Biostatistics (2018) **19**, 1, *pp.* 71–86
doi:10.1093/biostatistics/kxx017
Advance Access publication on May 24, 2017

A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data

QIANXING MO*

Division of Biostatistics, Dan L. Duncan Cancer Center, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA

qmo@bcm.edu

RONGLAI SHEN

Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, 485 Lexington Ave, New York, NY 10017, USA

CUI GUO

Department of Biostatistics, University of Michigan, 1415 Washington Heights, Ann Arbor, MI 48109, USA

MARINA VANNUCCI

Department of Statistics, Rice University, 6100 Main Street, Houston, TX 77030, USA

KEITH S. CHAN

Molecular & Cellular Biology/Scott Department of Urology, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA

SUSAN G. HILSENBECK

Division of Biostatistics, Dan L. Duncan Cancer Center, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA

Summary

Identification of clinically relevant tumor subtypes and omics signatures is an important task in cancer translational research for precision medicine. Large-scale genomic profiling studies such as The Cancer Genome Atlas (TCGA) Research Network have generated vast amounts of genomic, transcriptomic, epigenomic, and proteomic data. While these studies have provided great resources for researchers to discover clinically relevant tumor subtypes and driver molecular alterations, there are few computationally efficient methods and tools for integrative clustering analysis of these multi-type omics data. Therefore, the aim of this article is to develop a fully Bayesian latent variable method (called iClusterBayes) that can jointly model omics data of continuous and discrete data types for identification of tumor subtypes and relevant omics features. Specifically, the proposed method uses a few latent variables to capture the inherent structure of multiple omics data sets to achieve joint dimension reduction. As a result, the tumor samples can be clustered in the latent variable space and relevant omics features that drive the sample

*To whom correspondence should be addressed.

clustering are identified through Bayesian variable selection. This method significantly improve on the existing integrative clustering method iClusterPlus in terms of statistical inference and computational speed. By analyzing TCGA and simulated data sets, we demonstrate the excellent performance of the proposed method in revealing clinically meaningful tumor subtypes and driver omics features.

Keywords: Multi-type omics data; Integrative clustering; iCluster; iClusterPlus; iClusterBayes; Latent variable model; Bayesian variable selection.

1. INTRODUCTION

Cancer is a complex and heterogeneous disease, driven by alterations occurring at multiple levels, namely chromosomal rearrangements, epigenetic changes, somatic mutations, and gene expression. Tumors with similar histopathologic phenotypes but diverse genomic profiles could respond differently to the same treatment, leading to distinct clinical outcomes. Therefore, there is a clinical need to classify tumors into molecular subtypes and to identify driver (causative) molecular alterations that could be targeted for precision medicine. In an effort to comprehensively catalog transcriptomic, genomic and epigenomic alterations of cancers, national and international consortia such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) have been established to perform large-scale genomic profiling studies, which have been generating an unprecedented amount of data with multiple layers of genomic and genetic information for a variety of cancers. While these large-scale genomic studies have provided integrative data for researchers to discover tumor subtypes and their genomic signatures, most of the clustering analyses are performed at the single data set level and then integrated manually, which is partially due to the relatively slower development of integrative statistical methods and efficient software. It has become increasingly clear that separate analyses of individual genomic data set and a *post hoc* integration of the clustering results have difficulty capturing the correlated structure of cancer omics data and thus much of the potential for new insight might have gone unrealized. Therefore, there is a great need for developing statistical methods and tools for integrative analysis of omics data from multiple sources.

In an effort to identify clinically relevant tumor subtypes from TCGA data, Shen *and others* (2009) developed a method called iCluster that uses a Gaussian latent variable model to jointly model continuous genomic data such as gene expression, DNA methylation and copy number data. The latent variables in the iCluster model form a set of principle coordinates spanning a low dimensional subspace that can collectively capture the correlative structure of multi-omics data, and thus can be used for tumor sample clustering and integrated visualization. The iCluster model assumes that each data set is conditionally independent given the latent variables. An optimal number of integrative clusters can be found when the joint likelihood of the genomic data sets is maximized. In the iCluster model, an expectation-maximization (EM) algorithm is used for parameter estimation and a soft-thresholding method is used to induce sparsity in the model parameters in order to achieve better clustering results and distinguish informative features from non-informative features. Shen *and others* (2013) further extended the iCluster model by employing lasso, elastic net and fused lasso methods to allow feature selection in integrative clustering context. The overall aims of the iCluster method are to obtain a joint clustering of samples and identify cluster-relevant features across data sets. Besides the iCluster method, a few other model-based methods have been developed for integrative clustering analysis although their aims may not be exactly the same as the iCluster method. For example, Kormaksson *and others* (2012) developed a finite mixture model to perform joint clustering of gene expression and DNA methylation data. In contrast to the iCluster method, the Multiple Dataset Integration (MDI) developed by Kirk *and others* (2012) does not seek to find joint sample clusters. Instead, it allows different data sets to have different numbers of clusters and the clustering of genes in one data set is influenced by the clustering in another data set. The MDI method focuses more on identifying

clusters of genes with shared characteristics in genomic data such as chromatin immunoprecipitation-chip and microarray gene expression. Lock and Dunson (2013) developed the Bayesian consensus clustering (BCC) method to perform both data-specific clustering and consensus clustering. In the BCC framework, individual clustering is not independent and adhere loosely to an overall clustering. The iCluster and MDI methods can perform clustering and feature selection, but the other two aforementioned methods only perform clustering.

The iClusterPlus method developed by Mo *and others* (2013) is a significant enhancement of the iCluster method that can jointly model multi-type omics data including continuous, count, binary, and multi-categorical data. In the iClusterPlus framework, there is no closed-form solution for estimation of the model parameters. Therefore, a modified Monte Carlo Newton–Raphson algorithm is used to estimate the model parameters and lasso penalty is used to induce sparse estimation. The iClusterPlus method has been used to generate biologically meaningful cancer subtypes in large-scale cancer genomic studies including squamous cell lung cancers, lung adenocarcinoma, endometrial carcinoma and gastric adenocarcinoma studies (TCGA, 2012; 2013; 2014; 2014). Although the iClusterPlus method is widely adopted by research communities, there are two limitations in the statistical model that could limit its usage. First, in order to find an optimal solution for parameter estimation and subtype identification, it needs to tune the model parameters by testing hundreds of λ (Lasso shrinkage parameter) values for a few integrated data sets. This requires a lot of computation because it needs to run through all the computationally intensive steps of the modified Monte Carlo Newton–Raphson algorithm for each λ. Second, there is no evaluation of statistical significance for the selected features. A feature is selected if its associated parameter is not 0, according to the Lasso method (Tibshirani, 1996).

To address the challenges in integrative analysis of cancer omics data, we have developed a fully Bayesian integrative clustering method named iClusterBayes to model continuous and discrete omics data. This new method overcomes the limitations of the iClusterPlus method and is a valuable tool for cancer research. We organize this article as follows. In Section 2, we provide details for the iClusterBayes framework. In Section 3, we perform simulations to examine the performance of the proposed method. In Section 4, we show that the iClusterBayes method can reveal clinically meaningful cancer subtypes by analyzing TCGA glioblastoma and kidney cancer data. In Section 5, we summarize this article with a brief discussion.

2. Methods

2.1. *The integrative clustering framework*

The proposed method is designed to integrate continuous and discrete data, which represent the major forms of omics data. Figure 1 shows the core idea of the joint integrative clustering framework. Suppose we have n tumor samples and each of them is analyzed by m different types of techniques generating m sets of genomic data. For example, microarray-based platforms can generate continuous data for gene expression, DNA copy number and CpG site methylation, while sequencing-based platforms can generate count data for gene expression and binary data for DNA mutation. Let $\mathbf{y}_{it} = (y_{i1t}, \cdots, y_{ip_tt})^T$ denote a p_t-dimensional data vector. Each element y_{ijt} denotes the observed data associated with the jth ($j \in \{1, \cdots, p_t\}$) genomic feature in the ith ($i \in \{1, \cdots, n\}$) sample of the tth ($t \in \{1, \cdots, m\}$) data type. Depending on the type of data, a genomic feature can be a protein coding gene or non-gene-centric elements of interest (e.g., genomic region/location, CpG sites, mRNA, etc.). The tth data set can be represented by $\mathbf{Y}_t = (\mathbf{y}_{1t}, \cdots, \mathbf{y}_{nt})$, a matrix with dimension $p_t \times n$, and all the data sets can be represented by a multi-high dimensional genomic data space $\{\mathbf{Y}_t\}_{t=1}^m$. Assume there are $k+1 (k > 0)$ molecular subtypes in the tumor samples, and we hope to identify them from the multi-high dimensional data space $\{\mathbf{Y}_t\}_{t=1}^m$.

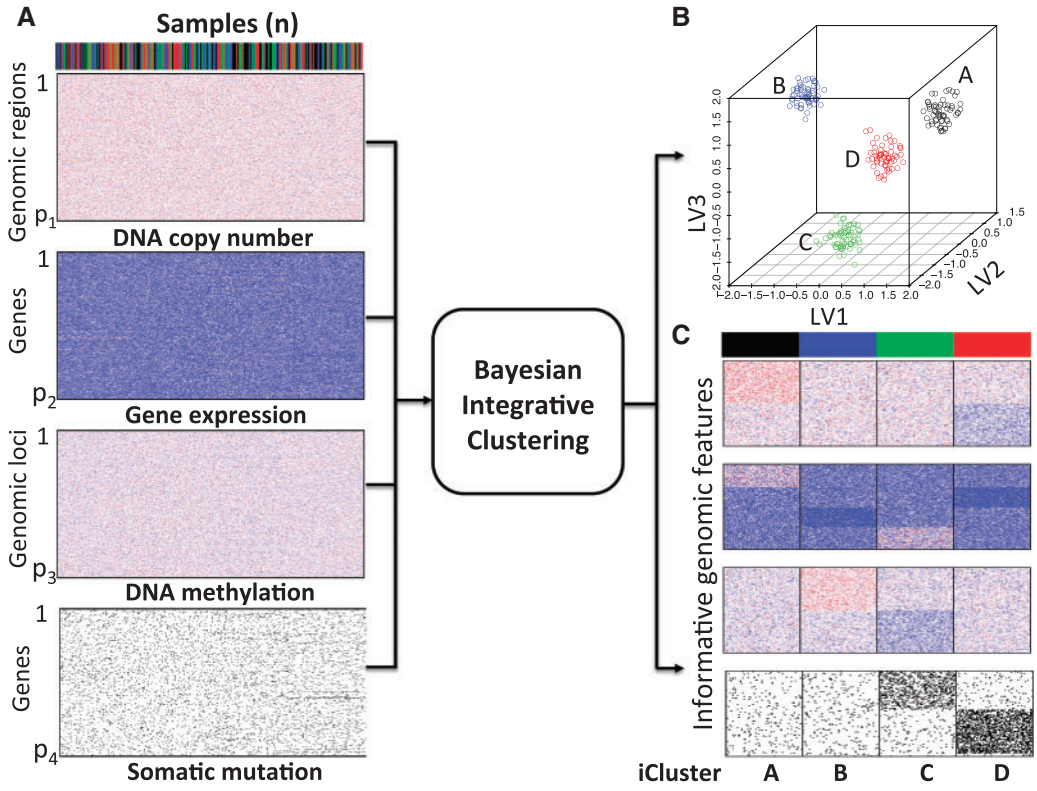

Fig. 1. The proposed Bayesian integrative clustering framework. Bayesian latent variable regression models are used to jointly model multiple genomic data sets (A) to identify common latent variables that can be used to cluster patient samples in a lower dimensional integrated latent variable space (B). Simultaneously, driver genetic/genomic features (e.g., DNA copy number, gene expression, DNA methylation and somatic mutation) that contribute to sample clustering are identified (C).

The core idea of the integrative clustering framework is to reduce the multi-high dimensional space to a low dimensional subspace that will collectively capture the major variations of the multiple genomic data sets. Therefore, the low-dimensional subspace can be used to cluster the tumor samples. This is similar to principle component analysis (PCA) when there is only one data type in the analysis. In PCA analysis, the high dimensionality of the data is reduced to a low-dimensional space that is represented by a set of new variables called principal components. Usually, the first few principal components (PCs) capture most of the variation in the original data set and thus they can be used to cluster the samples. Different profiling platforms could generate different types of genomic data. The units of measurement vary from one profiling platform to another, and thus it is not appropriate to directly pool them for PCA-like analyses. Borrowing the idea from PCA, we assume that we could project the multi-high dimensional space $\{\mathbf{Y}_t\}_{t=1}^m$ to a low-dimensional integrated subspace $\mathbf{Z}$ with dimension $n \times k$. In other words, we can say that each sample is associated with a latent variable $\mathbf{z}_i = (z_{i1}, \cdots, z_{ik})$, $(i \in \{1, \cdots, n\})$. We assume that $\mathbf{z}_i$ is a continuous variable and follows a standard multivariate normal distribution $\text{MVN}(\mathbf{0}, \mathbf{I}_k)$. The mean zero and identity covariance matrix are necessary constraints to make sure the proposed joint statistical model is identifiable. If there are $k + 1 (k > 0)$ molecular subtypes in the tumor samples, we can separate the samples using the latent variable $\mathbf{z}_i$ $(i \in \{1, \cdots, n\})$.

2.2. *Statistical model for continuous data*

Microarray-based technologies typically generate continuous data, which are measurements of signal intensities of probe-target hybridization on the microarrays. Through proper transformation, these data can be appropriately normal. Therefore, we first lay out our statistical framework for omics data with continuous measures, which is

$$y_{ijt} = \mathbf{x}_i \boldsymbol{\Gamma}_{jt} \boldsymbol{\beta}_{jt} + \varepsilon_{ijt},\ i = 1, \cdots, n,\ j = 1, \cdots, p_t,\ t \in (1, \cdots, m), \tag{2.1}$$

where $\boldsymbol{\beta}_{jt} = \left(\beta_{0jt}, \beta_{1jt}, \cdots, \beta_{kjt}\right)^T$ is the coefficient vector associated with the jth feature in the tth data set; $\mathbf{x}_i = (1, \mathbf{z}_i) = (1, z_{i1}, \cdots, z_{ik})$, is a vector in which the first component is 1 and the remaining components are exactly from vector $\mathbf{z}_i$; $\boldsymbol{\Gamma}_{jt} = \text{diag}(1, \gamma_{jt}, \cdots, \gamma_{jt})$ is a diagonal matrix whose first diagonal component is 1 and all the remaining k diagonal components are γ_{jt}; ε_{ijt} is the random error with mean 0 and variance σ^2_{jt}. The constant 1 in $\mathbf{x}_i$ and $\boldsymbol{\Gamma}_{jt}$ is designed to let the model (2.1) have the intercept β_{0jt}. In the model, γ_{jt} is an indicator variable with value 0 or 1, which is used for Bayesian variable selection (George and McCulloch, 1997). When $\gamma_{jt} = 0$, it indicates that the corresponding $\boldsymbol{\beta}_{jt}$ is small and thus the jth omics feature in the tth data set contributes little for the joint clustering. When $\gamma_{jt} = 1$, it indicates that the corresponding $\boldsymbol{\beta}_{jt}$ is large and thus the jth omics feature in the tth data set is a contributor for the joint clustering. Let $\mathbf{y}_{jt} = (y_{1jt}, \cdots, y_{njt})^T$ be the data vector of omics feature j for samples 1 to n. Let $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)^T$ be the design matrix with dimension $n \times (k+1)$, where row i is $\mathbf{x}_i$. The model for omics feature j in data set t can be written as

$$\mathbf{y}_{jt} = \mathbf{X} \boldsymbol{\Gamma}_{jt} \boldsymbol{\beta}_{jt} + \boldsymbol{\varepsilon}_{jt},\ j = 1, \cdots, p_t,\ t \in (1, \cdots, m), \tag{2.2}$$

where $\boldsymbol{\varepsilon}_{jt} = (\varepsilon_{1jt}, \varepsilon_{2jt}, \cdots, \varepsilon_{njt})^T$. To perform Bayesian analysis, we assume the following prior distributions for the model parameters

$$\boldsymbol{\beta}_{jt} \sim \text{MVN}(\boldsymbol{\beta}_{0t}, \Sigma_{0t}),\ \sigma^2_{jt} \sim \text{IG}(\nu_0/2,\ \nu_0\sigma_0^2/2),\ \gamma_{jt} \sim \text{Bernoulli}(q_t).$$

In words, we assume that the coefficient vector $\boldsymbol{\beta}_{jt}$ follows multivariate normal distribution with mean $\boldsymbol{\beta}_{0t}$ and covariance Σ_{0t}; σ^2_{jt} follows inverse-gamma distribution with shape parameter $\nu_0/2$ and scale parameter $\nu_0\sigma_0^2/2$; indicator variable γ_{jt} follows Bernoulli distribution with probability of an omics feature being selected as a driving factor for clustering is q_t. With these assumptions, we can derive the posterior distributions of the model parameters σ^2_{jt}, $\boldsymbol{\beta}_{jt}$ and γ_{jt}.

$$P(\sigma^2_{jt}|\mathbf{y}_{jt}, \mathbf{Z}, \gamma_{jt}, \boldsymbol{\beta}_{jt}) \sim \text{IG}\left(\frac{\nu_0 + n}{2}, \frac{\nu_0\sigma_0^2 + (\mathbf{y}_{jt} - \mathbf{X}\boldsymbol{\Gamma}_{jt}\boldsymbol{\beta}_{jt})^T(\mathbf{y}_{jt} - \mathbf{X}\boldsymbol{\Gamma}_{jt}\boldsymbol{\beta}_{jt})}{2}\right) \tag{2.3}$$

$$P(\boldsymbol{\beta}_{jt}|\mathbf{y}_{jt}, \mathbf{Z}, \sigma^2_{jt}, \gamma_{jt}) \sim \text{MVN}(\mathbf{m}, \mathbf{V}) \tag{2.4}$$

$$\mathbf{m} = (\boldsymbol{\Gamma}_{jt}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\Gamma}_{jt}/\sigma^2_{jt} + \Sigma_{0t}^{-1})^{-1}(\boldsymbol{\Gamma}_{jt}^T\mathbf{X}^T X_{jt}/\sigma^2_{jt} + \Sigma_{0t}^{-1}\beta_{0t})$$

$$\mathbf{V} = (\boldsymbol{\Gamma}_{jt}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\Gamma}_{jt}/\sigma^2_{jt} + \Sigma_{0t}^{-1})^{-1}.$$

$$P(\gamma_{jt}|\mathbf{y}_{jt}, \mathbf{Z}, \boldsymbol{\beta}_{jt}, \sigma^2_{jt}) \propto \exp\left(-\frac{(\mathbf{y}_{jt} - \mathbf{X}\boldsymbol{\Gamma}_{jt}\boldsymbol{\beta}_{jt})^T(\mathbf{y}_{jt} - \mathbf{X}\boldsymbol{\Gamma}_{jt}\boldsymbol{\beta}_{jt})}{2\sigma^2_{jt}}\right) P(\gamma_{jt}), \tag{2.5}$$

where $P(\gamma_{jt})$ is the prior probability of γ_{jt}. In words, the posterior distribution of σ^2_{jt} is inverse gamma distribution with scale parameter $\left(\nu_0\sigma_0^2 + (\mathbf{y}_{jt} - \mathbf{X}\boldsymbol{\Gamma}_{jt}\boldsymbol{\beta}_{jt})^T(\mathbf{y}_{jt} - \mathbf{X}\boldsymbol{\Gamma}_{jt}\boldsymbol{\beta}_{jt})\right)/2$ and shape parameter $(\nu_0 +$

$n)/2$; the posterior distribution of $\boldsymbol{\beta}_{jt}$ is multivariate normal with mean $\mathbf{m}$ and covariance $\mathbf{V}$. Since the posterior distributions of σ_{jt}^2 and $\boldsymbol{\beta}_{jt}$ are known, we can use the Gibbs sampling algorithm (Geman and Geman, 1984; Kuo and Mallick, 1998) to obtain samples from their posterior distributions. There is no closed form for parameter γ_{jt}. However, we know its distribution is proportional to the terms on the right side of $\propto$ in equation (2.5), thus we can sample from its posterior distribution using the Metropolis–Hasting algorithm (Metropolis *and others*, 1953; Hastings, 1970; Kuo and Mallick, 1998). There is no closed form for $\mathbf{z}_i$. We will show how to sample from its posterior distribution after we describe our modeling strategy for discrete data in the following sections. The latent variable z_i will be used for sample clustering, and omics features with high posterior probability of γ_{jt} being 1 will be selected as informative features.

2.3. *Statistical model for binary data*

Next generation sequencing is becoming less and less expensive, and more and more sequencing data are being generated, which are usually discrete. For example, exome sequencing data can be represented by binary values indicating mutation or no mutation for genes. RNA-seq often measures gene expression based on read count (the number of reads mapped to a given gene). Therefore, it is not optimal to use the above framework to model these discrete data. Considering the nature of these data, we propose the following model for the discrete data. If y_{ijt} is a binary variable (e.g., mutation or no-mutation), we model the data with the classical logistic regression,

$$\log\frac{P(y_{ijt}=1 \mid \mathbf{z}_i)}{1-P(y_{ijt}=1 \mid \mathbf{z}_i)} = \mathbf{x}_i\boldsymbol{\Gamma}_{jt}\boldsymbol{\beta}_{jt},\ i=1,\cdots,n,\ j=1,\cdots,p_t,\ t\in(1,\cdots,m), \tag{2.6}$$

where $P(y_{ijt}=1 \mid \mathbf{z}_i)$ is the probability of gene j being mutant in sample i given the value of the latent variable $\mathbf{z}_i$; $\mathbf{x}_i=(1,\mathbf{z}_i)$ is the same design vector as the one in model (2.1); $\boldsymbol{\Gamma}_{jt}$ and $\boldsymbol{\beta}_{jt}$ have the same components as those in model (2.1) while data type t refers to binary. We assume the prior distributions of $\boldsymbol{\beta}_{jt}$ and γ_{jt} follow MVN$(\boldsymbol{\beta}_{0t},\Sigma_{0t})$ and Bernoulli(q_t), respectively. Therefore, the posterior distributions of γ_{jt} and $\boldsymbol{\beta}_{jt}$ given the data and the other parameters in model (2.6) are

$$P(\boldsymbol{\beta}_{jt}|\mathbf{y}_{jt},\mathbf{Z},\gamma_{jt}) \propto \left(\prod_{i=1}^{n}\frac{\left(\exp(\mathbf{x}_i\boldsymbol{\Gamma}_{jt}\boldsymbol{\beta}_{jt})\right)^{y_{ijt}}}{1+\exp(\mathbf{x}_i\boldsymbol{\Gamma}_{jt}\boldsymbol{\beta}_{jt})}\right)\exp\left(-\frac{1}{2}(\boldsymbol{\beta}_{jt}-\boldsymbol{\beta}_{0t})^T\Sigma_{0t}^{-1}(\boldsymbol{\beta}_{jt}-\boldsymbol{\beta}_{0t})\right), \tag{2.7}$$

$$P(\gamma_{jt}|\mathbf{y}_{jt},\mathbf{Z},\boldsymbol{\beta}_{jt}) \propto \left(\prod_{i=1}^{n}\frac{\left(\exp(\mathbf{x}_i\boldsymbol{\Gamma}_{jt}\boldsymbol{\beta}_{jt})\right)^{y_{ijt}}}{1+\exp(\mathbf{x}_i\boldsymbol{\Gamma}_{jt}\boldsymbol{\beta}_{jt})}\right)P(\gamma_{jt}). \tag{2.8}$$

In words, the posterior distributions of γ_{jt} and $\boldsymbol{\beta}_{jt}$ are proportional to the terms on the right side of $\propto$, respectively.

2.4. *Statistical model for count data*

When y_{ijt} is a count variable (e.g., RNA-seq gene expression data), we model the data with the following Poisson regression

$$\log\left(\lambda(y_{ijt} \mid \mathbf{z}_i)\right) = \mathbf{x}_i\boldsymbol{\Gamma}_{jt}\boldsymbol{\beta}_{jt},\ i=1,\cdots,n,\ j=1,\cdots,p_t,\ t\in(1,\cdots,m), \tag{2.9}$$

where $\lambda(y_{ijt} \mid \mathbf{z}_i)$ is the predicted mean of y_{ijt} conditional on the latent variable $\mathbf{z}_i$; $\boldsymbol{\Gamma}_{jt}$ and $\boldsymbol{\beta}_{jt}$ are the corresponding indicator variable and parameter for the count data type as those for the continuous and binary cases. To perform Bayesian analysis, we use a multivariate prior MVN$(\boldsymbol{\beta}_{0t},\Sigma_{0t})$ for $\boldsymbol{\beta}_{jt}$ and Bernoulli(q_t)

for γ_{jt}. Therefore the posterior distributions of γ_{jt} and $\boldsymbol{\beta}_{jt}$ given the data and the other parameters in the Poisson regression model (2.9) are

$$P(\boldsymbol{\beta}_{jt}|\mathbf{y}_{jt},\mathbf{Z},\gamma_{jt}) \propto \left(\prod_{i=1}^{n}\left(\exp(\mathbf{x}_i\boldsymbol{\Gamma}_{jt}\boldsymbol{\beta}_{jt})\right)^{y_{ijt}}\exp\left(-\exp(\mathbf{x}_i\boldsymbol{\Gamma}_{jt}\boldsymbol{\beta}_{jt})\right)\right)\exp\left(-\frac{1}{2}(\boldsymbol{\beta}_{jt}-\boldsymbol{\beta}_{0t})^T\Sigma_{0t}^{-1}(\boldsymbol{\beta}_{jt}-\boldsymbol{\beta}_{0t})\right) \tag{2.10}$$

$$P(\gamma_{jt}|\mathbf{y}_{jt},\mathbf{Z},\boldsymbol{\beta}_{jt}) \propto \left(\prod_{i=1}^{n}\left(\exp(\mathbf{x}_i\boldsymbol{\Gamma}_{jt}\boldsymbol{\beta}_{jt})\right)^{y_{ijt}}\exp\left(-\exp(\mathbf{x}_i\boldsymbol{\Gamma}_{jt}\boldsymbol{\beta}_{jt})\right)\right)P(\gamma_{jt}). \tag{2.11}$$

There is no closed form for the posterior distributions of γ_{jt} and $\boldsymbol{\beta}_{jt}$ when the data are binary or count values, but we know they are proportional to some functions respectively. Therefore, we can use the Metropolis–Hasting algorithm to jointly sample $(\gamma_{jt},\boldsymbol{\beta}_{jt})$ from their posterior distributions for statistical inference (Savitsky *and others*, 2011).

2.5. *The joint likelihood*

As it can be seen from the above models, each of them has a common $\mathbf{x}_i = (1,\mathbf{z}_i)$, which is the key to our modeling strategy. We let sample i be associated with latent variable $\mathbf{z}_i$. Through joint modeling, $\mathbf{z}_i$ will collectively capture the major biological variations across the omics data of the cancer samples and provide a basis for predicting if an omics feature j in a given data set t is a driver for sample clustering. The joint model for the data types described above can be written as

$$P(y_{ijt},\mathbf{z}_i \mid \boldsymbol{\beta}_{jt},\gamma_{jt},i=1,\cdots,n,j=1,\cdots,p_t,t=1,\cdots,m) = \prod_{t}^{m}\prod_{i}^{n}\prod_{j}^{p_t}P\left(y_{ijt} \mid \mathbf{z}_i,\boldsymbol{\beta}_{jt},\gamma_{jt}\right)P(\mathbf{z}_i), \tag{2.12}$$

where the multiplication is due to the conditional independence assumption of y_{ijt} given $\mathbf{z}_i$; $P(\mathbf{z}_i)$ is the density function of the standard multivariate normal distribution $\mathrm{MVN}(\mathbf{0},\mathbf{I}_k)$; the conditional density function $P\left(y_{ijt} \mid \mathbf{z}_i,\boldsymbol{\beta}_{jt},\gamma_{jt}\right)$ has the form of normal, Bernoulli or Poisson depending on the type of omics data. More specifically,

$$P\left(y_{ijt} \mid \mathbf{z}_i,\boldsymbol{\beta}_{jt},\gamma_{jt}\right) \propto \begin{cases} \sigma_{jt}^{-1}\exp\left(-\left(y_{ijt}-\mathbf{x}_i\boldsymbol{\Gamma}_{jt}\boldsymbol{\beta}_{jt}\right)^2/(2\sigma_{jt}^2)\right), & \text{normal,} \\ \left(\exp(\mathbf{x}_i\boldsymbol{\Gamma}_{jt}\boldsymbol{\beta}_{jt})\right)^{y_{ijt}}\left(1+\exp(\mathbf{x}_i\boldsymbol{\Gamma}_{jt}\boldsymbol{\beta}_{jt})\right)^{-1}, & \text{binomial,} \\ \left(\exp(\mathbf{x}_i\boldsymbol{\Gamma}_{jt}\boldsymbol{\beta}_{jt})\right)^{y_{ijt}}\exp\left(-\exp(\mathbf{x}_i\boldsymbol{\Gamma}_{jt}\boldsymbol{\beta}_{jt})\right), & \text{Poisson.} \end{cases}$$

The latent variable $\mathbf{z}_i$ is not observable, thus we also need to use the Markov Chain Monte Carlo (MCMC) method to obtain samples from its posterior distribution for statistical inference. The exact posterior distribution of $\mathbf{z}_i$ is not known. However, we know it is proportional to the product of the density of $\mathbf{z}_i$ and the joint likelihood of the data as the following,

$$P\left(\mathbf{z}_i \mid \mathbf{y}_{jt},\boldsymbol{\beta}_{jt},\gamma_{jt}\right) \propto P(\mathbf{z}_i)\prod_{t}^{m}\prod_{j}^{p_t}P\left(y_{ijt} \mid \mathbf{z}_i,\boldsymbol{\beta}_{jt},\gamma_{jt}\right), \quad i=1,2,\cdots,n. \tag{2.13}$$

Therefore, the Metropolis–Hasting algorithm will be used to sample from its posterior distribution for statistical inference (Metropolis *and others*, 1953; Hastings, 1970). The mean value of the latent variables

are used for sample clustering. Specifically, following a general principle for separating g clusters among n data points, we use k-means clustering to divide the n samples into g clusters in the latent variable space, where $g = k + 1$ (Mo *and others*, 2013).

3. SIMULATION STUDIES

To demonstrate the feasibility of the integrative model, we performed simulation studies. We assumed there were 240 cancer samples analyzed by array comparative genomic hybridization (aCGH), RNA sequencing, DNA methylation assay and whole-exome sequencing (see Figure 1). These analyses generated DNA copy number, gene expression, DNA methylation and somatic mutation data, which were presented as continuous, count, continuous and binary data types, respectively. We assumed that these cancer samples belonged to four subtypes (A, B, C, and D) and each of the subtypes had 60 samples. We let each data set have 2000 genomic features and 5% of them were informative features that defined the subtypes. More specifically, we let subtype A be characterized by 50 genomic regions with amplification, which made 50 genes increase their expression. Subtype B was characterized by hypermethylation of 50 genomic loci, which decreased the expression of 50 genes. Subtype C was characterized by hypomethylation of 50 genes, which increased the expression of 50 genes. Subtype D was characterized by 50 genomic regions with deletion, which led to decreased expression of 50 genes. In addition, subtypes C and D had relatively high somatic mutation rates in 50 genes, respectively. The simulated data were generated according to these hypothesized biological events. Specifically, the DNA copy number data for genomic regions with normal copy number were randomly generated from a standard normal distribution $N(0, 1)$ and the data for genomic regions with amplification and deletion were randomly generated from $N(1, 1)$ and $N(-1, 1)$, respectively. The normal gene expression data were randomly generated from *Poisson*(2) (Poisson distribution with mean of 2), and the increased and decreased expression data were randomly generated from *Poisson*(4) and *Poisson*(1), respectively. By appropriate transformation (e.g., logit transformation of beta-values), the methylation data can be approximately normal. Thus, we used the data randomly generated from $N(0, 1)$ as the normal methylation data, and the data randomly generated from $N(1, 1)$ and $N(-1, 1)$ as the hypermethylation and hypomethylation data, respectively. For the somatic mutation data, we assumed the genes selected for analysis have a back ground mutation rate of 0.05, and the informative genes in subtypes C and D have mutation rates of 0.3 and 0.5, respectively. The mutation data were generated from Bernoulli distribution with the assumed mutation rates. The heatmaps of the four data sets are shown in Figure 1A. It can be seen that the patterns defining the subtypes are not clearly revealed on the heatmaps without clustering analysis.

In all the simulation and real data analyses reported in the paper, we set the priors for $\boldsymbol{\beta}_{jt}$ to the standard multivariate normal distribution MVN$(\mathbf{0}, \mathrm{I})$, the priors for σ_{jt}^2 to Inverse-gamma$(1, 1)$, which are uninformative priors for the model parameters. For the the indicator variable γ_{jt}, we used *Bernoulli*(0.5) as the prior for each variable. If we know the proportion of informative features, we can choose to use an informative prior, which is the case in the simulation study. In fact, we found there was very little effect on the results by choosing the prior in the range of 0.05–0.5. In our model, k is the dimension of the latent variable $\mathbf{z}_i$ ($i \in \{1, \cdots, n\}$). For a given k, the samples can be divided into $k + 1$ clusters. In real data analysis, we usually don't know the number of subtypes. Therefore, we need to try a sequence of small integers and check how well the model fits the data under different values. In this simulation study, we let $k = 1, 2, 3, 4, 5, 6$ and for each k we ran 22 000 MCMC iterations, of which the first 10 000 were discarded as burn-in.

Figure 2A and B shows the deviance ratios and Bayesian information criterion (BIC) values for the selected k. We can see the BIC value reaches the minimum and the deviance ratio reaches a plateau when k is equal to 3, which indicates the model fits the data best when the samples are divided into 4 subtypes. Therefore, BIC or deviance ratio can be used as a criterion for selecting an appropriate value for k, which

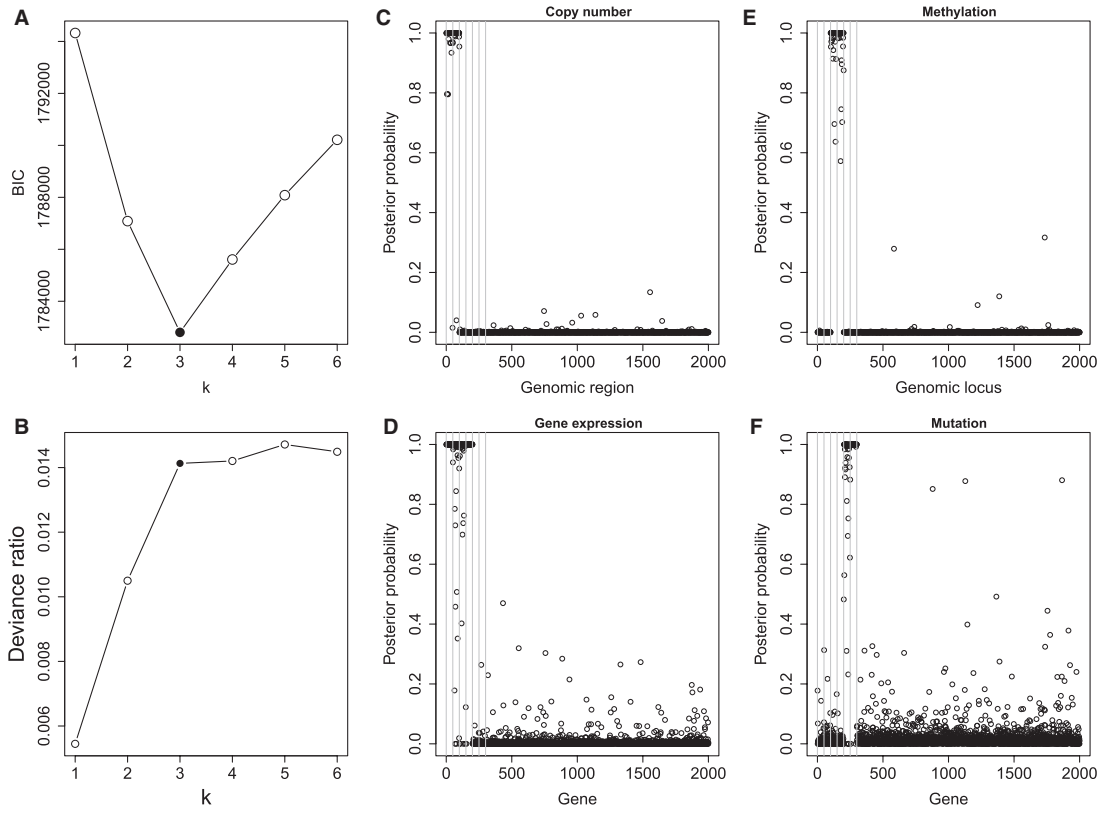

Fig. 2. Model and variable selection for the simulated data sets. (A) BIC and (B) deviance ratio at k = 1, 2, 3, 4, 5, and 6. The model fits the data best when k = 3. (C-F): Posterior probabilities of genomic features when k = 3. For easy comparison, the informative genomic features are arranged in the neighborhood of each other on the x-axes, which have high posterior probabilities of being the driver for the sample clustering. Copy number: genomic regions 1 − 50 and 51 − 100 are gain and loss regions, respectively. Methylation: genomic loci 101 − 150 and 151 − 200 are hypermethylated and hypomethylated, respectively. Gene expression: genes 1 − 200 are up- or down-regulated in response to the copy number and methylation alterations. Mutation: genes 201 − 300 are hypermutated.

defines an optimal number of clusters for the samples. Figure 2C-F shows the posterior probabilities of the genomic features. A high posterior probability suggests that the corresponding genomic feature is more likely to be a driver (informative) feature, which contributes to the integrative clustering. As it can be seen on Figure 2C-F, more than 95% of the informative features have posterior probabilities greater than 0.5 and more than 99% of the uninformative features have posterior probabilities less than 0.5. This demonstrates that the proposed method can achieve a high sensitivity and specificity in distinguishing informative features from uninformative features. In addition, all the samples cluster according to their subtypes in the three-dimensional latent variable space (Figure 1B). The genomic expression patterns of the subtypes were revealed by presenting the identified genomic features on the heat maps (Figure 1C). We also analyzed the simulated data sets using the iClusterPlus software which uses a MCMC algorithm to sample the latent variable and the L_1-norm penalized regression to induce sparsity in the model parameters. The iClusterPlus correctly clustered the samples. However, it was about 200 times slower than iClusterBayes using six cores of a Mac Pro computer for the computation (see supplementary material available at *Biostatistics* online.).

4. CASE STUDIES

4.1. *TCGA glioblastoma data application*

We first applied our method to TCGA glioblastoma (GBM) data sets. GBM was the most common brain tumor in adults and was the first cancer chosen for genomic characterization by TCGA (TCGA, 2008). The GBM data sets were downloaded from TCGA data portal (https://tcga-data.nci.nih.gov/tcga/) and the cBio Cancer Genomics Portal (http://cbioportal.org/) at the Memorial Sloan-Kettering Cancer Center. The data sets consisted of 84 samples that were measured by somatic mutation, DNA copy number and mRNA expression. The somatic mutation data were obtained by sequencing analysis of 601 selected genes in matched tumor-normal sample pairs. The data were summarized in a binary matrix (1: mutation, 0: no mutation) with the rows and columns corresponding to the samples and genes, respectively. We filtered out genes with mutation rate $\leq$ 2%. DNA copy number alterations (CNAs) measured by the Agilent 244k microarrays were used for clustering analysis. The level 3 normalized and segmented data were condensed to non-redundant regions as described by (Mo *and others*, 2013). Gene expression was measured by three microarray platforms (Affymetrix Human Exon 1.0 ST GeneChips, Affymetrix HT-HG-U133A GeneChips, and custom designed Agilent 244K array). For each gene, a unified gene expression was generated by integrating the measurements from the three platforms as described by (Verhaak *and others*, 2010). The 1740 most variable genes were used for our clustering analysis. Therefore, the GBM mutation, copy number and gene expression data were presented as binary, continuous, and continuous data types, respectively.

In this integrative clustering analysis, we used the prior $Bernoulli(0.5)$ for the indicator variable $\gamma_{jt}, j = 1, \cdots, p_t, t = 1, \cdots, m$, and tested the cluster number parameter k from 1 to 6. For each k, we ran 30 000 MCMC iterations, of which the first 18 000 were discarded as burn-in. Supplementary material available at *Biostatistics* online shows the BIC and deviance ratio at each k. We can see that the BIC as well as the deviance ratio reaches a transition point beyond which dividing samples into more clusters no longer provides significant improvement for model fitting. Therefore, we thought that $k = 3$ was an optimal parameter for the cluster number at which the samples could be divided into 4 clusters (subtypes). Figure 3B shows the 84 samples clustered in the three-dimensional latent variable space where the subtypes $1-4$ are indicated by blue, red, black and green colors respectively. Interestingly, the patients in the subtype 4 had significantly better survival than the other three subtypes (p-value = 0.019, Log-rank test; Figure 3C). In contrast, the overall survivals for the previously reported gene expression subtypes were not significantly different (p-value = 0.44, Log-rank test; Figure3D) (Verhaak *and others*, 2010).

Besides identifying the four biologically meaningful subtypes, we also identified important genes and genomic regions that contributed to the sample clustering. Figure 3A shows the genes and genomic regions with posterior probability greater than 0.5 from the three data sets. The mutated genes contributing to the subtype classification included tumor suppressor genes (NF1, TP53, MN1), genes involved in intracellular signaling cascades (NF1, TP53, MAPK9, MAPK7,PIK3R1), and genes involved in inflammatory and defense responses (A2M, ITGB2, FN1). These genes play important roles in controlling cell division and proliferation and immune response. It is well known that the accumulated mutations in these genes could lead to tumor initiation and progression. There were 161 genomic regions located on chromosomes 4, 7, 9, and 12 identified as the drivers for the subtype classification, which had distinct patterns of copy number alterations across the 4 subtypes (Figure 3A, the middle panel). These regions contain important tumor suppressor genes (e.g., CDKN2A and CDKN2B on chr9) and oncogenes that play an important role in the regulation of cell activation, division, and proliferation (e.g., PDGFRA on chr4; EGFR on chr7; TSPAN31, CDK4 and MDM2 on chr12). Deletion of tumor suppressor genes and amplification of oncogenes could lead to tumorigenesis and tumor growth. In the gene expression data set, A set of 711 genes were identified as the drivers for the subtype classification. These genes were grouped into two clusters, A and B (Figure 3, the bottom panel). Gene cluster A consisted of 204 genes whose top

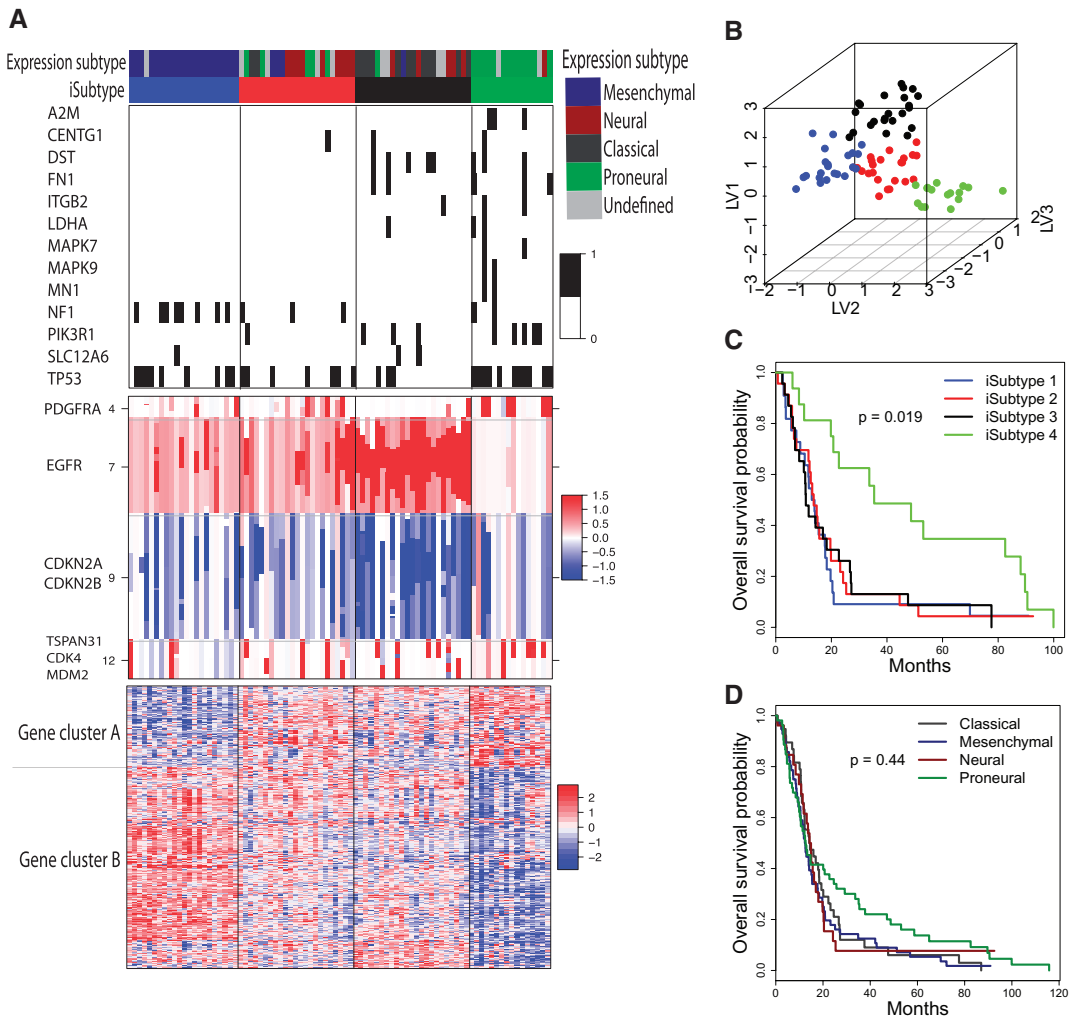

Fig. 3. GBM integrative subtypes. (A): Heatmaps of the genes and genomic regions with posterior probability > 0.5. The genomic patterns for the mutation (black, mutated; white, not mutated), copy number (red, amplification; white, normal; blue, deletion), and gene expression (red, high-level expression; blue, low-level expression) are shown on the top, middle and bottom panels, respectively. (B): GBM samples clustered in the three-dimensional latent variable space. (C): K–M survival curves for the four clusters (iSubtypes). (D): K–M survival curves for the previously reported gene expression subtypes (Verhaak *and others*, 2010). LV: latent variable.

enriched gene ontology (GO) terms included nervous system development, neurogenesis, gliogenesis, neuron differentiation, development and projection. Gene cluster B was made up of 507 genes whose top enriched GO terms included immune, wound, defense and inflammatory responses, cell activation, proliferation, migration and adhesion.

Using the unified expression data of the 1740 most variable genes, Verhaak *and others* (2010) clustered the GBM samples into 4 expression subtypes: Proneural, Neural, Classical and Mesenchymal. The integrative subtypes (iSubtypes) 1, 2, 3, and 4 were highly overlapped with the expression subtypes Mesenchymal, Neural, Classical and Proneural, respectively (See the color bars on Figure 3 A). For that reason,

the iSubtypes may be named as those used for the expression subtypes. However, the iSubtypes contained more information than the expression subtypes. Figure 3A shows the four iSubtypes with distinct genomic patterns. The iSubtype 1 (Mesenchymal) was characterized by relatively high mutation rates of NF1 and TP53, low-mutation rates of the other genes, moderate-level amplification of EGFR, moderate-level loss of CDKN2A and CDKN2B, low-level expression of gene cluster A, and high-level expression of gene cluster B. The iSubtype 2 (Neural) had a similar mutation and copy number patterns as the iSubtype 1, but it had a different expression pattern. The iSubtype 3 (Classical) was characterized by relatively low mutation rates of the genes including A2M, MAPK7, MAPK9, MN1 and NF1, high-level amplification of EGFR, high-level loss of CDKN2A and CDKN2B. The iSubtype 4 (Proneural) was a hypermutated subtype, but it had the lowest degree of copy number alteration of EGFR, CDKN2A and CDKN2B. In addition, it had high-level expression of gene cluster A and low-level expression of gene cluster B, which were negatively correlated with the expression patterns of the iSubtype 1 (Mesenchymal).

4.2. *TCGA kidney cancer data application*

Clear cell renal cell carcinoma (CCRCC) is the most common type of renal cell carcinoma (RCC), or kidney cancers. Recently, TCGA analyzed more than 400 CCRCC samples using multiple genomic platforms and reported 4 mRNA subtypes (m1–m4) and 4 miRNA subtypes (mi1-mi4) respectively. Although there were samples overlapping between mRNA and miRNA subtypes, coordinated genomic patterns across the genomic data sets were not self-evident. We hypothesized that an integrative clustering analysis would be more powerful in revealing coordinated genomic, epigenomic, and transcriptomic patterns in CCRCC subtypes. To this end, we performed an integrative clustering analysis of 241 CCRCC samples that had somatic mutation, copy number, mRNA expression, methylation, and miRNA expression data using our method. All the data used for the analysis were the Level 3 data, which were downloaded from http://firebrowse.org. Specifically, the somatic mutation data were generated by DNA sequencing and the mutation calls including deletion, insertion, missense, nonsense, RNA, splice site and translation start site mutations were summarized in the mutation annotation format (MAF). A gene by sample matrix of binary values (1, mutation; 0, no mutation) was generated from the mutation data. Genes with mutation rate $> 2\%$ were used for clustering. For the copy number data, the normalized and segmented data based on Affymetrix SNP Array 6.0 were used. The segmented data were further condensed to 4470 non-redundant copy number regions using the method described in Mo *and others* (2013). The gene expression data were generated by RNA sequencing and the level 3 expression data were the log2-transformed normalized count values. We used the top 20% (4106) most-variable genes for the clustering analysis. The DNA methylation level was measured using Illumina DNA methylation arrays and the methylation level at each CpG locus was summarized as a $\beta-$value ranging from 0 to 1, which can be interpreted as the percentage of methylation. We performed logit transformation for the $\beta-$values and then used the top 20% (3955) most-variable CpG sites that had the minimum correlation with the mRNA-seq data for the clustering analysis. The miRNA expression data were generated by miRNA sequencing and the level 3 data were the normalized count values represented as reads per million miRNA precursor reads (RPM). The data were log2 transformed and the bottom 20% least-variable were removed from the clustering analysis. In summary, the CCRCC mutation data were modeled as binary data types, and the other data were modeled as continuous data types.

As we analyzed the GBM data sets, we used the prior $Bernoulli(0.5)$ for the indicator variable γ_{jt} $(j = 1, \cdots, p_t, t = 1, \cdots, m)$ and tested the parameter k from 1 to 6. For each k, we ran 30 000 MCMC iterations including the initial 18 000 burn-in iterations. In our initial analysis, we included the DNA copy number data. However, we found that the copy number data contributed little to the integrative clustering because all the regions had low posterior probability (< 0.5). Therefore, we excluded the copy number data and only used the remaining four data sets for the final integrative analysis. Supplementary material

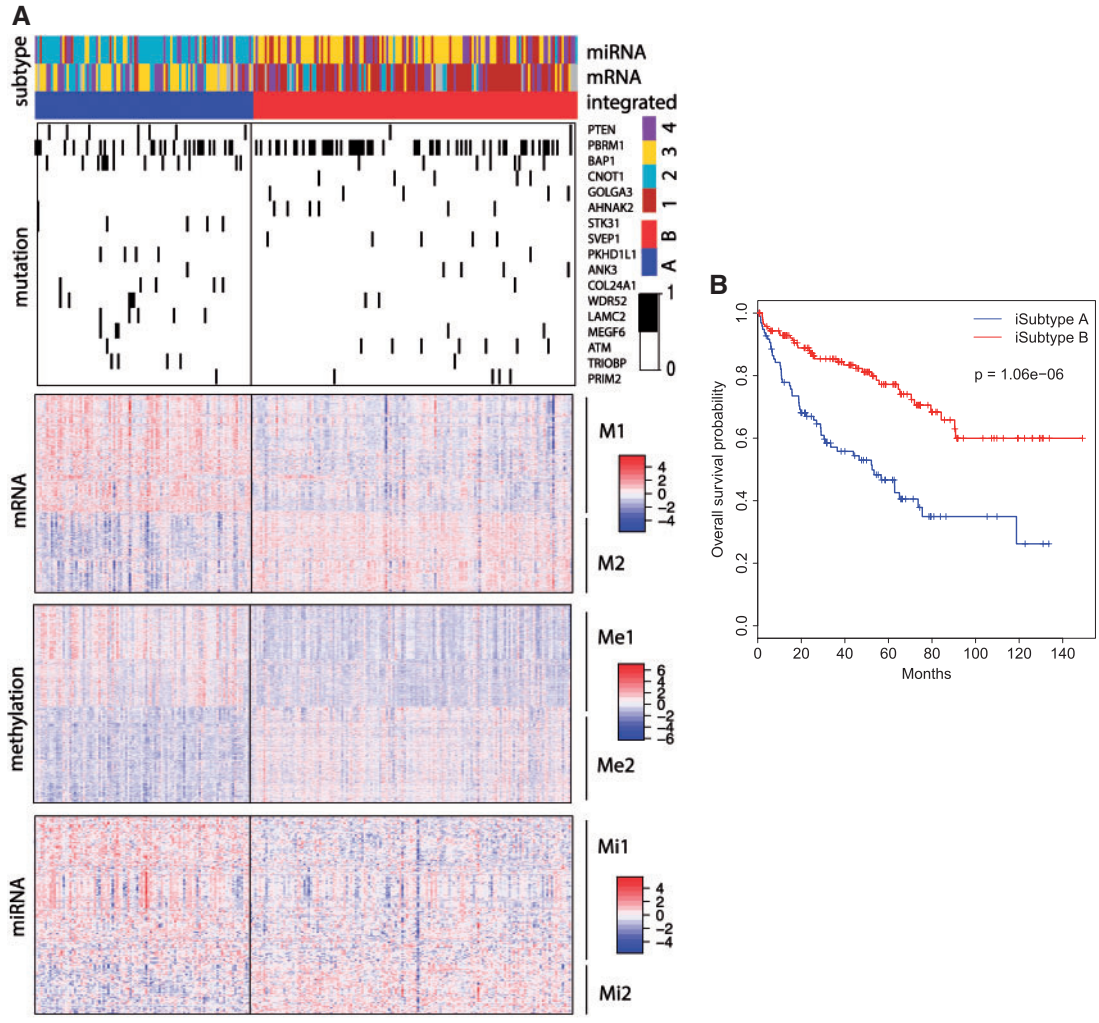

Fig. 4. CCRCC integrative subtypes. (A): Heatmaps of the genomic features with posterior probability > 0.5. The TCGA miRNA and mRNA subtypes are shown on the top of the heatmaps. For the mutation heatmap, black and white represent mutation and no mutation, respectively. For the mRNA, miRNA and methylation heatmaps, blue, white, and red represent low, middle, and high level of expression or methylation, respectively. The feature clusters are labeled as M1 and M2 for mRNA, Me1, and Me2 for methylation, and Mi1 and Mi2 for miRNA, respectively. (B): K–M survival curves for iSubtypes A and B.

available at *Biostatistics* online shows the BIC and deviance ratio for each k, respectively. We can see that the BIC is the minimum and the deviance ratio is the maximum when $k = 1$. Therefore, we thought that $k = 1$ was an optimal parameter for the integrative clustering, which implied that it was optimal to divide the samples into 2 iSubtypes. The iSubtype A was mainly overlapped with TCGA miRNA subtype 2 and mRNA subtype 3, and the iSubtype B was mainly overlapped with TCGA miRNA subtype 3 and mRNA subtype 1. The two iSubtypes showed different survival functions: the patients in the iSubtype B had significantly better survival than the patients in the iSubtype A (p-value = 1.06e–06, Log-rank test; Figure 4B).

Figure 4A shows distinct patterns of alterations in the two iSubtypes. There were 17 mutated genes identified as the drivers for the iSubtype classification, which showed different mutation rates in the two iSubtypes. The mutated genes include tumor suppressor PTEN, chromatin remodeler PBRM1 and BAP1 that can act as tumor suppressors by regulating gene expression through chromatin remodeling, transcription regulator CNOT1, Serine/Threonine kinases STK31 and ATM. Notably, ATM is an important cell cycle checkpoint kinase that regulates a variety of downstream targets including tumor suppressors BRCA1 and p53, cell cycle checkpoint proteins RAD17 and RAD9, checkpoint kinase CHK2 and DNA repair protein NBS1 (http://www.genecards.org). There were 2064 mRNAs, 2077 CpG sites and 229 miRNAs identified as the major contributors for the iSubtype classification. The genomic features in each of the three data sets appeared to form two clusters whose expression or methylation levels were relatively higher in one iSubtype and lower in the other iSubtype. These feature clusters were labeled as mRNA clusters M1 and M2, methylation clusters Me1 and Me2, and miRNA clusters Mi1 and Mi2, respectively (Figure 4A). Gene functional enrichment analysis using DAVID bioinformatics tools showed that gene ontology (GO) terms related to cell adhesion, immune, defense, inflammatory and wound responses were among the top enriched GO terms in the negatively correlated mRNA cluster M1 and methylation cluster Me2. GO terms related to cell motion and morphogenesis, regulation of cell migration and motion, positive regulation of DNA-dependent transcription and RNA metabolic process, and response to hormone, endogenous and extracellular stimuli were among the top enriched GO terms in the negatively correlated feature clusters M2 and Me1. The miRNAs contributing to the iSubtype classification also had important functions related to immune response, inflammation, apoptosis, cell proliferation, differentiation, death, and motility.

5. DISCUSSION

In the past years, an unprecedented amount of omics data have been generated by national and international cancer genome consortia, which have provided great resources for researchers to study cancer biology and to identify clinically relevant subtypes and biomarkers for precision medicine. These omics data are usually presented in different data types that pose a challenge for integrative analysis. For example, mutation data can be summarized in binary format with 1 and 0 indicating mutation and normal, respectively; microarray data are usually continuous; RNA-seq gene expression data are usually presented as count format. Different data types have different sources of variation, and need to be modeled differently. However, there is still a lack of data type flexible and computationally efficient methods for integrative analysis of multi-type omics data. In this article, we present a fully Bayesian latent variable model that can model any combination of major omics data including continuous, binary and count data for integrative clustering analysis. In our method, the latent variable designed to capture the inherent structure of multi-type omics data is used as the hub for statistical modeling. Conditioning on the latent variable, different types of data are assumed to be independent. Therefore, they can be modeled separately and then are integrated through the latent variable. In order to identify the features that drive the integrative clustering, Bayesian variable selection methods are naturally incorporated in the model. This new modeling approach has two advantages over the iClusterPlus method. First, it provides posterior probability estimation for each omics feature, which can be used as a criterion for feature selection. In contrast, the iClusterPlus method doesn't provide statistical inference (e.g., p-value or confidence interval) for variable selection due to the limitation of lasso-type penalized regression. Second, it significantly reduce the computational time, which make it possible to perform integrative clustering analysis on a single computer workstation. For example, it took iClusterBayes about 2 h to analyze the simulated data sets using 6 cores of a 2.62 GHz 12-core Mac Pro computer. However, it took the iClusterPlus about 391 h (see supplementary material available at *Biostatistics* online).

Through simulation studies, we demonstrated that our new method precisely identified the samples' subtypes and subtype-specific driver features. Using the TCGA GBM, and CCRCC data sets, we demonstrated that our new method was able to identify clinically relevant cancer subtypes and driver genes. For example, integrative clustering of the TCGA GBM data revealed four iSubtypes with distinct genomic patterns (Figure 3A). Interestingly, the GBM patients in the iSubtype 4 tended to have a better survival than the patients in the other three iSubtypes. Known disease-associated genes including TP53, PIK3R1, NF1, MN1, MAPK9, MAPK7, FN1, DST, and A2M were found to be the driver genes, which shown different mutation frequencies among the four iSubtypes. Tumor suppressor genes CDKN2A and CDKN2B, and oncogenes PDGFRA and EGFR were identified as driver genes that showed different copy number alteration patterns among the four iSubtypes. These genes are potential therapeutic targets for the subgroup of patients with abnormal expression or mutation. For the TCGA CCRCC data sets, we discovered two CCRCC iSubtypes that showed distinct mutation, methylation, mRNA and miRNA expression patterns (Figure 4A). These two iSubtypes were also clinically meaningful because their overall survivals were significantly different (Figure 4B). Brannon *and others* (2010) reported two gene expression subtypes named ccA and ccB. In terms of survival, the ccA and ccB are similar to the iSubtype A and B, respectively. Tumor suppressor genes including PTEN, PBRM1 and BAP1, and cell cycle checkpoint kinase ATM showed different mutation frequencies between the two iSubtypes, which could be potential therapeutic targets for subtype-specific therapy. In addition, a set of genes involved in immune response, transcription regulation, and cell adhesion and motion had different expression patterns between the two iSubtypes, which can also be potential therapeutic targets.

In summary, we have developed a fully Bayesian method for integrative clustering analysis of multi-type omics data. This new method significantly improved on the iClusterPlus method in terms of statistical computation. In addition, it provides a posterior probability estimation for each omics feature, which is a great advantage over the iClusterPlus method. It will provide researchers a powerful tool for deconvolution of cancer omics data and identification of clinically meaningful cancer subtypes and potential therapeutic targets.

6. Software

The iClusterBayes method was implemented in C language and wrapped in R code. User-friendly functions will be included in the iClusterPlus package https://www.bioconductor.org/packages/devel/bioc/html/iClusterPlus.html.

Supplementary material

Supplementary material is available online at http://biostatistics.oxfordjournals.org.

Acknowledgments

The authors thank the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing HPC resources. *Conflict of Interest*: None declared.

Funding

This work is supported by a seed grant to Q.M. from the Computational and Integrative Biomedical Research (CIBR) Center at Baylor College of Medicine, Houston, TX, USA. Q.M. and S.G.H. are supported in part by the cancer center support grant CA125123. Q.M. and K.C. are supported in part by R01 CA175397. R.S. is supported in part by P30-CA008748, P01-CA129243, CA195365, Movember Foundation-PCF Challenge Award GC226463. M.V. is supported in part by T32 CA096520.

REFERENCES

BRANNON, A. R., REDDY, A., SEILER, M., ARREOLA, A., MOORE, D. T., PRUTHI, R. S., WALLEN, E. M., NIELSEN, M. E., LIU, H., NATHANSON, K. L., LJUNGBERG, B., ZHAO, H. *and others*. (2010). Molecular stratification of clear cell renal cell carcinoma by consensus clustering reveals distinct subtypes and survival patterns. *Genes Cancer* **1**, 152–163.

GEMAN, S. AND GEMAN, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.

GEORGE, E. I. AND MCCULLOCH, R. E. (1997). Approaches of bayesian variable selection. *Statistica Sinica* **7**, 339–373.

HASTINGS, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika* **57**, 97–109.

KIRK, P., GRIFFIN, J. E., SAVAGE, R. S., GHAHRAMANI, Z. and WILD, D. L. (2012). Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics* **28**, 3290–3297.

KORMAKSSON, M., BOOTH, J. G., FIGUEROA, M. E. AND MELNICK, A. (2012). Integrative model-based clustering of microarray methylation and expression data. *Annals of Applied Statistics* **6**, 1327–1347.

KUO, L. AND MALLICK, B. K. (1998). Variable selection for regression models. *Sankhya: The Indian Journal of Statistics. Series B* **60**, 65–81.

LOCK, E. F. AND DUNSON, D. B. (2013). Bayesian consensus clustering. *Bioinformatics* **29**, 2610–2616.

METROPOLIS, N., ROSENBLUTH, A.W., ROSENBLUTH, M.N., TELLER, A.H. AND TELLER, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics* **21**, 1087–1092.

MO, Q., WANG, S., SESHAN, V. E., OLSHEN, A. B., SCHULTZ, N., SANDER, C., POWERS, R. S., LADANYI, M. and SHEN, R. (2013). Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proceedings of the National Academy of Sciences* **110**, 4245–4250.

SAVITSKY, T., VANNUCCI, M. AND SHA, N. (2011). Variable selection for nonparametric gaussian process priors: Models and computational strategies. *Statistical Science* **26**, 130–149.

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.

SHEN, R., WANG, S. AND MO, Q. (2013). Sparse integrative clustering of multiple omics data sets. *Annals of Applied Statistics* **7**, 269–294.

TCGA. (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068.

TCGA. (2012). Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519–525.

TCGA. (2013). Integrated genomic characterization of endometrial carcinoma. *Nature* **497**, 67–73.

TCGA. (2014*a*). Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* **513**, 202–209.

TCGA. (2014*b*). Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543–550.

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

VERHAAK, R. G. W., HOADLEY, K. A., PURDOM, E., WANG, V., QI, Y., WILKERSON, M. D., MILLER, C. R., DING, L., GOLUB, T., MESIROV, J. P. *and others*. (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in pdgfra, idh1, egfr, and nf1. *Cancer Cell* **17**, 98–110.

[*Received August 31, 2016; revised Februay 20, 2017; accepted for publication March 14, 2017*]