier. Thus, both the component-

wise and platform-specific IHs can contribute something distinctly meaningful to the biological understanding and interpretations of an integrative analysis.

6.4 Discussion

In the additive IH, the color intensities allow us to see how each gene's expression is decomposed into the pieces explained by each upstream platform. We can also identify several structural characteristics that characterize certain sample classes. When we use the componentwise IHs to zoom in on each partitioned piece, even more structural distinctions among the classes become evident. Then the platform-specific IHs allow us to make deeper biological connections by incorporating the information from the raw platform values, and we are able to compare the differences in structure between the raw values and the partitioned components.

In our example, the patterns seen in the partitioned components are much clearer than those in the raw values. This speaks to the advantages offered by integrating multiple platforms via the method described in Section 6.2. Although not discussed here, this method has general usefulness and independent value for model-building (see Wang et al. (2013); Jennings et al. (2013); McGuffey et al. (2015)), as it provides a cleaner signal than that available from the raw data. In this example, we considered the platforms of gene expression, methylation, and copy number, but other platforms such as protein expression and microRNA can also be accommodated (McGuffey et al., 2015).

The dendrograms produced by our IH code can also be used as input for MD Anderson Cancer Center's next-generation clustered heatmap application (MD Anderson Cancer Center, 2015). These next-generation heatmaps allow the user to zoom in on different parts of the heatmap and to display other information related to each gene and sample, such as gene pathway membership and clinical informa-

tion for each patient. This is an excellent tool for exploratory data analysis for high-dimensional data.

7. CONCLUSION

In this dissertation, we presented several statistical methods to analyze multi-platform genomic data, as well as a novel illustration tool useful for visualizing the integration of the data platforms, all with applications to cancer research. In Chapters 2-4, we presented a linear and a nonlinear formulation of iBAG, a hierarchical two-step Bayesian model that integrates multiple data types by modeling the known biological relationships among them, and flags genes important to a clinical outcome. The NG prior on the effects facilitates flexible shrinkage and induces sparsity, resulting in improved estimation efficiency and increased power to identify significant genes. Whereas we integrate three platforms as predictors in our data applications, we proposed natural extensions to incorporate miRNA and protein expression as predictor platforms; as long as the biological relationships among the platforms are understood, essentially any platform could be accommodated. Beyond identifying important genes, the structure of the data integration also provides mechanistic information regarding through which platform the gene expression is related to the clinical outcome. We applied iBAG to a glioblastoma multiforme (GBM) data set from The Cancer Genome Atlas (TCGA), using survival time as the clinical response, and identified several potential prognostic genes, some of them not previously implicated in GBM progression.

We presented piBAG, a pathway iBAG model, in Chapter 5. This method is also a two-step hierarchical Bayesian model, with similar integration and gene selection properties as the non-pathway iBAG model, but piBAG also incorporates gene pathway membership information. The model design borrows strength within each pathway to aid in estimation efficiency, and the results of the method include

flagging important genes and their driving platforms, as before, but also a list of pathways ranked by their importance to the clinical outcome through each platform. The piBAG method shares many of the advantages of the regular iBAG model discussed above, such as data integration, flexible shrinkage, efficient estimation, and high power to detect significant genes. However the additional information gained by implementing piBAG, specifically the pathway scores and a ranking of pathway clinical relevance, which are estimated simultaneously with the gene effects, sets this method apart. We demonstrated in a simulation study that the ideal scenario is to integrate multiple data platforms and fit piBAG; however, we also showed that fitting the pathway iBAG model to a single platform predictor provides the advantage of additional pathway-level results without losing power or accuracy on the gene level. Using a new subset of the TCGA GBM data, with survival time as the clinical outcome, we applied piBAG and identified four potential prognostic markers, as well as pathway rankings. We also compared the results to an application of the method without pathway information and thus provided more evidence of the increased power available through the pathway iBAG method.

Finally, in Chapter 6 we presented integrative heatmaps (IHs), a visualization tool to illustrate genomic data integration and provide insight into the genetic differences between cancer subtypes. The additive IH provides a big picture view and clearly portrays the partitioning of gene expression into the components explained by various upstream platforms. We then zoom in on each of the components in the componentwise IHs, allowing a more detailed view of where and how that particular partitioned piece differs between samples of differing subtype. The platform-specific IHs depict the raw data platforms, as opposed to the partitioned components, and facilitate a comparison of the information contained in the raw versus partitioned data. We showed these variations as applied to a colorectal cancer data set avail-

able from the TCGA repository, and we interpreted each of the three variations to demonstrate how they achieve their respective goals.

All of the methods presented in this dissertation provide information that is critical to the development of targeted cancer therapies. A final important note is that although the majority of the applications in this dissertation focus on a glioblastoma multiforme (GBM) data set, the proposed methods are not cancer-specific; they can be applied to any cancer type as long as the appropriate data is available.

REFERENCES

American Association of Neurological Surgeons (2012), "Glioblastoma Multiforme," http://www.aans.org/Patient%20Information/Conditions%20and%20Treatments/Glioblastoma%20Multiforme.aspx.

American Brain Tumor Association (2013), "Glioblastoma," http://www.abta.org/understanding-brain-tumors/types-of-tumors/glioblastoma.html.

American Cancer Society (2013), *American Cancer Society: Cancer Facts and Figures 2013*, American Cancer Society, Atlanta, GA.

— (2014), *American Cancer Society: Cancer Facts and Figures 2014*, American Cancer Society, Atlanta, GA.

— (2015a), *American Cancer Society: Cancer Facts and Figures 2015*, American Cancer Society, Atlanta, GA.

— (2015b), "Targeted Therapy," www.cancer.org/targeted-therapy-pdf.

Barbieri, M. M. and Berger, J. O. (2004), "Optimal predictive model selection," *Annals of Statistics*, 32, 870–897.

Bardelli, A., Parsons, D. W., Silliman, N., Ptak, J., Szabo, S., Saha, S., Markowitz, S., Willson, J. K. V., Parmigiani, G., Kinzler, K. W., Vogelstein, B., and Velculescu, V. E. (2003), "Mutational Analysis of the Tyrosine Kinome in Colorectal Cancers," *Science*, 300, 949.

Bell, D., Berchuck, A., Birrer, M., Chien, J., Cramer, D., Dao, F., Dhir, R., et al. (2011), "Integrated genomic analyses of ovarian carcinoma," *Nature*, 474, 609–615.

Benjamini, Y. and Hochberg, Y. (1995), "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society, Series B (Methodological)*, 57, 289–300.

Brennan, C., Momota, H., Hambardzumyan, D., Ozawa, T., Tandon, A., Pedraza, A., and Holland, E. (2009), "Glioblastoma subclasses can be defined by activity among signal transduction pathways and associated genomic alterations," *Public Library of Science One*, 4, e7752.

Cancer.net (2012), "Side Effects of Radiation Therapy," `http://www.cancer.net/navigating-cancer-care/how-cancer-treated/radiation-therapy/side-effects-radiation-therapy`.

Carapancea, M., Alexandru, O., Fetea, A. S., Dragutescu, L., Castro, J., Georgescu, A., Popa-Wagner, A., Backlund, M. L., Lewensohn, R., and Dricu, A. (2009), "Growth factor receptors signaling in glioblastoma cells: therapeutic implications," *Journal of Neurooncology*, 92, 137–147.

Chakravarti, A., Loeffler, J. S., and Dyson, N. J. (2002), "Insulin-like growth factor receptor I mediates resistance to anti-epidermal growth factor receptor therapy in primary human glioblastoma cells through continued activation of phosphoinositide 3-kinase signaling," *Cancer Research*, 62, 200–207.

Cooper, L. A., Gutman, D. A., Long, Q., Johnson, B. A., Cholleti, S. R., Kurc, T., Saltz, J. H., Brat, D. J., and Moreno, C. S. (2010), "The proneural molecular signature is enriched in oligodendrogliomas and predicts improved survival among diffuse gliomas," *Public Library of Science One*, 5, e12548.

Craig, D. W., O'Shaughnessy, J. A., Kiefer, J. A., Aldrich, J., Sinari, S., Moses, T. M., Wong, S., Dinh, J., Christoforides, A., Blum, J. L., Aitelli, C. L., Osborne, C. R., Izatt, T., Kurdoglu, A., Baker, A., Koeman, J., Barbacioru, C., Sakarya, O., De La Vega, F. M., Siddiqui, A., Hoang, L., Billings, P. R., Salhia, B., Tolcher, A. W., Trent, J. M., Mousses, S., Von Hoff, D., and Carpten, J. D. (2013), "Genome and transcriptome sequencing in prospective metastatic triple-negative breast cancer uncovers therapeutic vulnerabilities," *Molecular Cancer Therapeu-*

tics, 12, 104–116.

Dabney, A. R. (2006), "ClaNC: point-and-click software for classifying microarrays to nearest centroids," *Bioinformatics*, 22, 122–123.

Goldstein, M., Meller, I., and Orr-Urtreger, A. (2007), "FGFR1 over-expression in primary rhabdomyosarcoma tumors is associated with hypomethylation of a 5' CpG island and abnormal expression of the AKT1, NOG, and BMP4 genes," *Genes Chromosomes Cancer*, 46, 1028–1038.

Griffin, J. E. and Brown, P. J. (2010), "Inference with normal-gamma prior distributions in regression problems," *Baysian Analysis*, 5, 171–188.

— (2012), "Structuring shrinkage: some correlated priors for regression," *Biometrika*, 99, 481–487.

Guinney, J., Dienstmann, R., Wang, X., de Reynies, A., Schlicker, A., Soneson, C., Marisa, L., Roepman, P., Nyamundanda, G., Angelino, P., Bot, B., Morris, J., Simon, I., Gerster, S., Fessler, E., de Sousa e Melo, F., Missiaglia, E., Ramay, H., Barras, D., Homicsko, K., Maru, D., Manyam, G., Broom, B., Boige, V., Laderas, T., Salazar, R., Gray, J., Tabernero, J., Bernards, R., Friend, S., Laurent-Puig, P., Medema, J., Sadanandam, A., Wessels, L., Delorenzi, M., Kopetz, S., Vermeulen, L., and Tejpar, S. (2015), "The consensus molecular subtypes of colorectal cancer," Under review.

Hastie, T. and Tibshirani, R. (1986), "Generalized additive models," *Statistical Science*, 1, 297–310.

Hewish, M., Chau, I., and Cunningham, D. (2009), "Insulin-like growth factor 1 receptor targeted therapeutics: novel compounds and novel treatment strategies for cancer medicine," *Recent Patents on Anticancer Drug Discovery*, 4, 54–72.

Ikeda, K., Saitoh, S., Tsubota, A., Arase, Y., Chayama, K., Kumada, H., Watanabe, G., and Tsurumaru, M. (1993), "Risk factors for tumor recurrence and prognosis

after curative resection of hepatocellular carcinoma," *Cancer*, 71, 19–25.

Ingenuity Systems (2015), *Calculating and Interpreting the p-values for Functions, Pathways and Lists in IPA*, Ingenuity Systems, IPA 8.5 ed.

Jennings, E. M., Morris, J. S., Carroll, R. J., Manyam, G. C., and Baladandayuthapani, V. (2012), "Hierarchical Bayesian methods for integration of various types of genomics data," *Genomic Signal Processing and Statistics, (GENSIPS), 2012 IEEE International Workshop*, 5–8.

— (2013), "Bayesian methods for expression-based integration of various types of genomics data," *European Association for Signal Processing (EURASIP) Journal on Bioinformatics and Systems Biology*, 2013, 13.

Jiang, Y., Boije, M., Westermark, B., and Uhrbom, L. (2011), "PDGF-B Can sustain self-renewal and tumorigenicity of experimental glioma-derived cancer-initiating cells by preventing oligodendrocyte differentiation," *Neoplasia*, 13, 492–503.

Johnson, D. R. and O'Neill, B. P. (2012), "Glioblastoma survival in the United States before and during the temozolomide era," *Journal of Neurooncology*, 107, 359–364.

Kanu, O. O., Hughes, B., Di, C., Lin, N., Fu, J., Bigner, D. D., Yan, H., and Adamson, C. (2009), "Glioblastoma Multiforme Oncogenomics and Signaling Pathways," *Clinical Medicine: Oncology*, 3, 39–52.

Kapoor, G. S. and O'Rourke, D. M. (2010), "SIRPalpha1 receptors interfere with the EGFRvIII signalosome to inhibit glioblastoma cell transformation and migration," *Oncogene*, 29, 4130–4144.

Katoh, M. and Nakagama, H. (2013), "FGF receptors: cancer biology and therapeutics," *Medical Research Reviews*, 34, 280–300.

Khuri, F. R., Kim, E. S., Lee, J. J., Winn, R. J., Benner, S. E., Lippman, S. M., Fu, K. K., Cooper, J. S., Vokes, E. E., Chamberlain, R. M., Williams, B., Pajak, T. F., Goepfert, H., and Hong, W. K. (2001), "The impact of smoking status, disease

stage, and index tumor site on second primary tumor incidence and tumor recurrence in the head and neck retinoid chemoprevention trial," *Cancer Epidemiology, Biomarkers & Prevention*, 10, 823–829.

Kyung, M., Gill, J., Ghosh, M., and Casella, G. (2010), "Penalized Regression, Standard Errors, and Bayesian Lassos," *Bayesian Analysis*, 5, 369–412.

Lanckriet, G. R., De Bie, T., Cristianini, N., Jordan, M. I., and Noble, W. S. (2004), "A statistical framework for genomic data fusion," *Bioinformatics*, 20, 2626–2635.

Liu, D., Lin, X., and Ghosh, D. (2007), "Semiparametric Regression of Multidimensional Genetic Pathway Data: Least Squares Kernel Machines and Linear Mixed Models," *Biometrics*, 63, 1079–1088.

Loilome, W., Joshi, A. D., ap Rhys, C. M., Piccirillo, S., Vescovi, A. L., Angelo, V. L., Gallia, G. L., and Riggins, G. J. (2009), "Glioblastoma cell growth is suppressed by disruption of Fibroblast Growth Factor pathway signaling," *Journal of Neurooncology*, 94, 359–366.

Lowenstein, E. J., Daly, R. J., Batzer, A. G., Li, W., Margolis, B., Lammers, R., Ullrich, A., Skolnik, E. Y., Bar-Sagi, D., and Schlessinger, J. (1992), "The SH2 and SH3 domain-containing protein GRB2 links receptor tyrosine kinases to ras signaling," *Cell*, 70, 431–442.

Martini, N., Bains, M. S., Burt, M. E., Zakowski, M. F., McCormack, P., Rusch, V. W., and Ginsberg, R. J. (1995), "Incidence of local recurrence and second primary tumors in resected stage I lung cancer," *Journal of Thoracic and Cardiovascular Surgery*, 109, 120–129.

McGuffey, E. J., Morris, J. S., Manyam, G. C., Carroll, R. J., and Baladandayuthapani, V. (2015), "Bayesian models for flexible integrative analysis of multiplatform genomics data," in *Integrating Omics Data*, eds. Tseng, G. C., Ghosh, D., and Zhou, X. J., New York, NY: Cambridge University Press.

McLendon, R., Friedman, A., Bigner, D., Van Meir, E. G., Brat, D. J., Mastrogianakis, G. M., Olson, J. J., et al. (2008), "Comprehensive genomic characterization defines human glioblastoma genes and core pathways," *Nature*, 455, 1061–1068.

MD Anderson Cancer Center (2015), "Next Generation Clustered Heat Maps," `http://bioinformatics.mdanderson.org/main/NG-CHM:Overview`, Department of Bioinformatics and Computational Biology.

Memorial Sloan-Kettering Cancer Center (2012), "Pathway analysis of genetic alterations in glioblastoma (TCGA)," `http://cbio.mskcc.org/cancergenomics/gbm/pathways/`.

Michaelson, M. D. and Oh, W. K. (2013), "Treatment-related toxicity in men with testicular germ cell tumors," `http://www.uptodate.com`.

Muller, P., Parmigiani, G., and Rice, K. (2006), "FDR and Bayesian Multiple Comparisons Rules," Tech. rep., Johns Hopkins University, Dept. of Biostatistics Working Papers, Working Paper 115.

National Cancer Institute (2015), "Targeted Cancer Therapies," `http://www.cancer.gov/cancertopics/factsheet/Therapy/targeted`, National Cancer Institute Fact Sheets.

National Human Genome Research Institute (2015), "Biological Pathways," `http://www.genome.gov/27530687#al-8`.

Nazarenko, I., Hede, S. M., He, X., Hedren, A., Thompson, J., Lindstrom, M. S., and Nister, M. (2012), "PDGF and PDGF receptors in glioma," *Upsala Journal of Medical Sciences*, 117, 99–112.

Noushmehr, H., Weisenberger, D. J., Diefes, K., Phillips, H. S., Pujara, K., Berman, B. P., Pan, F., Pelloski, C. E., Sulman, E. P., Bhat, K. P., Verhaak, R. G., Hoadley, K. A., Hayes, D. N., Perou, C. M., Schmidt, H. K., Ding, L., Wilson, R. K., Van

Den Berg, D., Shen, H., Bengtsson, H., Neuvial, P., Cope, L. M., Buckley, J., Herman, J. G., Baylin, S. B., Laird, P. W., and Aldape, K. (2010), "Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma," *Cancer Cell*, 17, 510–522.

Park, T. and Casella, G. (2008), "The Bayesian Lasso," *Journal of the American Statistical Association*, 103, 681–686.

Prigent, S. A., Nagane, M., Lin, H., Huvar, I., Boss, G. R., Feramisco, J. R., Cavenee, W. K., and Huang, H. S. (1996), "Enhanced tumorigenic behavior of glioblastoma cells expressing a truncated epidermal growth factor receptor is mediated through the Ras-Shc-Grb2 pathway," *Journal of Biological Chemistry*, 271, 25639–25645.

Rebocho, A. P. and Marais, R. (2013), "ARAF acts as a scaffold to stabilize BRAF:CRAF heterodimers," *Oncogene*, 32, 3207–3212.

Ruano, Y., Mollejo, M., Ribalta, T., Fiano, C., Camacho, F. I., Gomez, E., de Lope, A. R., Hernandez-Moneo, J. L., Martinez, P., and Melendez, B. (2006), "Identification of novel candidate target genes in amplicons of Glioblastoma multiforme tumors detected by expression and CGH microarray profiling," *Molecular Cancer*, 5, 39.

Sawyers, C. (2004), "Targeted cancer therapy," *Nature*, 432, 294–297.

Sawyers, C. L. (2003), "Opportunities and challenges in the development of kinase inhibitor therapy for cancer," *Genes & Development*, 17, 2998–3010.

Shen, R., Olshen, A. B., and Ladanyi, M. (2009), "Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis," *Bioinformatics*, 25, 2906–2912.

Solomon, D. A., Kim, J. S., Jenkins, S., Ressom, H., Huang, M., Coppa, N., Mabanta, L., Bigner, D., Yan, H., Jean, W., and Waldman, T. (2008), "Identification of p18 INK4c as a tumor suppressor gene in glioblastoma multiforme," *Cancer Research*,

68, 2564–2569.

Suzuki, K., Momota, H., Tonooka, A., Noguchi, H., Yamamoto, K., Wanibuchi, M., Minamida, Y., Hasegawa, T., and Houkin, K. (2010), "Glioblastoma simultaneously present with adjacent meningioma: case report and review of the literature," *Journal of Neurooncology*, 99, 147–153.

The Cancer Genome Atlas (2012), "Program Overview," `http://cancergenome.nih.gov/abouttcga/overview`.

The Cancer Genome Atlas Data Portal (2013), "Data Levels and Data Types," `https://tcga-data.nci.nih.gov/tcga/tcgaDataType.jsp`.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society, Series B (Methodological)*, 58, 267–288.

Tyekucheva, S., Marchionni, L., Karchin, R., and Parmigiani, G. (2011), "Integrating diverse genomic data using gene sets," *Genome Biology*, 12, R105.

Verhaak, R. G., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., Miller, C. R., et al. (2010), "Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1," *Cancer Cell*, 17, 98–110.

Vogelstein, B. and Kinzler, K. W. (2004), "Cancer genes and the pathways they control," *Nature Medicine*, 10, 789–799.

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.

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

Welsh, J. B., Sapinoso, L. M., Su, A. I., Kern, S. G., Wang-Rodriguez, J., Moskaluk, C. A., Frierson, H. F., and Hampton, G. M. (2001), "Analysis of gene expres-

sion identifies candidate markers and pharmacological targets in prostate cancer," *Cancer Research*, 61, 5974–5978.

Wood, S. (2014), "Package `mgcv`," `http://cran.r-project.org/web/packages/mgcv/mgcv.pdf`.

Yin, D., Ogawa, S., Kawamata, N., Tunici, P., Finocchiaro, G., Eoli, M., Ruckert, C., Huynh, T., Liu, G., Kato, M., Sanada, M., Jauch, A., Dugas, M., Black, K. L., and Koeffler, H. P. (2009), "High-resolution genomic copy number profiling of glioblastoma multiforme by single nucleotide polymorphism DNA microarray," *Molecular Cancer Research*, 7, 665–677.

APPENDIX A

CHAPTER 3 SUPPLEMENT

A.1 Data Imputation

Since the percentage of missing data is so low ($\sim$ 5% for methylation and $\sim$ 0.1% for copy number), we choose to do imputation using the following algorithm for both the methylation data and the copy number data: (1) For each marker, replace any NA's with the mean of the other patients. Call this resulting matrix Temp. (2) Use Temp to calculate a correlation matrix between markers. (3) For each marker with missing value(s), regress it on the three markers which it is most highly positively correlated with (using the Temp matrix for the predictors to avoid further complications from missing data). (4) Substitute this predicted value for the missing value in the original matrix.

A.2 Complete Conditionals

$$
\begin{aligned}
\boldsymbol{\beta}|\text{rest} &\sim \text{Normal}\{(X^{\mathrm{T}}X+\sigma^2 D_\tau^{-1})^{-1}X^{\mathrm{T}}\mathbf{Y}, \sigma^2(X^{\mathrm{T}}X+\sigma^2 D_\tau^{-1})^{-1}\} \\
\sigma^2|\text{rest} &\sim \text{Inv.Gamma}(a=a+n/2, b=b+\{(\mathbf{Y}-X\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{Y}-X\boldsymbol{\beta})\}/2) \\
\psi_{j,i}|\text{rest} &\sim \text{Gen.Inv.Gaussian}(a=\gamma_j^{-2}, \quad b=\beta_{j,i}^2, p=\lambda_j-1/2), \text{where} \\
&\quad V=\text{Gen.Inv.Gaussian}(a,b,p) \text{ has density} \\
&\quad (a/b)^{p/2}v^{p-1}\exp\{-(av+b/v)/2\}/\{2K_p(\sqrt{ab})\}, \text{ where } K_p(\cdot) \text{ is a} \\
&\quad \text{modified Bessel function of the second kind.} \\
\lambda_j|\text{rest} &\sim (1/\lambda_j)^{\tilde{a}}\exp\{-\tilde{b}\gamma_j^{-2}/(2\lambda_j)-c\lambda_j\}\times\left(\prod_{i=1}^{p_j}\psi_{j,i}^{\lambda_j}\right)/[\{\Gamma(\lambda_j)\}^{p_j}(2\gamma_j^2)^{p_j\lambda_j}] \\
\gamma_j^{-2}|\text{rest} &\sim \text{Gamma}(a=p_j\lambda_j+\tilde{a}, b=(\tilde{b}/\lambda_j+\sum_{i=1}^{p_j}\psi_{j,i})/2)
\end{aligned}
$$

In the Metropolis-Hastings update step, the proposed value is $\lambda_j^* = \exp(\sigma_\lambda^2 z)\lambda_j$

where $z \sim \text{Normal}(0,1)$ and the tuning parameter σ^2_λ is chosen to result in an acceptance rate between 20% and 30%. The acceptance probability is then

$$\min\left\{1, \frac{\pi(\lambda_j^*)}{\pi(\lambda_j)}\left(\frac{\Gamma(\lambda_j)}{\Gamma(\lambda_j^*)}\right)^{p_j}\left((2\gamma_j^2)^{-p_j}\prod_{i=1}^{p_j}\psi_{j,i}\right)^{\lambda_j^*-\lambda_j}\left(\frac{\lambda_j^*}{\lambda_j}\right)\right\}$$

where $\pi(\lambda_j) = (1/\lambda_j)^{\tilde{a}}\exp\{-\tilde{b}\gamma_j^{-2}/(2\lambda_j) - c\lambda_j\}$, the prior for λ_j.

A.3 Initial Values and Hyperparameters

The initial values and hyperparameters are chosen as follows:

- The hyperparameters for σ^2 are $a = b = 0.001$, so as to be uninformative.
- The hyperparameter for λ_j is $c = 1$ (Griffin and Brown, 2010).
- The hyperparamters for γ_j^{-2} are $\tilde{a} = 2$ and $\tilde{b} =$ the mean of the least squares $\widehat{\beta^2_{j,i}}$ (Griffin and Brown, 2010).
- The initial $\boldsymbol{\beta}$ is the estimate from the frequentist lasso with a single shrinkage parameter.
- The initial σ^2 is the mean sum of squares from the frequentist lasso.
- Each initial λ_j, $\psi_{j,i}$, and γ_j^{-2} is set to 1.

APPENDIX B

CHAPTER 4 SUPPLEMENT

B.1 Complete Conditionals

$$
\begin{aligned}
\boldsymbol{\beta}|\text{rest} &\sim \text{Normal}\{(X^{\mathrm{T}}X+\sigma^2 D_\psi^{-1})^{-1}X^{\mathrm{T}}\mathbf{Y}, \sigma^2(X^{\mathrm{T}}X+\sigma^2 D_\psi^{-1})^{-1}\} \\
\sigma^2|\text{rest} &\sim \text{Inv.Gamma}(a=a+n/2, b=b+\{(\mathbf{Y}-X\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{Y}-X\boldsymbol{\beta})\}/2) \\
\psi_{jg}|\text{rest} &\sim \text{Gen.Inv.Gaussian}(a=\gamma_j^{-2}, \quad b=\beta_{jg}^2, p=\lambda_j-1/2), \text{where} \\
&\quad V=\text{Gen.Inv.Gaussian}(a,b,p) \text{ has density} \\
&\quad (a/b)^{p/2}v^{p-1}\exp\{-(av+b/v)/2\}/\{2K_p(\sqrt{ab})\}, \text{where } K_p(\cdot) \\
&\quad \text{is a modified Bessel function of the second kind.} \\
\lambda_j|\text{rest} &\sim (1/\lambda_j)^{\tilde{a}}\exp\{-\tilde{b}\gamma_j^{-2}/(2\lambda_j)-c\lambda_j\}\left(\prod_{g=1}^{p_j}\psi_{jg}^{\lambda_j}\right)/[\{\Gamma(\lambda_j)\}^{p_j}(2\gamma_j^2)^{p_j\lambda_j}] \\
\gamma_j^{-2}|\text{rest} &\sim \text{Gamma}(a=p_j\lambda_j+\tilde{a}, b=(\tilde{b}/\lambda_j+\sum_{i=1}^{p_j}\psi_{jg})/2)
\end{aligned}
$$

In the Metropolis-Hastings update step, the proposed value is $\lambda_j^* = \exp(\sigma_\lambda^2 z)\lambda_j$, where $z \sim \text{Normal}(0,1)$ and the tuning parameter σ_λ^2 is chosen to result in an acceptance rate between 20% and 30%. The acceptance probability is then

$$
\min\left\{1, \frac{\pi(\lambda_j^*)}{\pi(\lambda_j)}\left(\frac{\Gamma(\lambda_j)}{\Gamma(\lambda_j^*)}\right)^{p_j}\left((2\gamma_j^2)^{-p_j}\prod_{g=1}^{p_j}\psi_{jg}\right)^{\lambda_j^*-\lambda_j}\left(\frac{\lambda_j^*}{\lambda_j}\right)\right\},
$$

where $\pi(\lambda_j) = (1/\lambda_j)^{\tilde{a}}\exp\{-\tilde{b}\gamma_j^{-2}/(2\lambda_j)-c\lambda_j\}$, the prior for λ_j.

B.2 Partitioning Explained Variation

Regardless of linear or nonlinear formulation, after estimating the mechanistic model (for one gene), we have

$$
y = \widehat{b_0} + M + CN + O \tag{B.1}
$$

where y is the gene expression. We carry forward M, CN, O and disregard $\widehat{b_0}$ as a kind of mean centering. The sums of squares that we use to calculate proportions of explained variances are as follows:

$$SST = \sum_{i=1}^{n} (y_i^*)^2 \tag{B.2}$$

$$SSM = \sum_{i=1}^{n} (M_i)^2 \tag{B.3}$$

$$SSCN = \sum_{i=1}^{n} (CN_i)^2 \tag{B.4}$$

$$SSE = SST - SSM - SSCN \tag{B.5}$$

where $y^* = y - \widehat{b_0}$. Then the proportion of variance explained by methylation is SSM/SST; the proportion explained by copy number is $SSCN/SST$; and the proportion explained by something other than methylation or copy number is SSE/SST.

Note that this formulation holds for $\widehat{b_0} = \bar{y}$, which occurs when, in the mechanistic model, $\int f_{jgk}(R_{jgk}) = 0$ for all j, g, and k. This is accomplished in the linear case by centering each R_{jgk} prior to fitting the model. For the nonlinear case, this is a common identifiability assumption that is used in most GAM packages.

APPENDIX C

CHAPTER 5 SUPPLEMENT

C.1 Complete Conditional Distributions

$$
\begin{aligned}
\boldsymbol{\beta}|\text{rest} &\sim \text{Normal}\{(X^{\mathrm{T}}X + \tau^2 D_{\sigma^2}^{-1})^{-1}X^{\mathrm{T}}\mathbf{Y}, \quad \tau^2(X^{\mathrm{T}}X + \tau^2 D_{\sigma^2}^{-1})^{-1}\} \\
\tau^2|\text{rest} &\sim \text{Inv.Gamma}(a = a + n/2, \quad b = b + \{(\mathbf{Y} - X\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{Y} - X\boldsymbol{\beta})\}/2) \\
\sigma^2_{pkg}|\text{rest} &\sim \text{Gen.Inv.Gaussian}(a = \xi_{pk}^{-2}, \quad b = \beta^2_{pkg}, \quad p = \alpha - 1/2), \text{ where} \\
&\quad V = \text{Gen.Inv.Gaussian}(a, b, p) \text{ has density} \\
&\quad (a/b)^{p/2} v^{p-1} \exp\{-(av + b/v)/2\}/\{2K_p(\sqrt{ab})\}, \text{where } K_p(\cdot) \\
&\quad \text{is a modified Bessel function of the second kind.} \\
\xi^2_{pk}|\text{rest} &\sim \text{Gamma}(G_k\alpha + \tilde{a}, \quad [\tilde{b}/\lambda + \sum_{g=1}^{G_k} \sigma^2_{pkg}]/2) \\
\alpha|\text{rest} &\sim \exp(-\tilde{c}\alpha)[\Gamma(\alpha)]^{-P\mathcal{G}}(1/2)^{\alpha P\mathcal{G}} \prod_{p=1}^{P}\prod_{k=1}^{K}\prod_{g=1}^{G_k}(\xi_{pk}^{-2}\sigma^2_{pkg})^{\alpha} \quad =: \mathbf{p_0(\boldsymbol{\alpha})} \\
\lambda|\text{rest} &\sim \text{Gen.Inv.Gaussian}(a = 2\tilde{d}, \quad b = \tilde{b}\sum_{p=1}^{P}\sum_{k=1}^{K}\xi_{pk}^{-2}, \quad p = 1 - PK\tilde{a})
\end{aligned}
$$

In the Metropolis-Hastings update step, the proposed value is $\alpha^* = \exp(\sigma_\alpha^2 z)\alpha$, where $z \sim \text{Normal}(0, 1)$ and the tuning parameter σ_α^2 is chosen to result in an acceptance rate between 30% and 40%. The acceptance probability is then

$$\min\left\{1, \quad \frac{p_0(\alpha^*)}{p_0(\alpha)}\left(\frac{\alpha^*}{\alpha}\right)\right\}$$

where $p_0(\alpha)$ is the complete conditional of α as defined above.

C.2 Hyperparameters and Starting Values

The hyperparameters for τ^2 are $a = 1$, $b = 2$ in the simulation study and $a = 0.1$, $b = 2.2$ in the data application. These are set to be minimally informative while placing enough mass away from zero so that τ^2 does not get "stuck" at 0 in the

MCMC chain. The remaining hyperparameters $\tilde{a}$, $\tilde{b}$, $\tilde{c}$, and $\tilde{d}$ are set to 1.

All β_{pkg} parameters have a starting value of 0, and all other parameters updated in the Gibbs sampler (τ^2, σ^2_{pkg}s, ξ^{-2}_{pk}s, α, and λ) start at 1. Trace plots show that all parameters mix well and converge quickly.