Biometrics 75, 183–192
March 2019

DOI: 10.1111/biom.12962

Bayesian Negative Binomial Mixture Regression Models for the Analysis of Sequence Count and Methylation Data

Qiwei Li,[1] Alberto Cassese,[2] Michele Guindani,[3] and Marina Vannucci[4,*]

[1]Department of Clinical Sciences, University of Texas Southwestern Medical Center, Dallas, Texas, U.S.A.
[2]Department of Methodology and Statistics, Faculty of Psychology and Neuroscience, Maastricht University, Maastricht, The Netherlands
[3]Department of Statistics, University of California, Irvine, California, U.S.A.
[4]Department of Statistics, Rice University, Houston, Texas, U.S.A.
*email: marina@rice.edu

Summary. In this article, we develop a Bayesian hierarchical mixture regression model for studying the association between a multivariate response, measured as counts on a set of features, and a set of covariates. We have available RNA-Seq and DNA methylation data measured on breast cancer patients at different stages of the disease. We account for the heterogeneity and over-dispersion of count data (here, RNA-Seq data) by considering a mixture of negative binomial distributions and incorporate the covariates (here, methylation data) into the model via a linear modeling construction on the mean components. Our modeling construction includes several innovative characteristics. First, it employs selection techniques that allow the identification of a small subset of features that best discriminate the samples while simultaneously selecting a set of covariates associated to each feature. Second, it incorporates known dependencies into the feature selection process via the use of Markov random field (MRF) priors. On simulated data, we show how incorporating existing information via the prior model can improve the accuracy of feature selection. In the analysis of RNA-Seq and DNA methylation data on breast cancer, we incorporate knowledge on relationships among genes via a gene-gene network, which we extract from the KEGG database. Our data analysis identifies genes which are discriminatory of cancer stages and simultaneously selects significant associations between those genes and DNA methylation sites. A biological interpretation of our findings reveals several biomarkers that can help understanding the effect of DNA methylation on gene expression transcription across cancer stages.

Key words: Count data; Feature selection; Integrative analysis; Markov random field; Mixture models; Negative binomial.

1. Introduction

In recent years, RNA sequencing (RNA-Seq), also known as high throughput or next-generation sequencing, has emerged as a powerful biotechnology for quantifying gene expression (Wang et al., 2009). RNA-Seq data consist of non-negative counts as number of reads observed in a region of interest (e.g., gene or exon) after genome mapping. Compared to microarray data, RNA-Seq does not suffer from cross-hybridization and poor quantification of low- and high-expressed genes (Kukurba and Montgomery, 2015). However, RNA-Seq data require specialized methods to take into account the skewness and the heterogeneity typically observed in these data. Initial modeling efforts were focused on data normalization (Bullard et al., 2010; Hansen et al., 2012) and on the detection of differentially expressed genes based on univariate testing procedures (Anders and Huber, 2010; Robinson et al., 2010). Recent contributions have employed methods for sample classification and clustering (Witten, 2011) and empirical Bayes methods that use mixture models and hierarchical frameworks for robust inference on gene expression changes (Lee et al., 2015; Leng et al., 2015). A recent review on RNA-Seq data analysis methods can be found in Conesa et al. (2016).

To our knowledge, few rigorous integrative modeling approaches exist to study the relationship between RNA-Seq and other genomic data, such as DNA methylation data. Required for embryonic development, DNA methylation is a process by which methyl groups are added to DNA. It typically acts to suppress gene transcription when located in a gene promoter. Several epigenetic studies have revealed that DNA methylation plays an important role in cancer since it can influence gene expression, see for example Murrell et al. (2005). Integrative analyses were conducted for example by Ferrón et al. (2011), who linked DNA methylation to tissue-specific gene expression in mice, and Xie et al. (2011), who correlated DNA methylation variation between human tissues with gene expression levels. Their results indicated that DNA methylation influences tissue differentiation via regulating gene expression. More recently, Tang et al. (2017) have developed a Bayesian Gaussian regression model to measure the relationship among DNA methylation, differential gene expression and tumor suppressor gene status, and Ma et al. (2017) have employed a multiple network framework for epigenetic studies. As the field of epigenomics expands to study several types of normal and pathological processes, it has become increasingly significant to understand the role that global genome-wide DNA methylation patterns play in influencing RNA-Seq gene expression. However, the development of statistical models for the understanding of DNA methylation in regulating gene expression is still limited.

Motivated by data measured on breast cancer patients at different stages of the disease, in this article, we develop an integrative Bayesian hierarchical mixture regression model for studying the association between a multivariate response, measured as counts on a set of features (i.e., RNA-Seq genes), and a set of covariates (i.e., DNA probes). We account for the heterogeneity and over-dispersion of the count data by considering a mixture of negative binomial regressions, where the covariates are incorporated into the model via a linear modeling construction on the mean components. Our modeling framework includes several innovative characteristics. First, it employs selection techniques that allow the identification of a small subset of features that best discriminate the available sample groups, while simultaneously selecting a set of covariates associated to each feature. Second, it incorporates known dependencies into the feature selection process via the use of Markov random field (MRF) priors. On simulated data, we show how employing available information via the prior model can improve the accuracy of feature selection. In the analysis of RNA-Seq and DNA methylation data, we capture existing relationships among genes via a gene-gene network, which we extract from the KEGG database. Our data analysis identifies genes which are discriminatory of cancer stages and simultaneously selects significant associations of those genes with DNA methylation sites. A biological interpretation of our findings reveals several biomarkers that can help understanding the effect of DNA methylation on gene expression transcription across cancer stages.

The remainder of the article is organized as follows. In section 2, we introduce the model and the priors. In section 3, we investigate results on the data analysis from a case study on breast cancer. In Section 4, we assess performances on simulated data and carry out comparisons with two-stage approaches. Section 5 concludes the article with some remarks.

2. Model

Let $\boldsymbol{Y}$ indicate a matrix of multivariate count data measured on a set of p features (here, RNA-Seq data) on n subjects, and let $\boldsymbol{X}$ indicate a $n \times R$ matrix of observations on R covariates (here, DNA methylation data). Finally, let the vector $\boldsymbol{z}$, with $z_i = k$, for $k \in \{1, \ldots, K\}$, indicate the known sample allocations of the n subjects to K groups (here, cancer stages). A graphical representation of the proposed model is given in Figure 1.

2.1. *Negative Binomial Mixture Model with Feature Selection*

We start by modeling the over-dispersed counts by a negative binomial (NB) mixture model in which we incorporate feature selection. For this, we first introduce a $p \times 1$ vector $\boldsymbol{\gamma}$ of binary latent indicators, with $\gamma_j = 1$ if feature j discriminates the n samples within the K given groups, and $\gamma_j = 0$ otherwise. Then, for sample i and $\gamma_j = 0$, we write

$$y_{ij}|\gamma_j \overset{ind}{\sim} \text{NB}(y_{ij}; \lambda_{ij0}, \phi_j), \tag{1}$$

while for $\gamma_j = 1$ we have

$$y_{ij}|z_i = k, \gamma_j \overset{ind}{\sim} \text{NB}(y_{ij}; \lambda_{ijk}, \phi_j), \tag{2}$$

for $i = 1, \ldots, n$, with $\text{NB}(y; \lambda, \phi)$ denoting a negative binomial distribution for the random variable y, with expectation λ and dispersion $1/\phi$. With this parametrization, the variance of the distribution can be written as $\lambda + \lambda^2/\phi$, allowing to model over-dispersion. Our model formulation assumes that counts mapping to non-discriminatory features are drawn from a negative binomial distribution with mean λ_{ij0} ("null" model), while any count that maps to a discriminatory feature and belongs to group k is drawn from a negative binomial distribution with mean λ_{ijk}. In Section 2.2, we describe how to incorporate covariates into the modeling construction via the mean components. We specify a prior for ϕ_j as $\phi_j \sim \text{Ga}(a_\phi, b_\phi)$.

Model (1) and (2) require a prior on the γ_j's. A simple choice in variable selection is to assume independent Bernoulli priors, that is, $\gamma_j \sim \text{Bern}(\alpha)$, where α can be either a fixed hyperparameter or a random variable itself. For example, a beta hyperprior can be imposed on α which leads to a beta-binomial prior on the number of discriminatory features. Recent contributions in the Bayesian variable selection modeling of genomic data have made use of Markov random field (MRF) priors that allow to incorporate external information about dependencies among predictors (Li and Zhang, 2010; Stingo and Vannucci, 2011). Here, we consider a MRF prior on $\boldsymbol{\gamma}$ that takes into account dependencies among the features (i.e., genes) as captured by a gene-gene interaction network that we extract from the KEGG database (Zhang and Wiemann, 2009). Specifically, we write

$$p(\gamma_j|\boldsymbol{\gamma}_{-j}) = \frac{\exp\left(\gamma_j(d + f\sum_{j' \in N_j} \gamma_{j'})\right)}{1 + \exp\left(d + f\sum_{j' \in N_j} \gamma_{j'}\right)}, \tag{3}$$

with d and f hyperparameters to be chosen, $\boldsymbol{\gamma}_{-j}$ denoting the vector of $\boldsymbol{\gamma}$ excluding the j-th element, and N_j the set of direct "neighbors" of feature j, as defined by the KEGG network. Then, according to (3) neighboring features are more likely to be jointly discriminatory. The joint prior on $\boldsymbol{\gamma}$, up to its normalizing constant, is

$$p(\boldsymbol{\gamma}) \propto \exp(d\mathbf{1}_{1\times p}\boldsymbol{\gamma} + f\boldsymbol{\gamma}^T\boldsymbol{G}\boldsymbol{\gamma}),$$

with $\boldsymbol{G}$ a $p \times p$ symmetric matrix with $g_{jj'} = 1$ if gene j and j' have a direct link in the gene network, and $g_{jj'} = 0$ otherwise (Hammersley and Clifford, 1971). Here d controls the sparsity of the prior model, while f affects the probability of selection of a feature according to the status of its neighbors. Note that if a feature does not have any neighbor, its prior distribution reduces to an independent Bernoulli with parameter

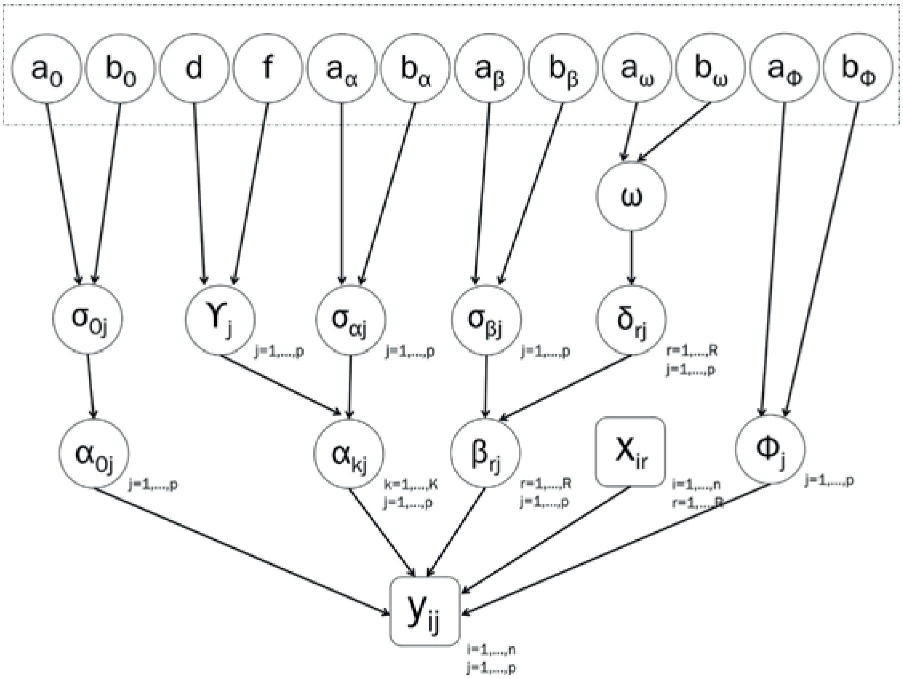


Figure 1. A graphical representation of the proposed Bayesian negative binomial mixture regression model. Each node in a circle refers to a parameter of the model and each node in a square refers to the observable data. Circle nodes in the dashed block are fixed hyperparameters. The link between two nodes represents a direct probabilistic dependence.

$\omega_\gamma = \exp(d)/(1+\exp(d))$, which is a logistic transformation of d.

2.2. *Mean Regression Model with Covariate Selection*

We incorporate the covariates into the modeling construction by specifying a log link model for the mean components, which we write as

$$\begin{cases} \log \lambda_{ij0} = \alpha_{0j} + \boldsymbol{x}_i^T \boldsymbol{\beta}_j & \text{if } \gamma_j = 0 \\ \log \lambda_{ijk} = \alpha_{0j} + \alpha_{kj} + \boldsymbol{x}_i^T \boldsymbol{\beta}_j & \text{if } \gamma_j = 1 \text{ and } z_i = k. \end{cases} \tag{4}$$

In this formulation, α_{0j} is the baseline process for feature j and it is shared by all observations. Note that $\exp(\alpha_{0j})$ can also be considered as a scaling factor adjusting for feature-specific levels across all observations. The group-specific parameter α_{kj} captures differential expression as a shift from the baseline, shared by all observations that belong to group k, when feature j is discriminatory; and it is set to zero, otherwise. To avoid identifiability problems arising from the sum of the components, we fix the mean shifts in the reference group, which is usually the first group, to zero, that is, $\alpha_{1j} = 0$, $j = 1, \ldots, p$.

The subject-specific coefficient vectors $\boldsymbol{\beta}_j$ in (4) describe the effect of the R covariates on the observed counts. Note that our model formulation assumes a relationship between the covariates and all the features, both discriminatory, that is, for $\gamma_j = 1$, and non-discriminatory, that is, for $\gamma_j = 0$. However, we allow different sets of covariates to contribute to each feature mean by specifying a *spike-and-slab* prior on each β_{rj} as

$$\beta_{rj} | \delta_{rj}, \sigma^2_{\beta j} \sim (1-\delta_{rj}) \mathrm{I}_0(\beta_{rj}) + \delta_{rj} \mathrm{N}(0, \sigma^2_{\beta j}), \tag{5}$$

with I_0 a point mass distribution at $\beta_{rj} = 0$ and δ_{rj} a binary indicator. If $\delta_{rj} = 1$, then covariate r is considered relevant to explain the observed counts for feature j, and irrelevant otherwise. In the application, this allows us to identify significant associations between feature expression j (RNA-Seq data) and covariate r (DNA methylation), via the selection of the non-zero β_{rj} coefficients, for all discriminatory and non-discriminatory features. We assume independent Bernoulli priors $\delta_{rj} | \omega \sim \mathrm{Bern}(\omega)$, with a beta hyperprior on ω.

We complete the model by imposing inverse-gamma hyperpriors on the hyperparameters σ^2_{0j}, $\sigma^2_{\alpha j}$, and $\sigma^2_{\beta j}$, which leads to marginal non-standardized Student's t-distributions.

2.3. *Model Fitting and Posterior Inference*

We design a Markov chain Monte Carlo (MCMC) algorithm based on stochastic search variable selection algorithms with within-model updates (Savitsky and Vannucci, 2010). Full details can be found in the Supplementary Material.

For posterior inference, our primary interest lies in the identification of the discriminatory features, via the vector $\boldsymbol{\gamma}$, and the selection of the important covariates for each feature, via the vectors $\boldsymbol{\delta}_j$, $j = 1, \ldots, p$. One way to summarize the posterior distribution of the parameters of interest is via *maximum-a-posteriori* (MAP) estimates. Alternatively, selection can be done by thresholding the estimated marginal posterior probabilities of inclusion (PPI) of single features, or covariates, obtained as the proportion of MCMC iterations, after burn-in, in which the corresponding γ_j, or δ_{rj}, are equal to 1. In choosing the threshold, we follow the procedure of Newton et al. (2004), which guarantees the expected Bayesian false discovery rate (BFDR) to be smaller than a pre-specified threshold.

3. Case Study on Breast Cancer

In this section, we illustrate an application of our method to RNA-Seq gene expression and DNA methylation data from a case study on breast cancer. Breast cancer is the most common cancer in women. The NCI currently estimates that 1 in 8 women will be diagnosed with breast cancer during their lifetime, up from 1 in 10 in 1970s (Harbeck and Gnant, 2017). When caught early, the 5-year survival rate of breast cancer is over 90%. The disease becomes deadly when it metastasizes, spreading to other organs or the bones. Cancer stages are defined by tumor size and spread, with higher stages referring to a more severe form of the disease and to different expected survival outcomes. Specifically, stage 0 and I breast cancer diagnoses have a 5-year survival rate close to 100%, stage II of about 93%, stage III of 72%, and, stage IV of 22%.

Here, we analyze data that we downloaded from The Cancer Genome Atlas (TCGA) data portal. In particular, we focused our interest on genes belonging to the 60 KEGG pathways identified by Jiao et al. (2017) as being involved in breast cancer. This resulted in the selection of RNA-Seq data on $p = 1,439$ genes. The dataset comprised $n = 78$ breast cancer patients, of which $n_1 = 19$ were in stage I, $n_2 = 27$ in stage II, $n_3 = 15$ in stage III and $n_4 = 17$ in stage IV. Additionally, data on $R = 29,779$ DNA methylation probes from the same patients were available through TCGA. In the analysis reported here, we focused in particular on associations of genes with DNA methylation probes that map to the same gene. These local associations are called associations in *cis*, as opposed to associations in *trans*, that is, associations of genes with probes that map far from the location of the gene of origin. Even though more than half of the genetically explained variance in gene expression is due to *trans* acting variants, many expression quantitative trait loci (eQTLs) studies have focused on *cis* eQTLs, since reliable detection of *trans* eQTLs has been challenging in humans (Pai et al., 2015). This is due to the smaller effect size of *trans*-acting variants and thus the necessity of a large sample size to establish statistical significance. In our model formulation, the specification of the MRF prior on $\boldsymbol{\gamma}$ requires a gene-gene interaction network. We extracted this network from the KEGG database. Specifically, we used the `R` package `KEGGgraph` of Zhang and Wiemann (2009). Among the genes we selected for our analysis, 256 did not have any neighbors. The resulting network is shown in the Supplementary Material.

For prior specification, we set the hyperparameters that control the MRF prior (3) to $d = -4$ and $f = 1$. Hence, for a gene with no neighbors in the network, or whose neighbors do not discriminate between the four groups, the prior probability of inclusion is $e^{-4}/(1 + e^{-4}) \approx 0.02$. This setting follows our assumption that only a small number of genes are differentially expressed across the different cancer stages. We refer to the simulation study and the sensitivity analysis reported in the Supplementary Material for more details on the choices of these parameters. As for the beta prior on the covariate selection parameters ω_δ, we set $a_\omega = 0.2$ and $b_\omega = 1.8$, which implies a 10% expected prior probability. We used the same flat hyperprior IG(2, 1), as prior choice for σ^2_{0j}, $\sigma^2_{\alpha j}$, and $\sigma^2_{\beta j}$. Finally, we set the prior which controls the dispersion of the negative binomial model to $\phi_j \sim \text{Ga}(a_\phi = 1, b_\phi = 0.01)$, which leads to an expected value of 100 and a relatively large variance (10, 000). We ran 200,000 iterations with 100,000 sweeps as burn-in. Since we noticed that γ was converging slower than δ, we performed multiple updates of γ within each MCMC iteration. As previously noted, we restricted the *add* and *delete* moves of the Metropolis search on δ_{rj} to in *cis* probes. There were on average about 21 probes per gene in the data, and a median of 16.

To assess convergence, we ran four independent MCMC chains. We computed pairwise Pearson correlation coefficients of the marginal posterior probabilities of inclusion between each pair of chains. The values ranged from 0.939 to 0.947 for the γ_j's, and from 0.950 to 0.955 for the δ_{rj}'s. The high correlations indicated substantial agreement between the four MCMC chains. We also used the Gelman and Rubin's convergence diagnostics (Gelman and Rubin, 1992) to inspect for signs of non-convergence of the individual parameters. We found that 95% of those statistics ranged from 1.003 to 1.101, clearly suggesting that the MCMC chains were run for a sufficient number of iterations. Results we report here were obtained by pooling together the outputs from the four chains.

3.1. *Results*

The primary interest of our analysis was to identify differential expression profiles associated to the cancer stages and to study the association of each gene expression with the covariates, that is, the DNA methylation probes. Accordingly, we focus our presentation of the results on the inference on the variable selection parameters, which allow the identification of the discriminating genes and their association with the DNA methylation probes. A plot of the the marginal posterior probabilities of inclusion of $\boldsymbol{\gamma}$, as estimated by our method, is reported in Figure 2. Based on those probabilities, a 5% BFDR threshold corresponded to a cut-off probability of 0.774 and selected 227 genes, while a 1% BFDR threshold corresponded to a cut-off probability of 0.949 and selected 149 genes. We focused on this smaller set of genes for a biological interpretation of the results. Marginal posterior probabilities of associations for this set of 149 genes with their corresponding DNA methylation probes are shown in Figure 2. The dots in the plot indicate the 37 associations, involving 32 genes, as selected with a 5% BFDR threshold. Estimates of the regression coefficients β_{rj} corresponding to the 37 selected associations are also reported in the Figure. A large number of these selected associations are negative. Indeed, for a specific gene, DNA methylation in the promoter is a potent mechanism for silencing gene expression and, thus, a negative association should be expected (Robertson, 2005). However, positive associations have also been found in the literature, especially in the gene body (Yang et al., 2014). We note that, even though it is of greater scientific interest to study whether there are DNA methylation effects which lead to a differential transcription across cancer stages, our method also allows inference on the association with DNA methylation probes for non-discriminating genes.

3.2. *Biological Findings*

In order to assess the biological relevance of our findings, we first focused on the list of 149 selected genes, identified by our method as discriminating features across the 4 cancer stages, and conducted an enrichment analysis by employing

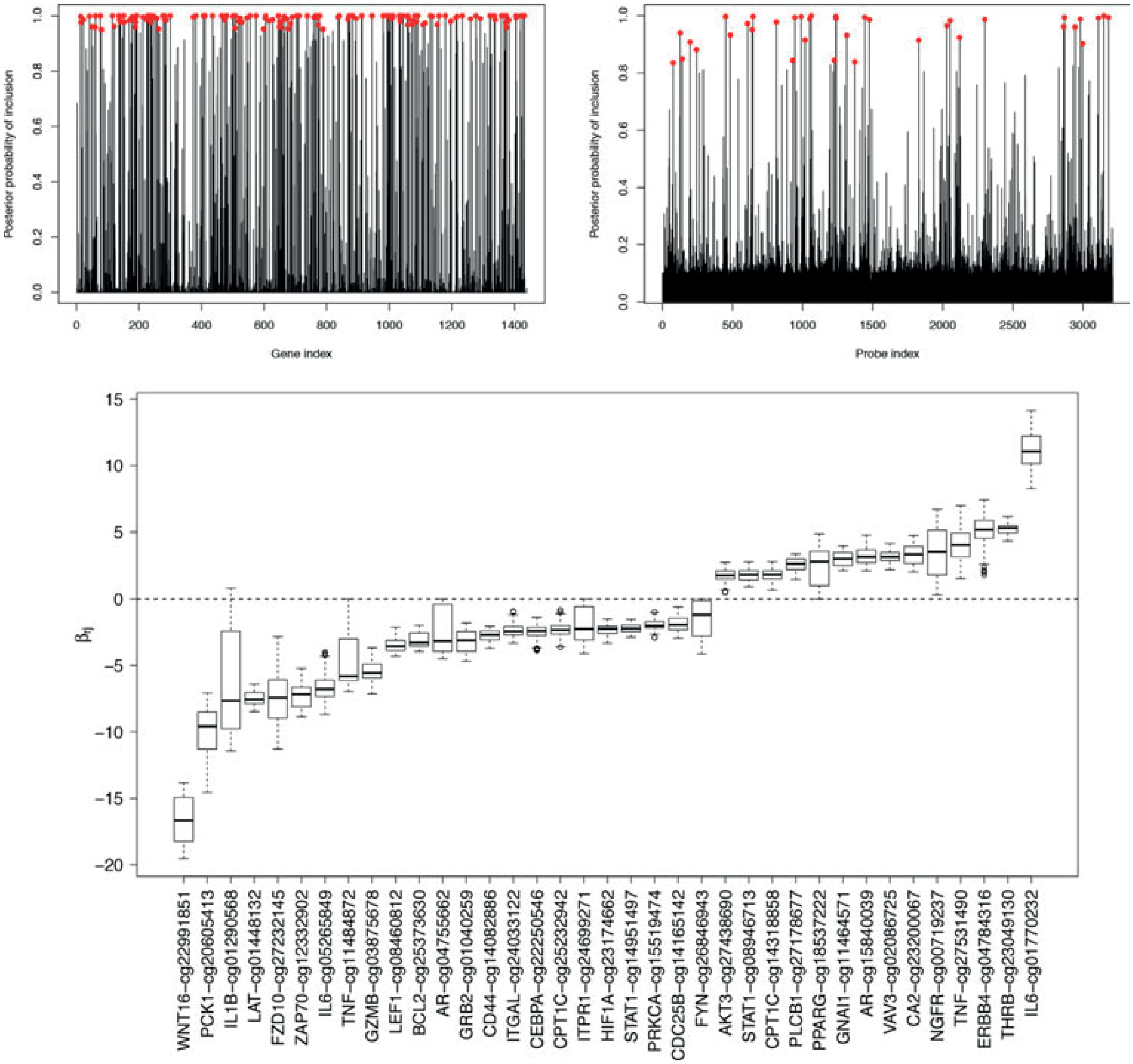


Figure 2. Case study. Top row: Marginal posterior probabilities of inclusion of genes $p(\gamma_j = 1|\cdot)$'s (*left*), with dots indicating the 149 genes selected as discriminatory across cancer stages at a 1% BFDR threshold, and marginal posterior probabilities of inclusion of DNA methylation probes for the 149 selected genes $p(\delta_{rj} = 1|\gamma_j = 1, \cdot)$'s (*right*), with dots marking the 37 probes selected at a 5% BFDR threshold. Bottom row: Boxplots of the strength of the 37 selected gene-probe associations. This figure appears in color in the electronic version of this article, and color refers to that version.

the database for annotation visualization and integrated discovery (DAVID) (Dennis et al., 2003). We selected significant annotation terms at a 0.05 threshold applied to corrected p-values (Benjamini and Hochberg, 1995). Several significant terms were identified. For example, under the Pathways category, focusing on the KEGG pathway sub-category, the term Pathways in cancer showed the smallest p-value (2.5×10^{-32}), with the terms Neurotrophin signaling pathway and T cell receptor signaling pathway following with the second and third smallest p-values (5.0×10^{-18} and 2.9×10^{-13}, respectively). The gene-gene networks of these 3 pathways are depicted in Figure 3. Genes highlighted in red correspond to those selected by our method as discriminatory of cancer stages, at a 1% BFDR cut-off level, and those highlighted in green are genes selected as discriminatory that, in addition, show significant associations with DNA methylation. We comment on interesting findings from these pathways in the Supplementary Material. Additionally, it is of great scientific interest to identify biomarkers that can help understanding the effect of DNA methylation on gene transcription. Our method has the desirable feature to be able to pinpoint at specific CpG sites which may play a principal role in the epigenetic mutation. Accordingly, we performed a literature search on the 32 genes that showed significant association with at least one DNA probe (see Figure 2). Results confirmed the relevance of our results, see the Supplementary Material. Finally, we conducted a comparative evaluation of the biological results obtained by our method and those from

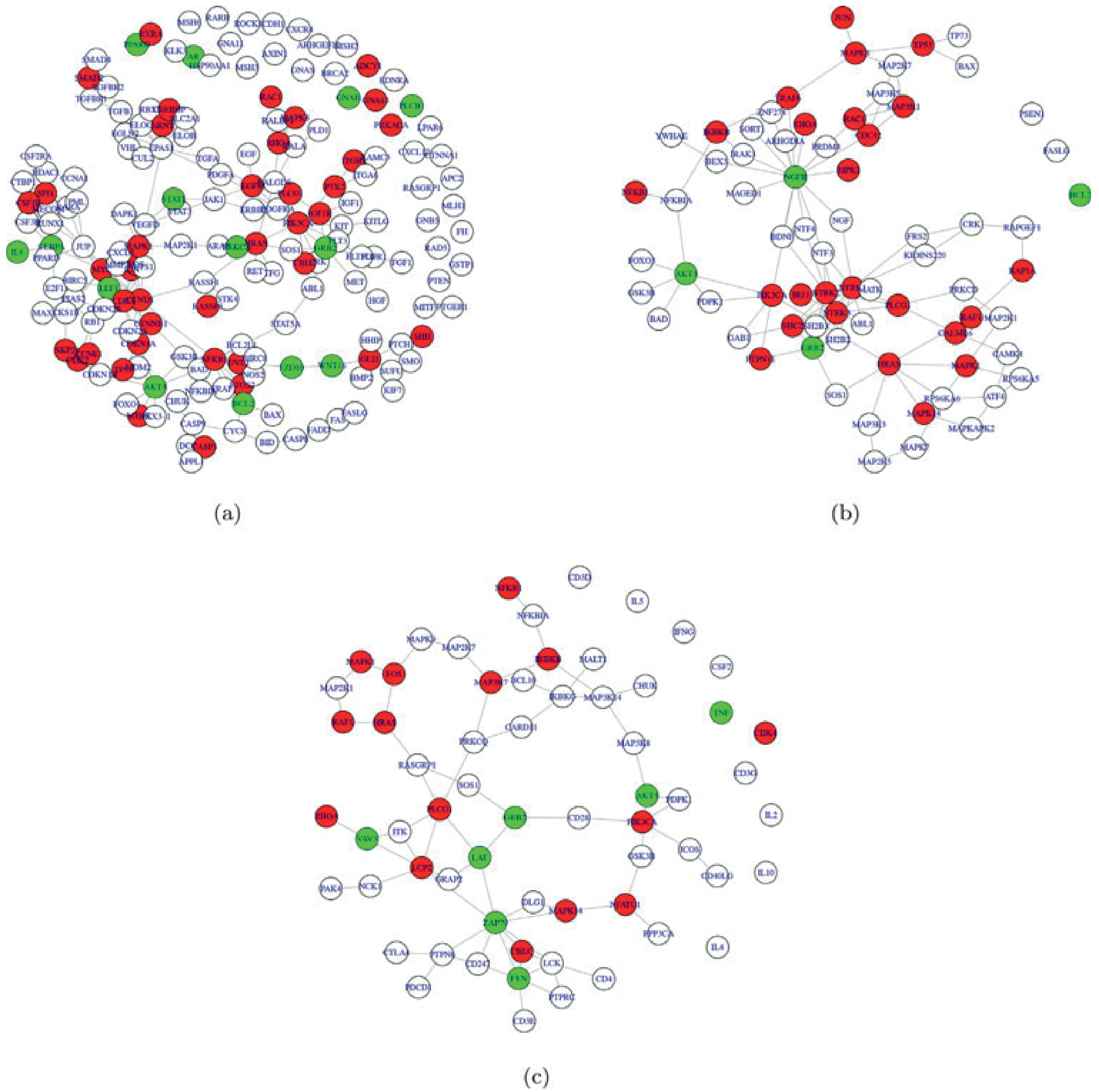
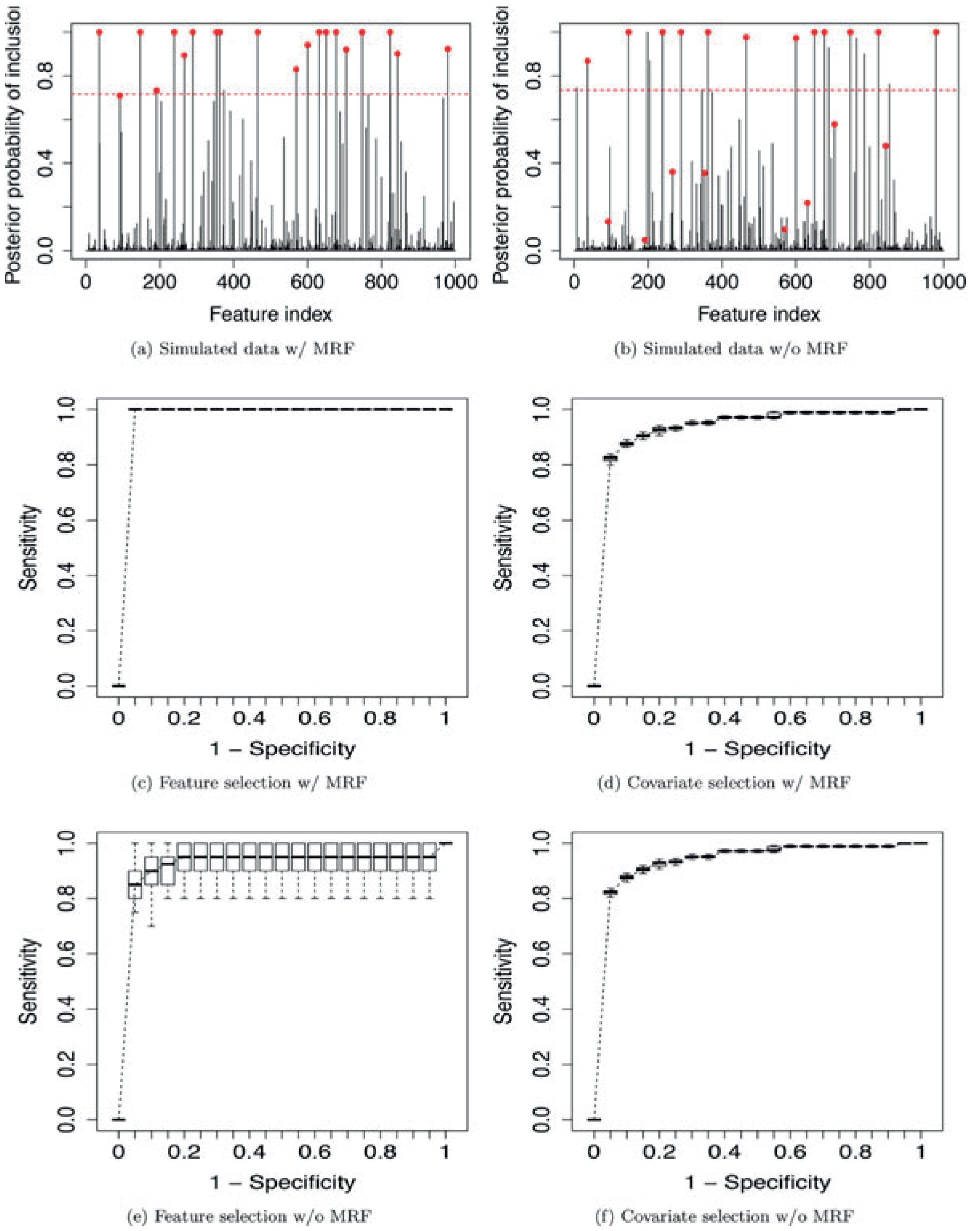


Figure 3. Case study: The gene-gene interaction networks for the three most enriched pathways: (a) Pathways in cancer, (b) Neurotrophin signaling pathway and (c) T cell receptor signaling pathway. The colored genes were selected by our method as discriminatory across cancer stages using a 1% BFDR threshold, with the green ones were those genes showed significant association to DNA methylation. This figure appears in color in the electronic version of this article, and color refers to that version.

alternative approaches, based on findings from the enrichment analyses. Such comparative study, reported in the Supplementary Material, showed an increased ability of our method to identify groups of genes belonging to pathways clearly related to cancer.

We conclude by comparing our results with those obtained using an independent Bernoulli prior on $\boldsymbol{\gamma}$, specifically $\gamma_j \sim \mathrm{Bern}(0.02)$. With a 1% BFDR threshold, the Bernoulli prior led to the selection of 24 discriminating genes, 18 of which were in the list obtained by the model with the MRF prior. Out of the 6 genes selected by the Bernoulli prior only, a literature search did not report any association to cancer for 3 of them. Additionally, an enrichment analysis on the whole set of 24 genes selected with the Bernoulli prior found only one statistically significant term. This was in contrast with the many cancer-related terms identified when employing the MRF prior, as described in the Supplementary Material. In summary, employing a KEGG-informed MRF prior aided the interpretation of the results and increased the ability to identify discriminating genes. This, however, does not affect the selection of gene expression-methylation associations since the model assumes that the DNA methylation effect on RNA-Seq gene expression is the same across different conditions.



(a) Simulated data w/ MRF

(b) Simulated data w/o MRF

(c) Feature selection w/ MRF

(d) Covariate selection w/ MRF

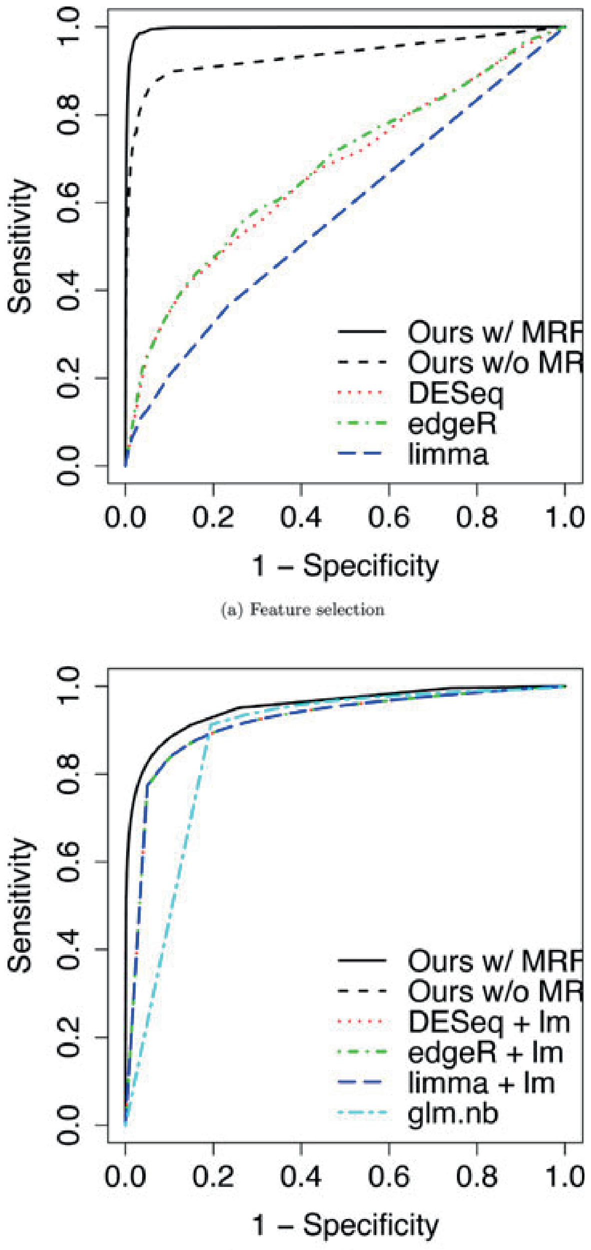
(e) Feature selection w/o MRF

(f) Covariate selection w/o MRF

Figure 4. Simulated data: Marginal posterior probabilities of inclusion $p(\gamma_j = 1|\cdot)$ with dots indicating truly discriminatory features and the horizontal dotted lines indicating a threshold for a 5% BFDR, using (a) the MRF prior and (b) the independent Bernoulli prior. Receiver operating characteristic (ROC) curves on the posterior probabilities of inclusion for features and covariates, for different values of the threshold and 30 replicated datasets, using (c)(d) the MRF prior and (e)(f) the independent Bernoulli prior. This figure appears in color in the electronic version of this article, and color refers to that version.

4. Simulation Studies

In this section we use simulated data to assess the performance of our model against alternative solutions and to investigate the sensitivity to the prior choices. The dimension of $\boldsymbol{Y}$ was set to $n = 78$ and $p = 1{,}000$. In order to test the ability of our method to discover relevant features in the presence of a good amount of noise, we focused on a scenario where only a few of the features were truly discriminatory ($p_\gamma = 20$). We generated a matrix $\boldsymbol{X}$ with $n = 78$ rows (i.e., samples) and $R = 5000$ columns (i.e., covariates) by sampling each element from a beta distribution $\mathrm{Be}(0.4, 0.6)$. As for the coefficient matrix $\boldsymbol{B}$, we arbitrarily set 10% of the entries to non-zero values that we generated from a mixture of two uniform distributions $0.7 \times \mathrm{U}(1,2) + 0.3 \times \mathrm{U}(-2,-1)$. In generating the response $\boldsymbol{Y}$, we assumed $K = 4$ groups for those counts that map to the 20 discriminatory features. Furthermore, we induced correlation structure among the discriminatory features as follows. We first generated the vectors $\boldsymbol{\lambda}_i^{(\gamma)}$, with (γ)



(a) Feature selection

(b) Covariate selection

Figure 5. Simulated data: Comparison between ROC curves for our methods and those obtained using four competing methods for different p-values. All curves are averaged over 30 replicated datasets. Results in plots (a) refer to the identification of discriminating features of sample groups, and those in plot (b) to the selection of significant associations to the covariates. This figure appears in color in the electronic version of this article, and color refers to that version.

indicating the elements of $\boldsymbol{\lambda}_i$ that correspond to $\gamma_j = 1$, from a mixture of four ($K = 4$) multivariate normal densities

$$\log \boldsymbol{\lambda}_i^{(\gamma)} \sim \boldsymbol{x}_i^T \boldsymbol{B}^{(\gamma)} + \mathrm{I}(1 \leq i \leq 19)\mathrm{MN}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + \mathrm{I}(20 \leq i \leq 46) \mathrm{MN}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) + \mathrm{I}(47 \leq i \leq 61)\mathrm{MN}(\boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3) + \mathrm{I}(62 \leq i \leq 78)\mathrm{MN}(\boldsymbol{\mu}_4, \boldsymbol{\Sigma}_4),$$

where the first 19 observations were drawn from the first distribution, the second 27 from the second distribution, the next 15 from the third and the last 17 from the fourth. We set the means of the multivariate normal distributions equal to $\boldsymbol{\mu}_1 = 3 \times \mathbf{1}_{p_\gamma}$, $\boldsymbol{\mu}_2 = 5 \times \mathbf{1}_{p_\gamma}$, $\boldsymbol{\mu}_3 = 4.5 \times \mathbf{1}_{p_\gamma}$, $\boldsymbol{\mu}_4 = 3.5 \times \mathbf{1}_{p_\gamma}$, where $\mathbf{1}_{p_\gamma}$ is a unit vector of dimension p_γ. We constructed the four covariance matrices of the 20 discriminatory features by first setting the diagonal elements to $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 2$ and $\sigma_4^2 = 3$. We then induced correlation among the 20 features by setting 0.5 to some of the off-diagonal elements of the four covariance matrices. This induced a network among the 20 features, with 4 features connected to two other features, 10 connected to three others and 6 connected to four others. For the 980 noisy features, we generated $\boldsymbol{\lambda}_i^{(\gamma^c)}$, with (γ^c) indicating the elements of $\boldsymbol{\lambda}_i$ that correspond to $\gamma_j = 0$, as $\boldsymbol{\lambda}_i^{(\gamma^c)} \sim \boldsymbol{x}_i^T \boldsymbol{B}^{(\gamma^c)} + \mathrm{MN}(4 \times \mathbf{1}_{p-p_\gamma}, \boldsymbol{\Sigma}_0)$, with the diagonal and the off-diagonal elements of the covariance matrix set to 1 and 0.1, respectively. Finally, we simulated the count data from Negative Binomial distributions with parameters λ_{ij} and ϕ_j with $\phi_j \sim \mathrm{Exp}(1/10)$.

We used the same prior and algorithm settings as described in Section 3. Figure 4 shows the marginal posterior probabilities of inclusion of single features when using the MRF prior with $(d = -4, f = 1)$ and the independent Bernoulli prior with $\omega = 0.02$. The dots indicate the truly discriminatory features and the horizontal dotted lines correspond to a threshold that ensures an expected Bayesian false discovery rate (BFDR) of 5%. This threshold resulted in a model that included 21 features with the MRF prior, 19 of which were in the set of truly discriminatory features, and 20 features with the independent prior, 12 of which were in the set of truly discriminatory features. We also observed that the PPIs of the discriminatory features obtained with the MRF prior were generally higher than those obtained with the independent Bernoulli prior, a finding which suggests that employing the MRF prior results in an increased ability to identify features with strong discriminatory power. Other authors have reported similar results, see for example Stingo and Vannucci (2011) and Li and Zhang (2010). Generally speaking, the effect of the MRF prior depends on the concordance of the prior network with the data, while an independent prior, obtained for $f = 0$, is expected to lead best results if features are independent. Figure 4 also reports receiver operating characteristic (ROC) curves on the posterior probabilities of inclusion of features and covariates under the two prior settings, averaged over 30 simulated datasets. These plots confirm the overall best performance of our method with the MRF prior. The area under curve (AUC) is 0.998 and 0.969, with and without MRF prior, respectively, for the discriminating features, and 0.979 and 0.979, respectively, for selection of the covariates.

Unlike existing approaches, which typically employ multistep analyses, a novel characteristic of our integrative modeling approach is the simultaneous identification of discriminating features and their associations to covariates. In setting up a comparison study, we therefore considered two-stage methods that first identify discriminatory features across sample groups and then regress them on the covariates. For the first stage, we employed three popular differential gene expression (DGE) analysis methods implemented by the R packages DESeq2 (Love et al., 2014), edgeR (McCarthy et al., 2012), and limma (Ritchie et al., 2015). The former two methods rely on the negative binomial distribution to model the over-dispersed raw counts. The latter one employs a GLM approach, assuming that counts are from a log-normal distribution. All three methods produced thresholded p-values, to control for false discover rate (FDR). In the second stage, we used the lm() function in R to calculate p-values. Specifically, for each competing method, DESeq2, edgeR and limma, we first identified those discriminating features whose p-values were smaller than 0.05. Then, for non-discriminating features, we centered the log-transformed counts to their mean across all the samples, for each feature, and then applied lm() to obtain the p-values for each covariate. For the discriminating features, we centered the log-transformed counts to their corresponding group means and then applied lm(). As an additional comparison, we also fit a negative binomial regression to each feature, that is, ignoring the mixture related to cancer stages on the intercept term, by applying glm.nb() to the raw counts.

Figure 5 shows the ROC curves obtained with DESeq2, edgeR and limma for different values of the threshold on the p-values and averaged over 30 replicated datasets, together with those from our methods with and without MRF prior. For the identification of the discriminating features, the AUCs were 0.659, 0.692, and 0.637, for DESeq2, edgeR, and limma, respectively. For the selection of the covariates, the AUCs were 0.952, 0.952 and 0.953, for DESeq2, edgeR and limma, respectively. A negative binomial regression fitted to the raw counts of each feature resulted in a AUC of 0.891. Our method therefore showed the best performance.

Results on the sensitivity of our model to different prior choices are reported in the Supplementary Material. These show that the model is considerably robust to the choice of d, when fixing f. As for the choice of f, even though the sensitivity analysis suggested increased performances for larger values, it is a general experience that allowing this parameter to vary widely can lead to a phase transition problem, that is, the expected number of variables equal to 1 can increase massively for small increments of f. This may lead to a drastic change in the proportion of selected genes. Thus, in our analyses we have set $f = 1$ as a conservative choice. As for the parameters (a, b) of the inverse-gamma hyperpriors on σ^2_{0j}, $\sigma^2_{\alpha j}$, and $\sigma^2_{\beta j}$, results show little sensitivity. In the applications of the article, we have set a to 2, as this is the smallest number such that the variance of the inverse-gamma is well-defined.

5. Conclusion

We have presented a Bayesian hierarchical mixture regression model for studying the association between a multivariate response, measured as counts on a set of features, and a set of covariates. Our motivation has come from the analysis of RNA-Seq and DNA methylation data from a breast cancer study. We have employed a mixture of negative binomial distributions, incorporating the covariates via a linear construction on the mean components. Our proposed approach allows a simultaneous selection of discriminatory features and relevant covariates. We have also incorporated structural dependencies among genes via the use of Markov random field priors. Our results have identified several biomarkers that can help understanding the effect of DNA methylation on gene transcription. We have also demonstrated improved performances over alternative approaches.

Several extensions of our model are worth investigating. First, even though in our application only 2.6% counts were zeros, RNA-Seq count data can sometimes have an excess of zeros, because of insufficient sequencing depth or a large amount of short RNAs. This feature can be taken into account by considering zero-inflated models. Next, our approach can be extended to infinite mixture models that cluster the observations, for the discovery of cancer sub-types, and estimate the number of clusters directly from the data (Muller et al., 2015). Also, in the data analysis, we restricted our attention to associations of genes with DNA methylation probes that map to the same gene. Other interesting analyses are possible, for example by considering pre-defined lists of promoter regions or data driven choices, such as regions of dense CpG clusters. When appropriate, correlation structures among methylation sites can be incorporated, for example, via selection priors similar to those used in Cassese et al. (2014), that allow for associations between individual genes and groups of correlated probes. Finally, model-based methods that characterize base-level RNA-Seq reads within each transcript (Hu et al., 2011; SUN et al., 2013) could be considered, to obtain a more accurate quantification of transcript-level data, and joint models of RNA-seq and DNA-methylation could be used for simultaneous inference on both types of variables/processes.

6. Supplementary Materials

Web appendices, tables, and figures referenced in Sections 2, 3, and 4 are available with this article at the *Biometrics* website on Wiley Online Library. Software coded in R/C++ is available at https://github.com/liqiwei2000/BayesNBMixReg.

References

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology* **11**, R106.

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* **57**, 289–300.

Bullard, J. H., Purdom, E., Hansen, K. D., and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* **11**, 94.

Cassese, A., Guindani, M., Tadesse, M., Falciani, F., and Vannucci, M. (2014). A hierarchical Bayesian model for inference on copy number variants and their association to gene expression. *Annals of Applied Statistics* **8**, 148–175.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., et al. (2016). A survey of best practices for RNA-Seq data analysis. *Genome Biology* **17**, 13.

Dennis, G., Sherman, B., Hosack, D., Yang, J., Gao, W., Lane, H., et al. (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biology* **4**, P3+.

Ferrón, S. R., Charalambous, M., Radford, E., McEwen, K., Wildner, H., Hind, E., et al. (2011). Postnatal loss of Dlk1 imprinting in stem cells and niche astrocytes regulates neurogenesis. *Nature* **475**, 381–385.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science* **7**, 457–472.

Hammersley, J. M. and Clifford, P. (1971). Markov fields on finite graphs and lattices. https://www.bibsonomy.org/bibtex/2199eae8bccbe1a7a1bffb1992ba0b394/brefeld

Hansen, K. D., Irizarry, R. A., and Zhijin, W. (2012). Removing technical variability in RNA-Seq data using conditional quantile normalization. *Biostatistics* **13**, 204–216.

Harbeck, N. and Gnant, M. (2017). Breast cancer. *The Lancet* **389**, 1134–1150.

Hu, M., Zhu, Y., Taylor, J. M., Liu, J. S., and Qin, Z. S. (2011). Using Poisson mixed-effects model to quantify transcript-level gene expression in RNA-Seq. *Bioinformatics* **28**, 63–68.

Jiao, Y., Hidalgo, M. R., Cubuk, C., Amadoz, A., Carbonell-Caballero, J., Vert, J.-P., and Dopazo, J. (2017). Signaling pathway activities improve prognosis for breast cancer. *bioRxiv*. https://www.biorxiv.org/content/early/2017/04/29/132357

Kukurba, K. R. and Montgomery, S. B. (2015). RNA sequencing and analysis. *Cold Spring Harbor Protocols* **2015**, **11**, 951–969.

Lee, J., Ji, Y., Liang, S., Cai, G., and Müller, P. (2015). Bayesian hierarchical model for differential gene expression using RNA-Seq data. *Statistics in Biosciences* **7**, 48–67.

Leng, N., Li, Y., McIntosh, B. E., Nguyen, B. K., Duffin, B., Tian, S., et al. (2015). Ebseq-hmm: A Bayesian approach for identifying gene-expression changes in ordered RNA-Seq experiments. *Bioinformatics* **31**, 2614–2622.

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

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *Genome Biology* **15**, 550.

Ma, X., Liu, Z., Zhang, Z., Huang, X., and Tang, W. (2017). Multiple network algorithm for epigenetic modules via the integration of genome-wide DNA methylation and gene expression data. *BMC Bioinformatics* **18**, 72.

McCarthy, D. J., Chen, Y., and Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research* **40**, 4288–4297.

Muller, P., Quintana, F. A., Jara, A., and Hanson, T. (2015). *Bayesian Nonparametric Data Analysis.* Switzerland: Springer International Publishing.

Murrell, A., Rakyan, V. K., and Beck, S. (2005). From genome to epigenome. *Human Molecular Genetics* **14**, R3–R10.

Newton, M. A., Noueiry, A., Sarkar, D., and Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* **5**, 155–176.

Pai, A., Pritchard, J., and Gilad, Y. (2015). The genetic and mechanistic basis for variation in gene regulation. *PLoS Genetics* **11**, e1004857.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* **43**, e47–e47.

Robertson, K. D. (2005). DNA methylation and human disease. *Nature Reviews Genetics* **6**, 597.

Robinson, M., McCarthy, D., and Smyth, G. (2010). EdgeR: A bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140.

Savitsky, T. and Vannucci, M. (2010). Spiked dirichlet process priors for Gaussian process models. *Journal of Probability and Satistics* **2010**, 201489.

Stingo, F. C. and Vannucci, M. (2011). Variable selection for discriminant analysis with Markov random field priors for the analysis of microarray data. *Bioinformatics* **27**, 495–501.

Sun, Z., Wu, H., Qin, Z., and Zhu, Y. (2013). Model-based methods for transcript expression-level quantification in RNA-Seq. *Advances in Statistical Bioinformatics: Models and Integrative Inference for High-Throughput Data*. Cambridge University Press.

Tang, B., Zhou, Y., Wang, C.-M., Huang, T. H.-M., and Jin, V. X. (2017). Integration of DNA methylation and gene transcription across nineteen cell types reveals cell type-specific and genomic region-dependent regulatory patterns. *Scientific Reports* **7**, 1–11.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10**, 57–63.

Witten, D. M. (2011). Classification and clustering of sequencing data using a Poisson model. *The Annals of Applied Statistics* **5**, 2493–2518.

Xie, L., Weichel, B., Ohm, J. E., and Zhang, K. (2011). An integrative analysis of DNA methylation and RNA-Seq data for human heart, kidney and liver. *BMC Systems Biology* **5**, 1.

Yang, X., Han, H., De Carvalho, D. D., Lay, F. D., Jones, P. A., and Liang, G. (2014). Gene body methylation can alter gene expression and is a therapeutic target in cancer. *Cancer cell* **26**, 577–590.

Zhang, J. D. and Wiemann, S. (2009). KEGGgraph: A graph approach to KEGG pathway in R and Bioconductor. *Bioinformatics* **25**, 1470–1471.

Received September 2017. *Revised May* 2018.
Accepted July 2018.