A Bayesian framework for pathway-guided identification of cancer subgroups by integrating multiple types of genomic data

Sun, Z., Chung, D., Neelon, B., Millar-Wilson, A., Ethier, S. P., Xiao, F., Zheng, Y., Wallace, K., & Hardiman, G. (2023). A Bayesian framework for pathway-guided identification of cancer subgroups by integrating multiple types of genomic data. *Statistics in Medicine*, *42*(28), 5266-5284. https://doi.org/10.1002/sim.9911

Published in:
Statistics in Medicine

Document Version:
Publisher's PDF, also known as Version of record

Queen's University Belfast - Research Portal:
Link to publication record in Queen's University Belfast Research Portal

Received: 19 March 2021 | Revised: 15 July 2023 | Accepted: 5 September 2023

DOI: 10.1002/sim.9911

RESEARCH ARTICLE

Statistics in Medicine WILEY

A Bayesian framework for pathway-guided identification of cancer subgroups by integrating multiple types of genomic data

Zequn Sun[1] | Dongjun Chung[2,3] | Brian Neelon[4] | Andrew Millar-Wilson[5] | Stephen P. Ethier[6] | Feifei Xiao[7] | Yinan Zheng[1] | Kristin Wallace[4] | Gary Hardiman[4,8]

[1]Department of Preventive Medicine, Northwestern University, Chicago, Illinois,

[2]Department of Biomedical Informatics, The Ohio State University, Columbus, Ohio,

[3]The Pelotonia Institute for Immuno-Oncology, The Ohio State University Comprehensive Cancer Center, Columbus, Ohio,

[4]Department of Public Health Sciences, Medical University of South Carolina, Charleston, South Carolina,

[5]School of Biological Sciences, Queen's University Belfast, Belfast, UK

[6]Department of Pathology and Laboratory Medicine, Medical University of South Carolina, Charleston, South Carolina,

[7]Department of Biostatistics, University of Florida, Gainesville, Florida,

[8]Faculty of Medicine, Health and Life Sciences, School of Biological Sciences and Institute for Global Food Security, Queen's University Belfast, Belfast, UK

Correspondence
Zequn Sun, Department of Preventive Medicine, Northwestern University, Chicago, IL, 60611, USA.
Email: zequn.sun@northwestern.edu

Funding information
National Cancer Institute, Grant/Award Number: R21-CA209848; National Institute of General Medical Sciences, Grant/Award Number: R01-GM122078; National Institute on Drug Abuse, Grant/Award Number: U01-DA045300

In recent years, comprehensive cancer genomics platforms, such as The Cancer Genome Atlas (TCGA), provide access to an enormous amount of high throughput genomic datasets for each patient, including gene expression, DNA copy number alterations, DNA methylation, and somatic mutation. While the integration of these multi-omics datasets has the potential to provide novel insights that can lead to personalized medicine, most existing approaches only focus on gene-level analysis and lack the ability to facilitate biological findings at the pathway-level. In this article, we propose Bayes-InGRiD (Bayesian Integrative Genomics Robust iDentification of cancer subgroups), a novel pathway-guided Bayesian sparse latent factor model for the simultaneous identification of cancer patient subgroups (clustering) and key molecular features (variable selection) within a unified framework, based on the joint analysis of continuous, binary, and count data. By utilizing pathway (gene set) information, Bayes-InGRiD does not only enhance the accuracy and robustness of cancer patient subgroup and key molecular feature identification, but also promotes biological understanding and interpretation. Finally, to facilitate an efficient posterior sampling, an alternative Gibbs sampler for logistic and negative binomial models is proposed using Pólya-Gamma mixtures of normal to represent latent variables for binary and count data, which yields a conditionally Gaussian representation of the posterior. The R package "INGRID" implementing the proposed approach is currently available in our research group GitHub webpage (https://dongjunchung.github.io/INGRID/).

KEYWORDS
Bayesian model, biological pathway, clustering, integrative analysis, variable selection

1 | INTRODUCTION

In cancer genomics, it is of critical interest to identify cancer patient subgroups as it can facilitate the development of personalized medicine. The identification of novel molecular features associated with these patient subgroups can potentially lead to a novel biomarker for prognosis, diagnosis, and novel therapeutic targets. The emergence of comprehensive cancer genomics platforms, such as The Cancer Genome Atlas (TCGA),[1] opened unprecedented opportunities for such investigation by providing researchers an enormous amount of high throughput genomic datasets for each patient, including gene expression, DNA copy number alterations, DNA methylation, somatic mutation, miRNA, and proteomics.[1] At the same time, the availability of large-scale high-throughput multi-omics data sets requires the development of novel data integration methods that can effectively detect interactions and shared information among multiple data sets. Moreover, these datasets consist of both continuous and discrete forms, and hence, there is a need for a statistical approach that is also capable of handling various types of data.

Traditionally, principal component analysis (PCA)[2] has been used to decipher a single data set in a continuous form. As PCA achieves dimension reduction, its latent components can be used to identify patient subgroups. However, approaches such as PCA are no longer adequate for the integrative analysis of multiple data sets, since the latent components induced from PCA will be distinct between data types. To integrate multiple continuous data, joint latent factor models have been proposed to study both common and unique variations across different data sets, for example, iCluster,[3] iNMF[4] and JIVE.[5] Specifically, iCluster features a joint latent variable model and cancer subgroups can be identified by applying a clustering algorithm on the shared latent factors,[3] while key genes can be identified from multiple genomic platforms through regularization on the factor loading. JIVE and iNMF further extended iCluster by introducing a data-specific term, which affects the estimation of shared structures.[4,5] Although integrative approaches like JIVE and iNMF promoted the understanding of individual data structures, they still lack guidance on a meaningful patient subgroup clustering.

iCluster+ overcame the limitation of integrating only continuous data and implements the joint analysis of continuous, binary, counts, and categorical data using a latent factor model.[6,7] Recently, Mo and others improved iCluster+ and developed a Bayesian sparse latent factor model to integrate multiple types of omics data, called iClusterBayes.[8] The advantages of this Bayesian framework in data integration are three-fold: (i) it has flexibility in the specification of distributional assumptions on multiple types of data sets, as well as on the correlations among data sets; and (ii) it allows them to avoid complicated parameter tuning required when a penalization algorithm is used; and (iii) one could incorporate prior biological expert knowledge. iClusterBayes enables a posterior probability estimation for gene selection and improves the iCluster+ method regarding computational speed significantly.[8]

While integration approaches such as iCluster+ and iClusterBayes help us capture molecular interactions among different omics datasets, most existing methods only focus on gene-level analysis and lack the ability to facilitate biological findings at the pathway-level. The pathway-level analysis provides information about natural grouping structure and key insights to guide factor definition.[9] iFad[10] and PacFad[11] enable the incorporation of prior knowledge and represent biological pathways as latent factors in the Bayesian sparse factor analysis models. However, iFad and PacFad are only applicable to continuous data and the problem of an excessive number of latent factors in the model still remains a challenge that leads to a higher computational burden. InGRiD[12] is another approach that examines the genetic features at the pathway-level. To promote a robust interpretation of the pathway-level analysis results, Wei and colleagues built pathway-level latent components using sparse partial least squares Cox regression model,[13] an approach that allows simultaneous identification of key genes and pathways without a need for separate downstream gene set enrichment analysis. However, this approach can only be applied to single continuous data. To fully understand tumorigenesis at the system level, it is necessary to integrate the changes found in multiple types of omics data (ie, continuous, discrete) at the pathway level.

Another limitation of the iClusterBayes approach is the need for Metropolis-Hastings sampling[14] for the Bayesian inference. There are no close forms for the posterior distributions of multiple parameters, especially in the models derived for binary and count data. Hence, an alternative posterior sampling strategy without the need of using Metropolis-Hasting sampling can be of great interest. Recently, Polson and colleagues proposed an alternative Gibbs sampler that introduces a vector of latent variables that are scale mixtures of normals with independent Pólya-Gamma precision terms.[15] Pillow and Scott further extended the model to handle negative binomial (NB) case[16] for the count data. The application of Pólya-Gamma mixtures of normals leads to simple, effective methods for posterior inference and boosts fully automatic Gibbs sampler to avoid parameter tuning.

To overcome these limitations, here we propose a novel pathway-guided Bayesian sparse latent factor method, named Bayes-InGRiD (Bayesian Integrative Genomics Robust iDentification of cancer subgroups). Bayes-InGRiD can jointly model continuous, binary, and count omics data within a unified framework and can simultaneously identify patient subgroups and key molecular features. In addition, Bayes-InGRiD employs Pólya-Gamma mixtures of normal for binary and count data to promote an exact and fully automatic posterior sampling. Finally, pathway information is used to guide latent factor construction, provides information about natural grouping structure, and facilitates biological understanding and interpretation.

2 | METHOD

Here our main goals include (i) construction of a natural and unified framework for integrative analysis; (ii) incorporating prior biological knowledge; (iii) implementing efficient posterior sampling; and (iv) simultaneously identification of patient subgroups and key molecular features/pathways. We developed our model based on a Bayesian sparse latent factor model equipped with the Pólya-Gamma approach and the prior-guided latent factor structure to achieve the goals. Suppose we have m types of genomic data for n patients. We define $\mathbf{y}_{it} = (y_{i1t}, \dots, y_{ip_t t})^T$ to be the data vector, where y_{ijt} denotes genomic measurement for the jth molecular feature ($j = 1, \dots, p_t$) of the ith sample ($i = 1, \dots, n$) in the tth data type ($t = 1, \dots, m$). Modeling the high-dimensional space $\{\mathbf{Y}_t\}_{t=1}^m$, as a sparse linear combination of latent factors induces dimensionality reduction to a low-dimensional subspace $\mathbf{Z} = (\mathbf{z}_1, \dots, \mathbf{z}_n)^T$. We define $\mathbf{z}_i = (z_{i1}, \dots, z_{ik})$, $i = 1, \dots, n$, where $\mathbf{z}_i$ is a continuous latent variable from a standard multivariate normal distribution $MVN(0, I_k)$ and k is the number of independent latent components. The latent factor space $\mathbf{Z}$ captures the hidden structure shared among different data types in integrated data analysis that can be used for patient subgroup clustering.

2.1 | Bayesian latent factor model

In this section, we will first introduce the Bayesian latent factor model framework, which is motivated by the iCluster-Bayes[8] approach. If y_{ijt} is a continuous variable, we assume the following model,

$$y_{ijt} = \mathbf{z}_i \mathbf{\Gamma}_{jt} \boldsymbol{\beta}_{jt} + \epsilon_{ijt}, i = 1, \dots, n, j = 1, \dots, p_t, t = 1, \dots, m, \tag{1}$$

where $\mathbf{z}_i = (1, z_{i1}, \dots, z_{ik})$ is a latent factor vector of the ith sample; $\boldsymbol{\beta}_{jt} = (\beta_{0jt}, \beta_{1jt}, \dots, \beta_{kjt})^T$ denote the coefficient vector of the jth feature in the tth data set. We assume $\epsilon_{ijt} \sim N(0, \sigma_{jt}^2)$. $\mathbf{\Gamma}_{jt} = diag(1, \gamma_{jt}, \dots, \gamma_{jt})$ is a diagonal matrix serving as an indicator variable, where γ_{jt} takes values of either 0 or 1 for variable selection;[17] γ_{jt} indicates whether the corresponding feature can be decomposed as a linear combination of the latent structure or not. This, in turn, implies that the feature is key/not key for subgroup identification since the latent structure guides clustering. The model is designed so that $y_{ijt} = \beta_{0jt} + \epsilon_{ijt}$ if $\gamma_{jt} = 0$, which means that the corresponding feature is not selected as a key molecular feature for the patient subgroup identification. If $\gamma_{jt} = 1$, then $y_{ijt} = \beta_{0jt} + \beta_{1jt} z_{i1} + \cdots + \beta_{kjt} z_{ik} + \epsilon_{ijt}$, which means the corresponding $\boldsymbol{\beta}_{jt}$ is sufficiently away from zero and thus the corresponding feature contributes to the patient subgrouping. Next, if y_{ijt} is a binary variable, we assume the following logistic regression model.

$$log\left(\frac{P(y_{ijt} = 1|\mathbf{z}_i)}{1 - P(y_{ijt} = 1|\mathbf{z}_i)}\right) = \mathbf{z}_i \mathbf{\Gamma}_{jt} \boldsymbol{\beta}_{jt}, i = 1, \dots, n, j = 1, \dots, p_t, t = 1, \dots, m. \tag{2}$$

Moreover, if y_{ijt} is a count variable, we assume the following NB regression model.

$$\begin{aligned} P(y_{ijt}|r_{jt}, \mathbf{z}_i) &= \frac{\Gamma(y_{ijt} + r_{jt})}{\Gamma(r_{jt}) y_{ijt}!}(1 - \psi_{ijt})^{r_{jt}} \psi_{ijt}^{y_{ijt}}, r_{jt} > 0, j = 1, \dots, p_t, t = 1, \dots, m \\ \text{logit}(\psi_{ijt}) &= \mathbf{z}_i \mathbf{\Gamma}_{jt} \boldsymbol{\beta}_{jt}, \end{aligned} \tag{3}$$

where r_{jt} is the dispersion parameter. Finally, the joint model for the latent factor model for data integration is as follows

$$P(y_{ijt}, \mathbf{z}_i|\boldsymbol{\beta}_{jt}, \boldsymbol{\Gamma}_{jt}) \propto \prod_{t=1}^{m}\prod_{i=1}^{n}\prod_{j=1}^{p_t} P(y_{ijt}|\mathbf{z}_i, \boldsymbol{\beta}_{jt}, \boldsymbol{\Gamma}_{jt})P(\mathbf{z}_i), \tag{4}$$

where $\mathbf{z}_i$ follows a standard multivariate normal distribution $MVN(\mathbf{0}, \mathcal{I}_k)$; $P(y_{ijt}|\mathbf{z}_i, \boldsymbol{\beta}_{jt}, \boldsymbol{\Gamma}_{jt})$ is the conditional density function, where the form of distribution of $P(y_{ijt}|\mathbf{z}_i, \boldsymbol{\beta}_{jt}, \boldsymbol{\Gamma}_{jt})$ can be Gaussian, Bernoulli or NB depending on the data type; and the conditional independence of y_{ijt} is assumed given $\mathbf{z}_i$.

To achieve computationally efficient posterior sampling for the Bayesian inference, we will modify the above models by introducing scale mixtures of normals using Pólya-Gamma in the following two subsections.

2.2 | Bayesian latent factor model for binary data

To devise an alternative Gibbs sampler for logistic models, we apply scale mixtures of normals with independent Pólya-Gamma precision terms proposed by Polson and colleagues.[16] Assuming a random variable ω has a Pólya-Gamma distribution, an important property of the $PG(b, 0)$ density—is that for $a \in \mathfrak{R}$, $b > 0$ and $\eta \in \mathfrak{R}$,

$$\frac{(e^{\eta})^a}{(1+e^{\eta})^b} = 2^{-b}e^{\kappa\eta}\int_0^{\infty} e^{-\omega\eta^2/2}p(\omega|b, 0)d\omega, \tag{5}$$

where $\kappa = a - b/2$ and $p(\omega|b, 0)$ denotes a $PG(b, 0)$ density. Specifically, under the logistic model, the conditional likelihood for the binary response variable y_{ijt} is

$$P(y_{ijt}|\mathbf{z_i}) = \frac{(e^{\mathbf{z}_i\boldsymbol{\Gamma}_{jt}\boldsymbol{\beta}_{jt}})^{y_{ijt}}}{1+e^{\mathbf{z}_i\boldsymbol{\Gamma}_{jt}\boldsymbol{\beta}_{jt}}} = \frac{(e^{\eta_{ijt}})^{y_{ijt}}}{1+e^{\eta_{ijt}}}, i = 1, \ldots, n, j = 1, \ldots, p_t, t = 1, \ldots, m, \tag{6}$$

where $\eta_{ijt} = \mathbf{z}_i\boldsymbol{\Gamma}_{jt}\boldsymbol{\beta}_{jt}$. With $a = Y_{ijt}$ and $b = 1$, we can re-write the Bernoulli likelihood in terms of the Pólya-Gamma random variables $\boldsymbol{\Omega}_{jt} = \text{diag}(\omega_{1jt}, \ldots, \omega_{njt})$ as

$$P(y_{ijt}|r_{jt}, \boldsymbol{\beta}) = e^{\kappa_{ijt}\eta_{ijt}}\int_0^{\infty} e^{-\omega_{ijt}\eta_{ijt}^2/2}p(\omega_{ijt}|1, 0)d\omega_{ijt}, \tag{7}$$

where $\kappa_{ijt} = y_{ijt} - 1/2$ and the ω_{ijt} 's are independently distributed according to $PG(1, \eta_{ijt})$. By using the properties (5) to (7) of Polson et al[16] related to the Pólya-Gamma distribution, we can show the conditional distribution of $\boldsymbol{\beta}$ for binary response under the logistic model is

$$P(\boldsymbol{\beta}_{jt}|\mathbf{Z}, \mathbf{y}_{jt}, \boldsymbol{\Omega}_{jt}, \boldsymbol{\Gamma}_{jt}) \propto P(\boldsymbol{\beta}_{jt})\exp\left\{-\frac{1}{2}(\mathbf{u}_{jt} - \mathbf{Z}\boldsymbol{\Gamma}_{jt}\boldsymbol{\beta}_{jt})^T\boldsymbol{\Omega}_{jt}(\mathbf{u}_{jt} - \mathbf{Z}\boldsymbol{\Gamma}_{jt}\boldsymbol{\beta}_{jt})\right\}, \tag{8}$$

where $\mathbf{u}_{jt} = (u_{1jt}, \ldots, u_{njt})$ is a length n vector and its ith element $u_{ijt} = (y_{ijt} - 1/2)/(\omega_{ijt})$. This property leads to an augmentation step in the MCMC, allowing for a Gibbs instead of MH step on some of the parameters of the model although extra steps are needed to sample $\boldsymbol{\Omega}_{\mathbf{jt}}$.

2.3 | Bayesian latent factor model for count data

By parameterizing the NB probability parameter ψ_{ijt} with the *expit* function, where $expit(x) = 1/(1 + exp(-x))$, we can apply the same properties of the Pólya-Gamma density as in the logistic case.[15] Exploiting the earlier property of the Pólya-Gamma distribution with Equation (5), it follows that $\kappa_{ijt} = (y_{ijt} + r_{jt})/2$ and the ω_i 's are independently distributed according to $PG(y_{ijt} + r_{jt}, \eta_{ijt})$.[15] The parameter r_{jt} is used to capture the over-dispersion in count data. In particular, the

counts become increasingly dispersed relative to the Poisson distribution when $r_{jt} \to 0$. The full conditional for $\boldsymbol{\beta}_{jt}$ is

$$P(\boldsymbol{\beta}_{jt}|\mathbf{Z}, \mathbf{y}_{jt}, \boldsymbol{\Omega}_{jt}, \boldsymbol{\Gamma}_{jt}) \propto P(\boldsymbol{\beta}_{jt}) \exp\left\{-\frac{1}{2}(\mathbf{u}_{jt} - Z\boldsymbol{\Gamma}_{jt}\boldsymbol{\beta}_{jt})^T \boldsymbol{\Omega}_{jt}(\mathbf{u}_{jt} - Z\boldsymbol{\Gamma}_{jt}\boldsymbol{\beta}_{jt})\right\}, \tag{9}$$

where $\mathbf{u}_{jt} = (u_{1jt}, \ldots, u_{njt})$ is a length n vector and its ith element $u_{ijt} = (y_{ijt} - r_{jt})/(2\omega_{ijt})$.

To promote conjugate Gibbs update for dispersion parameter r_{jt} in the NB process, we use Chinese restaurant table (CRT) distribution for the sampling of r_{jt}.[18,19] The approach introduces a sample of latent counts, l_{ijt}, underlying each observed count y_{ijt}. Regarding sampling of the over-dispersion parameter, conditional on y_{ijt} and r_{jt}, l_{ijt} has a distribution defined by a CRT distribution as shown in Zhou and Carin:[18]

$$\begin{aligned} l_{ijt} &= \sum_{d=1}^{y_{ijt}} \mu_d \\ \mu_d &\sim Bern\left(\frac{r_{jt}}{r_{jt} + d - 1}\right), \end{aligned} \tag{10}$$

where $\mu_d = 1$ if a new customer sits in an unoccupied table in a Chinese restaurant (according to a so-called "Chinese restaurant process"), and l_{ijt} is the total number of occupied tables in the restaurant after y_{ijt} customers. $l_{ijt} \le y_{ijt}$ because there is at least 1 customer for each occupied table. By applying the two-step conjugate Gibbs update for r_{jt},[18] we first draw l_{ijt} according to this CRT distribution. Next, NB distribution can be derived from a random convolution of logarithmic random variables.[19] Specially, Zhou and Carin[18] note that, conditional on r_{jt} and ψ_{ijt},

$$\begin{aligned} l_{ijt} &\sim \text{Poisson}[-r_{jt} ln(1 - \psi_{ijt})] \\ \psi_{ijt} &\sim \frac{e^{\mathbf{z}_i \boldsymbol{\Gamma}_{jt} \boldsymbol{\beta}_{jt}}}{1 + e^{\mathbf{z}_i \boldsymbol{\Gamma}_{jt} \boldsymbol{\beta}_{jt}}}. \end{aligned} \tag{11}$$

Thus, if we assume a $Ga(e,f)$ prior for r_{jt}, then the full conditional for r_{jt} is

$$r_{jt}|\mathbf{l}_{jt}, \boldsymbol{\psi}_{jt} \sim Ga\left[e + \sum_{i=1}^{n} l_{ijt}, f - \sum_{i=1}^{n} ln(1 - \psi_{ijt})\right],$$

the Gibbs update first draws l_{ijt} independently from a CRT distribution, and then r_{jt} from its full conditional Gamma distribution given $\mathbf{l}_{jt}$ and $\boldsymbol{\psi}_{jt}$.

2.4 | Pathway-level data integration

One of the most important features of our approach is the utilization of pathway-level information. For the pathway-level analysis, we define G as the collection of pathways, where $G = \{G_1, \ldots, G_s\}$ for s pathways and G_i is the number of genes in the ith pathway ($i = 1, \ldots, s$). In our pathway model, we incorporate prior biological knowledge by specifying the factor loading matrix based on the known pathway annotation. Specifically, if the pathway annotations are disjointed, we have factor loading matrix $\boldsymbol{\beta}_{jt}$ of dimension p by $s + 1$, where $p = \sum_i^s G_i$. p is the total number of unique genes in s pathways. Then we put constraints on the factor loading matrix, where the genes in $\boldsymbol{\beta}_{jt}$ that belong to certain pathways are free to update, while the remaining elements in $\boldsymbol{\beta}_{jt}$ are forced to be zero. For example, we only update the first G_1 elements of the first latent factor in the factor loading matrix $\boldsymbol{\beta}$ from $N(0, 1)$, and force the remaining elements of the first latent factor to be zero. Identifiability is the main statistical issue in latent factor models, Lopes and West[20] argued that one should enforce the constraint on the loading matrix. However, recent publications seem to avoid this constrain[21,22] arguing that in practice the sparsity of the loadings is sufficient to enforce the identifiability of the latent factor model. To address the issue of identifiability, we carry out two strategies to check if the sparsity of the loadings is sufficient to enforce the identifiability of the latent factor model: (i) no constraint is put on the loading matrix; and (ii) we update the first nonzero elements of each latent factor using the truncated Gaussian distribution $TN(0, 1, 0, \infty)$. In the pathway-level analysis model, next, we implement the gene-cluster approach proposed by Wei et al[12] to deal with the issue of overlapping genes among the pathways and it provides nonoverlapping gene sets, that is, each gene belongs to one pathway only.

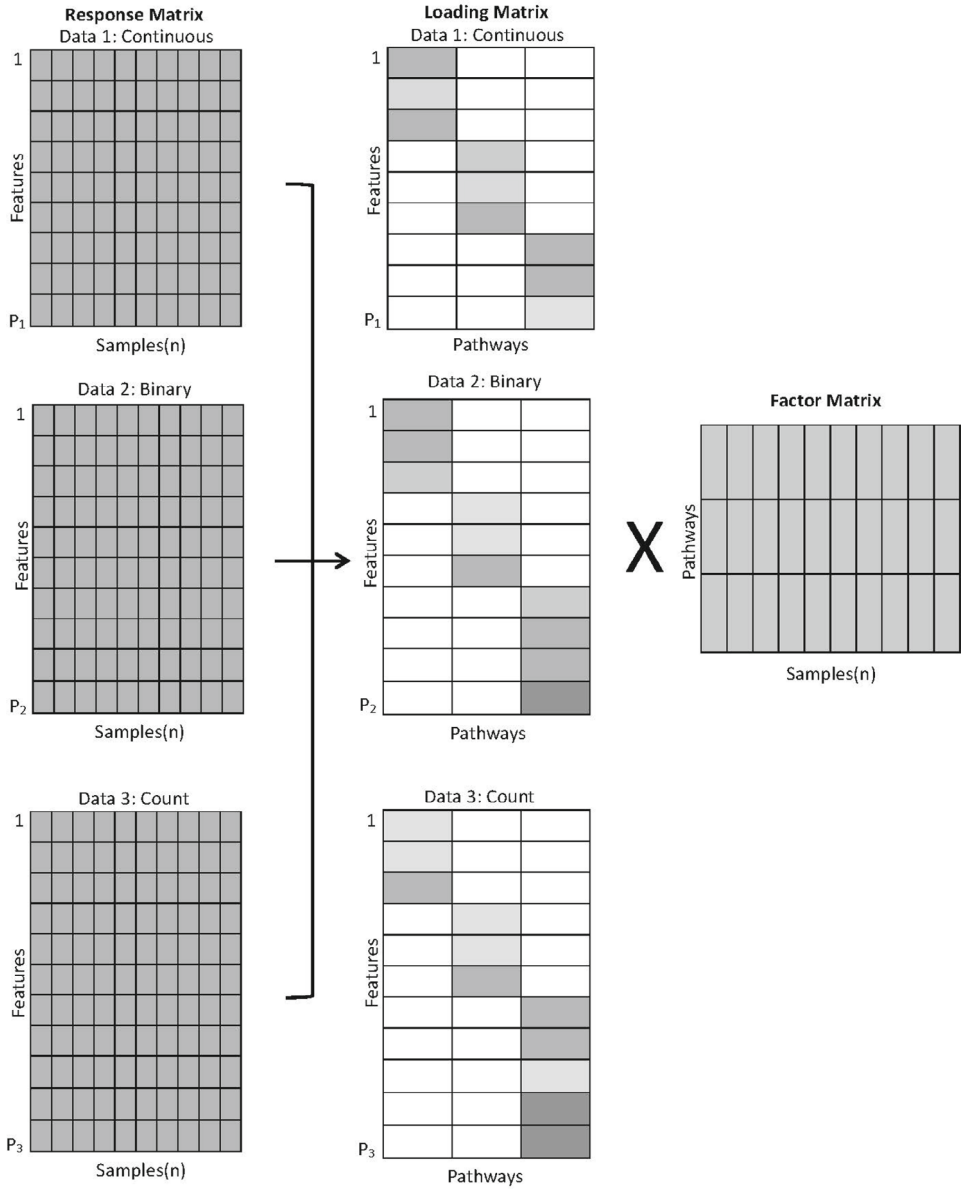

FIGURE 1 Illustration of the Bayes-InGRiD framework. We use the factor loading matrix for key feature selection, where we set the number of latent factors equal to the number of pathways assuming the pathways are independent. We identify patient subgroups by applying a *k*-means approach to the posterior estimate of the shared latent factor matrix. The pathway-level analysis setting helps make the factor loading matrix significantly sparser and addresses the challenging issue of selecting the number of factors.

Based on this rationale, we set the number of latent factors k equal to the number of pathways assuming the pathways are independent. Finally, we cluster the patients into c subgroups by applying a k-means approach on the posterior estimate of the shared latent factor matrix $\mathbf{Z}$. Notice that the pathway-level analysis setting helps make the factor loading matrix significantly sparser and addresses the challenging issue of selecting the number of factors. A graphical description of the model is shown in Figure 1.

For joint analysis of continuous, binary, and count data, we first focus on the common latent factor matrix $\mathbf{Z}$. The latent factor matrix $\mathbf{Z}$ is used to capture the shared structure of multiple data sets to achieve joint dimension reduction. This is an appealing approach due to the interdependency among genomic features. For example, copy number alterations, somatic mutations, and DNA methylations regulate gene expression, which again in turn affect protein expression.[1] Hence, joint dimension reduction allows effective information sharing among these interdependent genomic features and will lead to more accurate and robust patient clustering compared to the case that only gene expression data is used, which has been traditionally implemented. Based on this rationale, by joint analysis of multiple data, the cancer patients can be clustered

in the latent variable space and key molecular features that drive the patient clustering are identified through variable selection on the loading matrix in a given set t. In particular, the model for feature j of ith sample is given as

$$\mathbf{u}_{ij} - \boldsymbol{\beta}_{0j} = \boldsymbol{\Gamma}_j \boldsymbol{\beta}_j z_i + \boldsymbol{\epsilon}_{\mathbf{ij}}, i = 1, \ldots, n, j = 1, \ldots, p_t, \tag{12}$$

where $\mathbf{u}_{ij} = (u_{ij1}, \ldots, u_{ijm})$ is a length m vector, and its tth element depends on the tth data type as

$$u_{ijt} = \begin{cases} y_{ijt}, & \text{if } t\text{th data type is continuous,} \\ \dfrac{y_{ijt} - \frac{1}{2}}{\omega_{ijt}}, & \text{if } t\text{th data type is binary,} \\ \dfrac{y_{ijt} - r_{jt}}{2\omega_{ijt}}, & \text{if } t\text{th data type is count.} \end{cases} \tag{13}$$

We let $\beta_{0j} = (\beta_{0j1}, \ldots, \beta_{0jm})^T$ be the intercept vector; $\mathbf{z}_i = (z_{i1}, \ldots, z_{ik})$ be the latent variable for patient i; $\boldsymbol{\Gamma}_j = \text{diag}(\gamma_{j1}, \ldots, \gamma_{jm})$ be a diagonal matrix and its tth diagonal element γ_{jt} depends on the tth data type; ϵ_{ijt} be the error term with mean 0 and its variance depends on the tth data type as

$$\epsilon_{ijt} \begin{cases} N\left(0, \sigma_{jt}^2\right), & \text{if } t\text{th data type is continuous,} \\ N\left(0, \omega_{ijt}^{-1}\right), & \text{if } t\text{th data type is binary,} \\ N\left(0, \omega_{ijt}^{-1}\right), & \text{if } t\text{th data type is count.} \end{cases} \tag{14}$$

We define $\boldsymbol{\Sigma} = \text{diag}(Var(\epsilon_{ij1}), \ldots, Var(\epsilon_{ijm}))$, where Σ is the diagonal variance-covariance matrix whose diagonal components are the variance of random errors. As for prior distributions of model parameters, we assume $\boldsymbol{\beta}_{jt} \sim MVN(\boldsymbol{\beta}_{0t}, \boldsymbol{\Sigma}_{0t})$, $\sigma_{jt}^2 \sim IG(v_0/2, v_0\sigma_0^2)$, and $\gamma_{jt} \sim Bernoulli(q_t)$, where the coefficient vector $\boldsymbol{\beta}_{jt}$, σ_{jt}^2, the indicator variable γ_{jt} follow a multivariate normal distribution, inverse-gamma distribution, and Bernoulli distribution, respectively. Hence, we have the conditional posterior distributions of variance term depending on the tth data type as the following:

$$P(\sigma_{jt}^2|\mathbf{Z}, \mathbf{y}_{jt}, \boldsymbol{\beta}_{jt}, \boldsymbol{\Gamma}_{jt}) = IG\left(\frac{V_0 + n}{2}, \frac{V_0\sigma_0^2 + \left(\mathbf{y}_{jt} - \mathbf{Z}\boldsymbol{\Gamma}_{jt}\boldsymbol{\beta}_{jt}\right)^T \left(\mathbf{y}_{jt} - \mathbf{Z}\boldsymbol{\Gamma}_{jt}\boldsymbol{\beta}_{jt}\right)}{2}\right), \text{if } t\text{th data type is continuous,}$$

$$P(\omega_{ijt}|\mathbf{Z}, \boldsymbol{\beta}_{jt}, \boldsymbol{\Gamma}_{jt}) \sim PG(1, \mathbf{z}_i\boldsymbol{\Gamma}_{jt}\boldsymbol{\beta}_{jt}), \quad \text{if } t\text{th data type is binary,}$$

$$P(\omega_{ijt}|\mathbf{Z}, \mathbf{y}_{jt}, \boldsymbol{\beta}_{jt}, \boldsymbol{\Gamma}_{jt}, r_{jt}) \sim PG(y_{ijt} + r_{jt}, \mathbf{z}_i\boldsymbol{\Gamma}_{jt}\boldsymbol{\beta}_{jt}), \quad \text{if } t\text{th data type is count.}$$

We have the conditional posterior distributions of $\boldsymbol{\beta}_{jt}$ as follows.

$$\begin{aligned} P\left(\boldsymbol{\beta}_{jt}|\mathbf{Z}, u_{ijt}, \boldsymbol{\Sigma}, \boldsymbol{\Gamma}_{jt}\right) &\sim MVN\left(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta\right), \text{where} \\ \boldsymbol{\mu}_\beta &= \left(\boldsymbol{\Gamma}_{jt}^T\mathbf{Z}^T\boldsymbol{\Sigma}^{-1}\mathbf{Z}\boldsymbol{\Gamma}_{jt} + \boldsymbol{\Sigma}_{0t}^{-1}\right)^{-1}\left(\boldsymbol{\Gamma}_{jt}^T\mathbf{Z}^T\boldsymbol{\Sigma}^{-1}u_{ijt} + \boldsymbol{\Sigma}_{0t}^{-1}\boldsymbol{\beta}_{0t}\right), \\ \boldsymbol{\Sigma}_\beta &= \left(\boldsymbol{\Gamma}_{jt}^T\mathbf{Z}^T\boldsymbol{\Sigma}^{-1}\mathbf{Z}\boldsymbol{\Gamma}_{jt} + \boldsymbol{\Sigma}_{0t}^{-1}\right)^{-1}. \end{aligned}$$

Next, we define $\boldsymbol{\beta}_j$ is an $m \times k$ matrix in which the tth row is $(\beta_{1jt}, \ldots, \beta_{kjt})$. In words, it is $\boldsymbol{\beta}_{jt}$ without its intercept. By utilizing the Pólya-Gamma mixture of Normal distributions, we can derive the exact posterior distribution of $\mathbf{z}_i$ as the following:

$$\begin{aligned} P(\mathbf{z}_i|\boldsymbol{\beta}_j, \mathbf{y}_{ij}, \boldsymbol{\Sigma}, \boldsymbol{\Gamma}_j) &\sim MVN(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n), \text{where} \\ \boldsymbol{\mu}_n &= \left\{\sum_j^p (\boldsymbol{\Gamma}_j\boldsymbol{\beta}_j)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Gamma}_j\boldsymbol{\beta}_j + \mathcal{I}\right\}^{-1}\left\{\sum_j^p (\boldsymbol{\Gamma}_j\boldsymbol{\beta}_j)^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}_{ij} - \boldsymbol{\beta}_{0j})\right\}, \\ \boldsymbol{\Sigma}_n &= \left\{\sum_j^p (\boldsymbol{\Gamma}_j\boldsymbol{\beta}_j)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Gamma}_j\boldsymbol{\beta}_j + \mathcal{I}\right\}^{-1}. \end{aligned}$$

For parameter $\mathbf{\Gamma}_{jt}$, we use the Bayes rule to obtain samples from their posterior distributions, where we take $\mathbf{\Gamma}_{jt}$ from the previous iteration. We define $\widetilde{\mathbf{\Gamma}_{jt}} = \text{diag}(1, 1-\gamma_{jt}, \ldots, 1-\gamma_{jt})$ as a $(k+1)\times(k+1)$ diagonal matrix, and $\widetilde{\psi_{ijt}} = e^{\mathbf{z}_i\widetilde{\mathbf{\Gamma}_{jt}}\boldsymbol{\beta}_{jt}}/(1+e^{\mathbf{z}_i\widetilde{\mathbf{\Gamma}_{jt}}\boldsymbol{\beta}_{jt}})$. Finally, we have the conditional posterior distributions of the indicator variable term $\mathbf{\Gamma}_{jt}$ depending on the tth data type as follows.

$$P(\mathbf{\Gamma}_{jt}|\boldsymbol{\beta}_{jt}, \mathbf{Z}, \mathbf{y}_{jt}, \sigma^2_{jt}) = \frac{\exp\left\{-\frac{1}{2\sigma^2_{jt}}(\mathbf{y}_{jt}-\mathbf{Z}\mathbf{\Gamma}_{jt}\boldsymbol{\beta}_{jt})^T(\mathbf{y}_{jt}-\mathbf{Z}\mathbf{\Gamma}_{jt}\boldsymbol{\beta}_{jt})\right\}}{exp\left\{-\frac{1}{2\sigma^2_{jt}}(\mathbf{y}_{jt}-\mathbf{Z}\mathbf{\Gamma}_{jt}\boldsymbol{\beta}_{jt})^T(\mathbf{y}_{jt}-\mathbf{Z}\mathbf{\Gamma}_{jt}\boldsymbol{\beta}_{jt})\right\}) + exp\left\{-\frac{1}{2\sigma^2_{jt}}(\mathbf{y}_{jt}-\mathbf{Z}\widetilde{\mathbf{\Gamma}_{jt}}\boldsymbol{\beta}_{jt})^T(\mathbf{y}_{jt}-\mathbf{Z}\widetilde{\mathbf{\Gamma}_{jt}}\boldsymbol{\beta}_{jt})\right\}},$$

if tth data type is continuous,

$$P(\mathbf{\Gamma}_{jt}|\boldsymbol{\beta}_{jt}, \mathbf{Z}, \mathbf{y}_{jt}) = \frac{\prod_{i=1}^n \frac{exp(\mathbf{z}_i\mathbf{\Gamma}_{jt}\boldsymbol{\beta}_{jt})^{y_{ijt}}}{1+exp(\mathbf{z}_i\mathbf{\Gamma}_{jt}\boldsymbol{\beta}_{jt})}}{\prod_{i=1}^n \frac{exp(\mathbf{z}_i\mathbf{\Gamma}_{jt}\boldsymbol{\beta}_{jt})^{y_{ijt}}}{1+exp(\mathbf{z}_i\mathbf{\Gamma}_{jt}\boldsymbol{\beta}_{jt})} + \prod_{i=1}^n \frac{exp(\mathbf{z}_i\widetilde{\mathbf{\Gamma}_{jt}}\boldsymbol{\beta}_{jt})^{y_{ijt}}}{1+exp(\mathbf{z}_i\widetilde{\mathbf{\Gamma}_{jt}}\boldsymbol{\beta}_{jt})}}, \text{ if } t\text{th data type is binary,}$$

$$P(\mathbf{\Gamma}_{jt}|\boldsymbol{\beta}_{jt}, \mathbf{Z}, \mathbf{y}_{jt}, r_{jt}) = \frac{\prod_{i=1}^n \frac{\Gamma(y_{ijt}+r_{jt})}{\Gamma(r_{jt})y_{ijt}!}(1-\psi_{ijt})^{r_{jt}}\psi_{ijt}^{y_{ijt}}}{\prod_{i=1}^n \frac{\Gamma(y_{ijt}+r_{jt})}{\Gamma(r_{jt})y_{ijt}!}(1-\psi_{ijt})^{r_{jt}}\psi_{ijt}^{y_{ijt}} + \prod_{i=1}^n \frac{\Gamma(y_{ijt}+r_{jt})}{\Gamma(r_{jt})y_{ijt}!}(1-\widetilde{\psi_{ijt}})^{r_{jt}}\widetilde{\psi_{ijt}}^{y_{ijt}}}, \text{ if } t\text{th data type is count.}$$

By introducing Pólya-Gamma latent variables, we derived all the posterior distributions in a closed form, thus we can use the Gibbs sampling algorithm to obtain samples from their posterior distributions in MCMC.

3 | SIMULATION STUDIES

To compare feature selection performance between the pathway-level and the gene-level analyses, we performed simulation studies on separate data types including continuous, binary, and count data. We constructed each data set with 120 molecular features and 50% of them are informative features to define the patient subgroups. We assumed that the samples were from three subgroups of patients (A, B, and C) and each of the subgroups had 20 samples.

For continuous data, we let patient subgroup A be characterized by the first 20 genes with an amplified signal. Patient subgroup B was characterized by the second 20 genes with reduced signal, and subgroup C was characterized by the third 20 genes with the amplified signal. To be specific, features with amplified and reduced signals were randomly generated from $N(\mu, 1)$ and $N(-\mu, 1)$, respectively. We used different signal levels to evaluate model performance, where we let $\mu = 0.8, 1, 1.2, 1.5$. The background noise was randomly generated from $N(0, 1)$.

For the pathway-level analysis, we used the prior knowledge that matches with our gene specification in the simulation setting. We define the first 20 genes as pathway 1, the second 20 genes as pathway 2, the third 20 genes as pathway 3, and the remaining 60 genes as pathway 4. In our pathway model, we set the number of latent factors equal to the number of pathways (4). To specify the factor loading matrix based on the known pathway annotation, we only updated the first 20 elements of the first latent factor, the second 20 elements of the second latent factor, the third 20 elements of the third latent factor, and the last 60 elements of the fourth latent factor in factor loading matrix β from $N(0, 1)$, while forcing the rest of elements of the factor loading matrix to be zero. Identifiability is the main statistical issue in latent factor models, Lopes and West[20] argued that one should enforce the constraint on the loading matrix. However, recent publications seem to avoid this constrain[21,22] arguing that in practice the sparsity of the loadings is sufficient to enforce the identifiability of the latent factor model. To address the issue of identifiability, we carried out two strategies to check if the sparsity of the loadings is sufficient to enforce the identifiability of the latent factor model: (i) no constraint was put on the loading matrix; and (ii) $TN(0, 1, 0, \infty)$ was used to update the first nonzero element of each latent factor, which is the first element of the first latent factor, 21st element of the second latent factor, 41st element of the third latent factor, and 61st element of the last latent factor in the factor loading matrix. As a result, there was minimal difference between

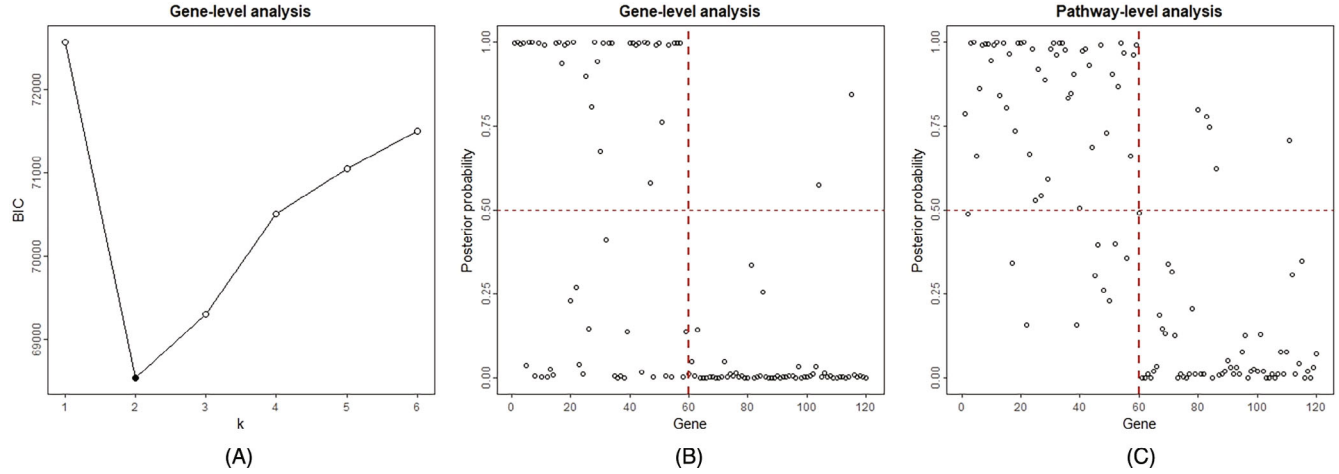

FIGURE 2 Model and variable selection for the continuous data with $N(1, 1)$ and $N(-1, 1)$ as the signal, and $N(0, 1)$ as the background. (A) The BIC curve as a function of the number of latent components (k). The gene-level analysis model fit the data best when $k = 2$. (B) Posterior probabilities of being informative features when $k = 2$ for gene-level analysis. Genes with posterior probabilities greater than 0.5 were considered key features. (C) Posterior probabilities of being informative features for pathway-level analysis.

TABLE 1 Feature selection performance for continuous data.

	Bayes-InGRiD gene-level		Bayes-InGRiD pathway-level		iClusterBayes	
Signal level	**Sensitivity (%)**	**Specificity (%)**	**Sensitivity (%)**	**Specificity (%)**	**Sensitivity (%)**	**Specificity (%)**
μ=0.8	30.4	96.7	60.7	89.9	26.7	95.0
μ=1.0	57.1	96.7	82.1	91.5	61.7	96.7
μ=1.2	82.1	96.7	96.4	96.7	83.3	96.7
μ=1.5	100.0	96.7	96.4	96.7	100.0	96.7

Note: Sensitivity and specificity for the continuous data with $N(\mu, 1)$ and $N(-\mu, 1)$ as the signal, $N(0, 1)$ as the background.

results obtained with and without constraint on the loading matrix. We used the uninformative priors for σ_{jt}^2 and γ_{jt}, where we set Inverse-gamma(1, 1) for σ_{jt}^2 and *Bernoulli*(0.5) for the indicator variable γ_{jt}. In each simulation, we ran 20 000 MCMC iterations, and the first 10 000 were removed as burn-in. In general, we observed very fast convergence of the Markov chains within 100 iterations (Figure S7).

Cluster membership was assigned by applying a standard k-means clustering on the posterior mean of the latent factor matrix $\mathbf{Z}$. In other words, cluster partition in the final step is performed in the integrated latent variable subspace of dimension n by $k + 1$. We use k-means clustering to divide the n samples into c clusters in the latent variable space, where $c = k + 1$. Figure 2A shows the Bayesian information criterion (BIC) values for the gene-level analysis of the continuous data with $N(1, 1)$ and $N(-1, 1)$ as the signal, and $N(0, 1)$ as the background. We observed the minimum BIC value when $k = 2$, which leads to the number of clusters c = 3. In summary, our results actually coincide with the ground truth. For all the gene-level analysis models, we used BIC to determine the optimal choice of k. Figure 2B,C present the posterior probabilities of the genomic features for the continuous data with $N(1, 1)$ and $N(-1, 1)$ as the signal, and N(0, 1) as the background. Table 1 illustrates when signal level $\mu = 1$, and pathway-level analysis showed a significantly higher level of sensitivity (82.1%) compared to gene-level analysis (57.1%) while specificity was comparable between two cases (96.7% and 91.5% for gene- and pathway-level analyses, respectively). It showed that the pathway-level model performs better in detecting informative features especially when signals are weak. This occurred because the information sharing using the pathway information improved the statistical power to detect the true signals. As the signal gets stronger, the performance of the gene-level analysis improves and becomes comparable to the pathway-level analysis in the sense of sensitivity. The gene-level analysis provides slightly higher specificity in general but with a significant sacrifice of sensitivity.

To set the driver features in the simulation study for binary data, we had the first 20 genes in patient subgroup A, the second 20 genes in patient subgroup B, and the third 20 genes in patient subgroup C to be characterized by a

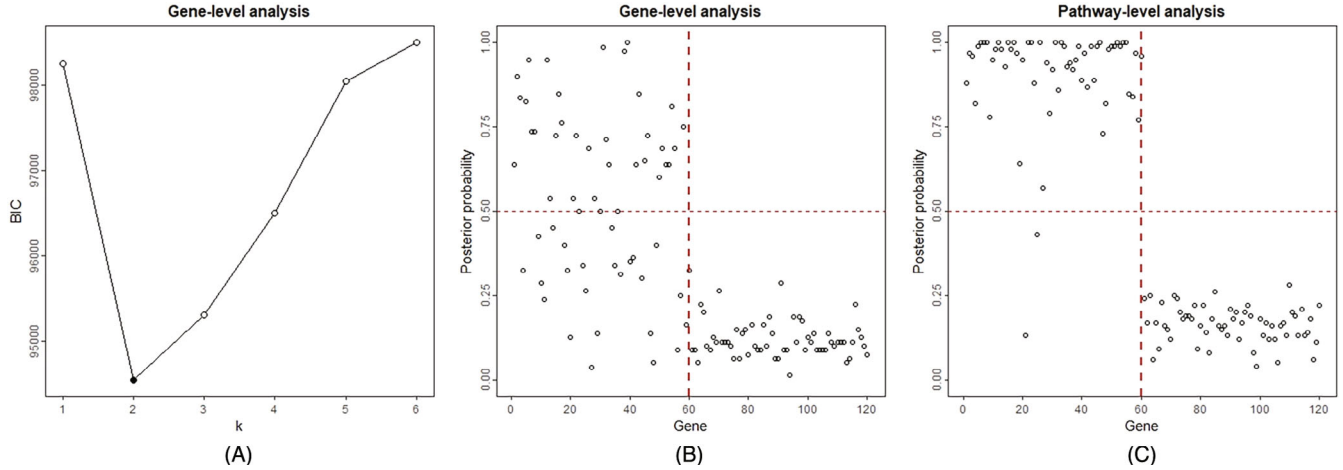

FIGURE 3 Model and variable selection for the binary data with *Bernoulli*(0.4) as signal and *Bernoulli*(0.02) as the background. (A) The BIC curve as a function of the number of latent components (k). The gene-level analysis model fit the data best when $k = 2$. (B) Posterior probabilities of being informative features when $k = 2$ for gene-level analysis. Genes with posterior probabilities greater than 0.5 were considered informative features. (C) Posterior probabilities of being informative features for pathway-level analysis.

TABLE 2 Feature selection performance for binary data.

	Bayes-InGRiD gene-level		Bayes-InGRiD pathway-level		iClusterBayes	
Signal level	**Sensitivity (%)**	**Specificity (%)**	**Sensitivity (%)**	**Specificity (%)**	**Sensitivity (%)**	**Specificity (%)**
S=0.4	57.1	100.0	96.4	100.0	75.0	96.7
S=0.5	89.4	100.0	100.0	100.0	81.6	100.0
S=0.6	96.4	100.0	100.0	100.0	93.3	100.0
S=0.7	100.0	100.0	100.0	100.0	100.0	100.0

Note: Sensitivity and specificity for the binary data with *Bernoulli*(*S*) as the signal, *Bernoulli*(0.02) as the background.

higher probability of being 1. Specifically, the genes with a higher probability of being 1 were randomly generated from *Bernoulli*(*P*). We let $P = 0.4, 0.5, 0.6, 0.7$ to check the model performance. The background genes were randomly generated from *Bernoulli*(0.02). Figure 3A shows the BIC values for the gene-level analysis of the binary data with *Bernoulli*(0.4) as the signal and *Bernoulli*(0.02) as the background. The best model fit was obtained when $k = 2$ based on BIC. Figure 3B,C demonstrate that the pathway-level model provided higher sensitivity in distinguishing informative features from uninformative features compared to the gene-level model when signals were generated from *Bernoulli*(0.4). Table 2 further indicates that the proposed pathway-level method can achieve high sensitivity and specificity in detecting the true signals compared to the gene-level approach.

For the simulation of count data, we let patient subgroup A be characterized by the first 20 genes with amplified signal, patient subgroup B be characterized by the second 20 genes with reduced signal, and subgroup C be characterized by the third 20 genes with the amplified signal. Specifically, the count data for genes with amplified- and reduced-signal were randomly generated from $NB(\mu = \mu_1)$ and $NB(\mu = \mu_2)$, respectively. We set different signal levels to evaluate model performance, where we set $\mu_1 = 7, 9, 11, 13$ and $\mu_2 = 1$. The data for background genes were randomly generated from a $NB(\mu = (\mu_1 + \mu_2)/2)$. Figure 4A shows the BIC values for the gene-level analysis of the count data with $NB(\mu = 11)$ and $NB(\mu = 1)$ as the signal, $NB(\mu = 6)$ as the background, respectively. The best model fit was obtained when $k = 2$ based on BIC. Figure 4B,C demonstrate better performance for the pathway-level model in detecting true signal genes. Table 3 further confirms this observation across different signal-to-noise ratios.

Next, we performed a simulation study for the joint analysis of continuous, binary, and count data. We used the same setting as the separate data analyses above, where each data set had 120 genomic features and 50% of them were informative features to define the patient subgroups. We defined that these samples were from three patient subgroups (A, B, and C) and each of the subgroups has 20 samples. To set the signal genes, we set the first 20 genes in patient subgroup

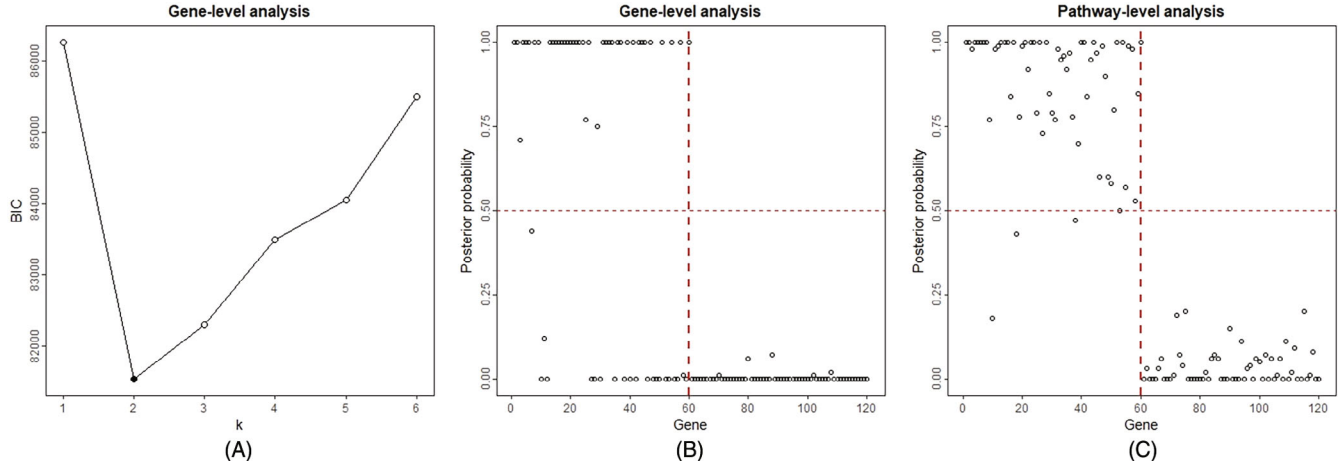

FIGURE 4 Model and variable selection for the count data with $NB(\mu = 11)$ and $NB(\mu = 1)$ as the signal, $NB(\mu = 6)$ as the background. (A) The BIC curve as a function of the number of latent components (k). The model fit the data best when $k = 2$. (B) Posterior probabilities of being informative features when $k = 2$ for gene-level analysis. Genes with posterior probabilities greater than 0.5 were considered informative features. (C) Posterior probabilities of being informative features for pathway-level analysis.

TABLE 3 Feature selection performance for count data.

	Bayes-InGRiD gene-level		Bayes-InGRiD pathway-level		iClusterBayes	
Signal level	**Sensitivity (%)**	**Specificity (%)**	**Sensitivity (%)**	**Specificity (%)**	**Sensitivity (%)**	**Specificity (%)**
$\mu_1 = 7, \mu_2 = 1$	55.4	100.0	87.5	98.9	86.7	85.0
$\mu_1 = 9, \mu_2 = 1$	76.8	100.0	89.3	100.0	90.0	83.3
$\mu_1 = 11, \mu_2 = 1$	83.9	100.0	92.9	100.0	95.0	83.3
$\mu_1 = 13, \mu_2 = 1$	92.9	98.3	96.4	100.0	95.0	81.6

Note: Sensitivity and specificity for the count data with $NB(\mu = \mu_1)$ and $NB(\mu = \mu_2)$ as the signal, $NB(\mu = (\mu_1 + \mu_2)/2)$ as the background.

A, the second 20 genes in patient subgroup B, and the third 20 genes in patient subgroup C to be characterized by the signal for each data. Specifically, we used $N(0.8, 1)$ and $N(-0.8, 1)$ as the signal, and $N(0, 1)$ as the background for the continuous data; *Bernoulli*(0.6) as the signal, and *Bernoulli*(0.02) as the background for the binary data; $NB(\mu = 11)$ and $NB(\mu = 1)$ as the signal, and $NB(\mu = 6)$ as the background for the count data. For the gene-level analysis of the integrated data analysis, we observed the minimum BIC value when the number of latent components k is equal to 2. For the pathway-level analysis, we set $s = 4$ as the number of pathways. Table 4 presents feature selection performance comparing integrated data analysis to separate data analysis. We observed higher sensitivity and specificity for integrated data analysis overall, demonstrating the benefit of added information through joint data analysis. In addition, pathway-level analysis was superior in selecting key molecular features compared to separate data analysis, especially when it came to separating out the true signal from the background. Furthermore, we carried out a comparison using k-means clustering on the posterior median of the latent factor matrix Z instead of the posterior mean and reached the same clustering results. It demonstrated the clustering results are robust to the choice of the posterior quantities of Z|Y.

Finally, we compared the feature selection performance of our proposed gene-level and pathway-level models with a cutting-edge approach, iClusterBayes.[8] Specifically, Tables 1–3 show the feature selection performance of iClusterBayes for the separate data analysis of continuous data, binary data, and count data, respectively. And Table S1 in the supplementary materials presents the feature selection performance of iClusterBayes for the integrated data analysis. Likewise, Figures S4-S6 show the discrimination curves comparison for continuous data, binary data, and count data, respectively. The sensitivity and specificity in Tables 1–3, S1, and the C-statistics in Figures S4-S6 demonstrate the superiority of the Bayes-InGRiD pathway-level analyses over Bayes-InGRiD gene-level analyses and iClusterBayes in identifying key molecular features, especially when the signal-to-noise ratio is low (eg, $N(0.8, 1)$ and $N(1.0, 1)$ as the signal for continuous data, *Bernoulli*(0.4) and *Bernoulli*(0.5) as the signal for binary data, and $NB(\mu_1 = 7, \mu_2 = 1)$ and

TABLE 4 Feature selection performance for integrated data analysis and separate data analysis.

	Separated data analysis				Integrated data analysis			
	Gene-level		Pathway-level		Gene-level		Pathway-level	
Signal level	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
Continuous data: $\mu = 0.8$	30.4	96.7	60.7	89.9	50.9	96.7	71.9	92.9
Binary data: P = 0.6	96.4	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Count data: $\mu = 11, \mu = 1$	69.6	100.0	92.9	100.0	96.4	98.3	98.3	93.3

Note: Samples are drawn from $N(0.8, 1)$ and $N(-0.8, 1)$ as the signal, $N(0, 1)$ as the background for continuous data; *Bernoulli*(0.6) as the signal, and *Bernoulli*(0.02) as the background for binary data; $NB(\mu = 11)$ and $NB(\mu = 1)$ as the signal, $NB(\mu = 6)$ as the background for count data, respectively. For gene-level analysis, we observed the optimal number of clusters equal to 3 for each signal level. For pathway-level analysis, the number of latent components k was set as 4.

$NB(\mu_1 = 9, \mu_2 = 1)$ as the signal for count data). In this case, we observe higher sensitivity, higher specificity, and higher C-statistics for pathway-level analyses. The performance of the gene-level analysis improves and becomes comparable to the pathway-level analysis as the signal gets stronger. In addition, the performance of Bayes-InGRiD gene-level analyses and iClusterBayes are comparable in most cases. We also compared the feature selection performance for the integrative data analysis (Tables 4 and S1) and we observe similar pattern as separate data analysis. Specifically, Bayes-InGRiD pathway-level analyses outperforms Bayes-InGRiD gene-level analyses and iClusterBayes in identifying key molecular features.

4 | REAL DATA ANALYSIS

In this section, we used a cohort of high-grade serous ovarian cancer (HGSOC) patients from the TCGA project[1] to demonstrate the benefit of the proposed Bayes-InGRiD approach. Specifically, gene expression (z-scores) and copy number alteration measurements (relative linear copy-number values) for 489 patients were obtained from the cBio Cancer Genomics Portal (http://cbioportal.org/). For pathway information, we used KEGG pathway annotations from the MSigDB database.[23,24] In this analysis, we considered only the 1045 genes from the 15 previously profiled cancer signaling pathways,[25] and the importance of these pathways has been discussed in the previous literature.[26,27]

To deal with the issue of overlapping genes among the 15 signaling pathways, we implemented the gene-cluster approach employed by InGRiD.[12] Specifically, if a gene is shared by multiple gene sets, this gene would be reallocated to a new pathway using the partitioning around medoids algorithm.[28] Two additional gene sets were identified by applying the gene-cluster approach, which are defined as "MAPK & APOPTOSIS" and "WNT & HEDGEHOG" gene sets. The "MAPK & APOPTOSIS" gene set mainly contains genes from the MAPK (62 genes) and the APOPTOSIS pathways (42 genes), and most genes in the "WNT & HEDGEHOG" gene set are from the WNT_SIGNALING (46 genes) and the HEDGEHOG_SIGNALING (32 genes) pathways. The gene lists of the two additional gene sets can be found in Tables S2 and S3 of the supplementary materials. We used the prior *Bernoulli*(0.5) for the indicator variable $\Gamma_{jt}, j = 1, \ldots, p_t, t = 1, \ldots, m$, and we ran 20 000 MCMC iterations with the first 10 000 iterations considered as burn-in. In general, we observed Markov chains converged after a couple of hundred iterations (Figure S8). The optimal number of clusters $c = 2$ was the one that maximizes the average silhouette over a range of possible values. We tested the cluster number parameter from 1 to 10.

The Bayes-InGRiD results for the mRNA expression data are presented in Table 5. Gene sets are selected if more than one gene is selected. We ranked the pathway based on the weighted averages of factor loadings of selected genes, namely "pathway coefficient." In Bayes-InGRiD, both gene-level and pathway-level analyses are performed simultaneously within the unified model, since prior pathway knowledge is embedded in the latent factor setting. In addition, pathway information guides the factor loading specification and the number of latent factors in the model. We use factor loading to select key genes, and we use the weighted average of absolute factor loading values to determine pathway ranking. Bayes-InGRiD identified 387 unique genes from 14 gene sets based on the mRNA expression data. CELL_CYCLE and CELL_ADHESION_MOLECULES_CAMS pathways are the two pathways with the highest pathway coefficient and the number of genes selected.

To make sense of the patient subgrouping, we used the four expression subtypes classified in annotated TCGA subtypes by Noushmehr and Malta,[29] where the patients are clustered into differentiated, immunoreactive, mesenchymal,

TABLE 5 Top pathways and genes selected for the mRNA expression data.

Pathways selected	Genes selected	Pathway coefficient	Top three genes		
CELL_ADHESION_MOLECULES_CAMS	76 (122)	0.524	*HLA.DRB1*	*HLA.DPB1*	*HLA.DPA1*
CELL_CYCLE	69 (86)	0.506	*ORC1*	*CDC25C*	*PLK1*
NUCLEOTIDE_EXCISION_REPAIR	6 (20)	0.493	*CUL4A*	*ERCC5*	*ERCC1*
MAPK_SIGNALING_PATHWAY	129 (192)	0.431	*FGF4*	*PLA2G12B*	*CACNG3*
MISMATCH_REPAIR	4 (8)	0.367	*MSH6*	*MSH2*	*MSH3*
APOPTOSIS	7 (39)	0.353	*BIRC2*	*ENDOD1*	*BIRC3*
WNT_SIGNALING_PATHWAY	5 (21)	0.321	*APC*	*CTNNBIP1*	*CSNK2B*
MTOR_SIGNALING_PATHWAY	8 (31)	0.301	*PRKAA1*	*RICTOR*	*RPTOR*
NOTCH_SIGNALING_PATHWAY	11 (37)	0.280	*DLL3*	*PSENEN*	*PSEN2*
WNT&HEDGEHOG	11 (85)	0.237	*PCNA*	*PLCB1*	*PLCB4*
MAPK&APOPTOSIS	20 (74)	0.224	*PPP3CA*	*NFKB1*	*CASP3*
JAK_STAT_SIGNALING_PATHWAY	18 (121)	0.214	*JAK2*	*IFNB1*	*IFNE*
PHOSPHATIDYLINOSITOL_SIGNALING_SYSTEM	14 (59)	0.196	*ITPR2*	*DGKH*	*PIP5K1A*
HEDGEHOG_SIGNALING_PATHWAY	9 (20)	0.187	*SUFU*	*GAS1*	*STK36*
BASE_EXCISION_REPAIR	0 (25)				
NON_HOMOLOGOUS_END_JOINING	0 (13)				
TGF_BETA_SIGNALING	0 (51)				

Note: Pathways were ranked based on the "pathway coefficient," the weighted averages of the factor loadings for the selected genes. "Genes selected" refers to the number of genes selected in each pathway, where the total number of genes in each pathway was also included within the parenthesis in the column "genes selected." Genes that rank top three in coefficient estimates would be shown in column "top three genes."

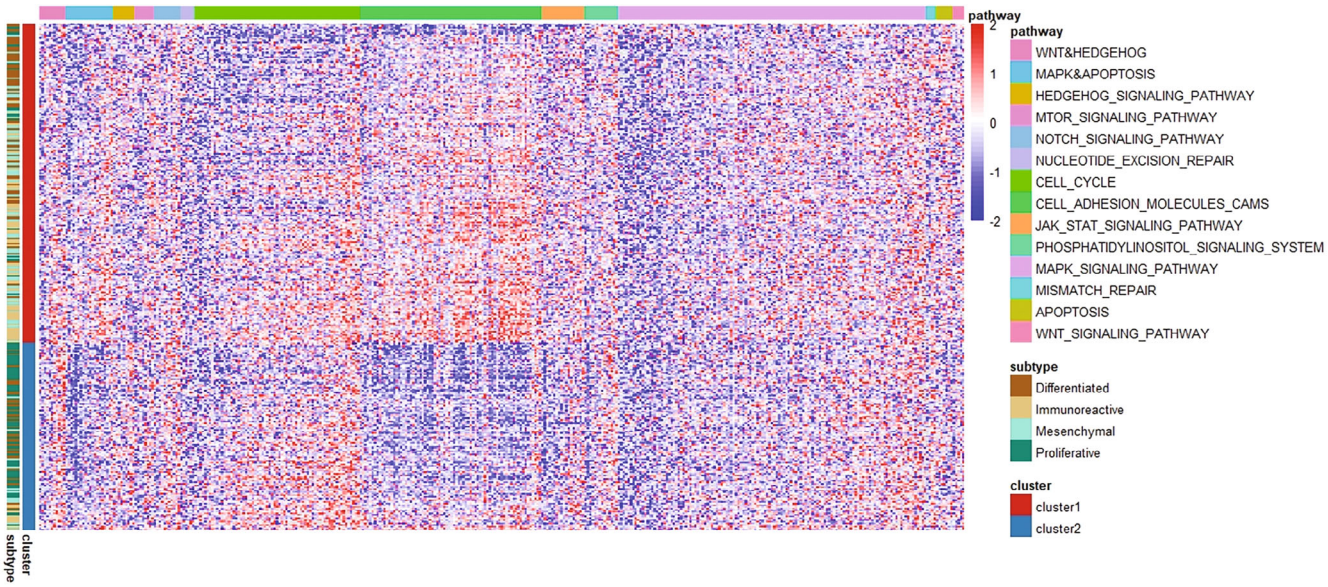

FIGURE 5 Heatmap of the selected genes for mRNA data. The genomic pattern for gene expression is shown in the heatmap (red, high-level expression; blue, low-level expression). The color bar on the left labeled "subtype" contains four expression subtypes classified in annotated TCGA subtypes including differentiated, immunoreactive, mesenchymal, and proliferative. The color bar on the left labeled "cluster" shows the patient subgroups identified using Bayes-InGRiD. The color bar on the top shows the selected 14 pathways.

TABLE 6 Top pathways and genes selected for the copy number data.

Pathways selected	Genes selected	Pathway coefficient	Top three genes		
MISMATCH_REPAIR	2 (8)	0.574	*MSH2*	*MSH6*	
JAK_STAT_SIGNALING_PATHWAY	24 (121)	0.438	*IFNA6*	*IFNA2*	*IFNE*
APOPTOSIS	6 (39)	0.354	*BIRC2*	*BIRC3*	*ENDOD1*
MTOR_SIGNALING_PATHWAY	3 (31)	0.352	*RICTOR*	*PRKAA1*	*MTOR*
WNT_SIGNALING_PATHWAY	2 (21)	0.284	*APC*	*CAMK2A*	
PHOSPHATIDYLINOSITOL_SIGNALING_SYSTEM	13 (59)	0.250	*PLCZ1*	*PIK3C2G*	*ITPR2*
NUCLEOTIDE_EXCISION_REPAIR	5 (20)	0.209	*CUL4A*	*ERCC5*	*ERCC2*
WNT&HEDGEHOG	20 (85)	0.205	*PLCB4*	*PLCB1*	*BMP2*
HEDGEHOG_SIGNALING_PATHWAY	5 (20)	0.191	*PTCH1*	*GAS1*	*SUFU*
NOTCH_SIGNALING_PATHWAY	23 (37)	0.182	*DLL3*	*PSENEN*	*NUMB*
MAPK&APOPTOSIS	27 (74)	0.127	*NFKB1*	*PPP3CA*	*MAPK10*
MAPK_SIGNALING_PATHWAY	8 (192)	0.126	*RASGRP4*	*MAP4K1*	*PTPRR*
CELL_CYCLE	6 (86)	0.103	*CCNE1*	*CDK1*	*FZR1*
CELL_ADHESION_MOLECULES_CAMS	22 (122)	0.079	*CDH4*	*CLDN23*	*ICAM1*
BASE_EXCISION_REPAIR	0 (25)				
NON_HOMOLOGOUS_END_JOINING	0 (13)				
TGF_BETA_SIGNALING	0 (51)				

Note: Pathways are ranked based on the "pathway coefficient," the weighted averages of the factor loadings for the selected genes. "Genes selected" refers to the number of genes selected in each pathway, where the total number of genes in each pathway is also included within the parenthesis in the column "genes selected." Genes that rank top three in coefficient estimates would be shown in column "top three genes."

and proliferative subtypes. Figure 5 demonstrates the heatmap of the selected genes for mRNA gene expression data. Integrative cluster 1 is highly correlated with immunoreactive and mesenchymal expression subtypes based on the overlapping color bars in Figure 5. Integrative cluster 2 is strongly correlated with the expression subtypes differentiated and proliferative. Figure 5 also presents different patterns of alterations across the two clusters especially in MAPK & APOPTOSIS gene set, CELL_CYCLE pathway, and CELL_ADHESION_MOLECULES_CAMS pathway. Figure S2 in the supplementary materials indicates the coefficients for the selected genes in each pathway.

Pathway and gene selection results of the copy number data are presented in Table 6. Bayes-InGRiD identified 166 unique genes from the 14 gene sets using the copy number data. While CELL_ADHESION_MOLECULES_CAMS and CELL_CYCLE pathways are the two pathways with the highest pathway coefficient in the gene expression data, they have the lowest pathway coefficient in copy number data. Figure S3 in the supplementary materials indicates that the coefficients are low for almost all the selected genes in those two pathways. Furthermore, it demonstrates the importance of incorporating prior biological knowledge into our estimate for the posterior inference of genes and pathways. More specifically, pathways such as CELL_ADHESION_MOLECULES_CAMS and CELL_CYCLE with dense and weak signals can be paid less attention to although 22 out of 122 genes are selected for that pathway. In contrast to the dense weak signal we found in CELL_ADHESION_MOLECULES_CAMS and CELL_CYCLE pathways, MISMATCH_REPAIR, APOPTOSIS, MTOR_SIGNALING_PATHWAY, and WNT_SIGNALING_PATHWAY are the pathways with only a few genes selected, however, the relatively high coefficients for the selected genes indicate sparse and strong signals for these pathways. Figure 6 shows the heatmap of the selected genes for copy number alteration data, we can observe different patterns of alterations in the two clusters, especially in "MTOR_SIGNALING" pathway.

HGSOC is among the most lethal human cancers and its prognosis remains extremely poor.[30] Tumor heterogeneity and rapid acquisition of resistance to conventional chemotherapy contribute to poor patient outcomes.[31] From the perspective of biological interpretation of the Bayes-InGRiD output data, it is important to consider that the outputs are highly relevant in the context of cancer and particularly HGSOC. Tables 5 and 6 list the top pathways and genes resulting

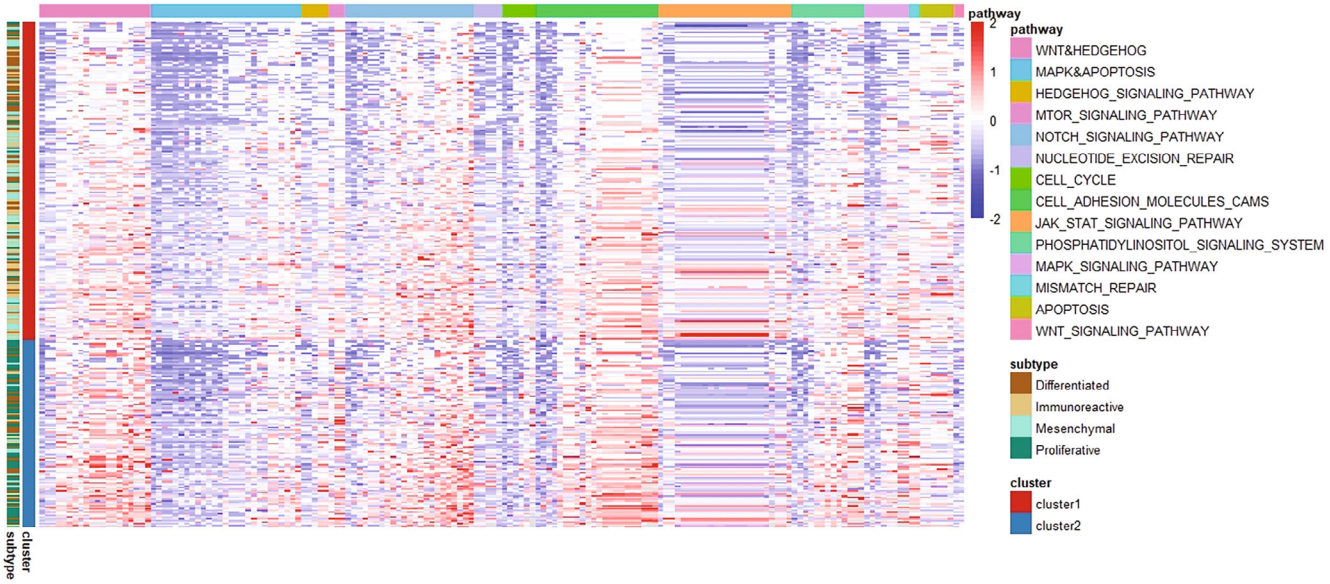

FIGURE 6 Heatmap of the selected genes for the copy number alteration data. The genomic pattern for gene expression is shown in the heatmap (red, high-level expression; blue, low-level expression). The color bar on the left labeled "subtype" contains four expression subtypes classified in annotated TCGA subtypes including differentiated, immunoreactive, mesenchymal, and proliferative. The color bar on the left labeled "cluster" shows the patient subgroups identified using Bayes-InGRiD. The color bar on the top shows the 14 selected pathways.

from the analysis of HGSOC mRNA expression data and copy number variation data respectively derived from the TCGA encompassing 489 patient samples.

In terms of CNV-identified biological pathways, these include MISMATCH_REPAIR, JAK_STAT_SIGNALING_PATHWAY, APOPTOSIS, and MTOR_SIGNALING_PATHWAY all of which are highly relevant to HGSOC. Malfunctioning of the MISMATCH_REPAIR pathway increases the mutational burden of specific cancers and is often involved in disease etiology, on occasion as an influential bystander and at times the main driving force.[32] The JAK_STAT_SIGNALING_PATHWAY is a key regulatory signaling cascade with roles in immune regulation, cell proliferation, cell survival, and apoptosis. Several studies have highlighted the involvement of the JAK_STAT_SIGNALING_PATHWAY in ovarian cancer pathogenesis. Additionally, aberrantly activated JAK/STAT signaling has been reported in several ovarian cancer cell lines as well as clinical tissue samples.[33] APOPTOSIS is a genetically programmed mechanism to eliminate damaged cells. Anti-apoptotic protein inhibition has emerged as a promising therapeutic direction for recurrent ovarian cancer.[34] The MTOR_SIGNALING_PATHWAY involves the Mammalian target of rapamycin (MTOR) which regulates cell proliferation, autophagy, and apoptosis. This pathway is frequently dysregulated in different subtypes of ovarian cancer.[35] At the individual gene-level both *MSH2* and *MSH6* encode proteins that play important roles in DNA repair. *MSH6* forms a dimer with *MSH2* and identifies locations in DNA where errors have occurred during the DNA replication process.[36] A significantly higher lifetime risk of the development of ovarian cancer has been reported in 10.4% of *MSH2* mutation carriers. The onset of ovarian cancer is more frequent (33%) in families with an *MSH6* mutation.[37] *IFNA6*, *IFNA2*, and *IFNE* are interferon proteins and members of the JAK_STAT_SIGNALING_PATHWAY. Copy number alterations of the IFN gene cluster is associated with increased mortality and decreased overall survival in cancer and currently a promising target for personalized immunotherapy.[38] *BIRC2* and *BIRC3* proteins play integral roles in regulation of apoptosis. *RICTOR*, *PRKAA1*, and *MTOR* encode proteins belonging to the MTOR_SIGNALING_PATHWAY. *RICTOR* contributes to cisplatin resistance in human ovarian cancer cells and represents a potential therapeutic target for chemoresistant ovarian cancer.[39] *PRKAA1* (protein kinase AMP-activated catalytic subunit α 1) is a catalytic subunit of AMP-activated protein kinase (AMPK), and functions in regulating cellular energy metabolism through phosphorylation. AMPK is a negative regulator of the Warburg Effect and has been shown to suppress tumor growth in vivo.[39] Targeting AMPK signaling in combating ovarian cancers represents another opportunity for therapeutic intervention in ovarian cancer.[40]

In terms of mRNA-identified biological pathways, these include CELL_ADHESION_MOLECULES_CAMS, CELL_CYCLE, NUCLEOTIDE_EXCISION_REPAIR, MAPK_SIGNALING_PATHWAY, MISMATCH_REPAIR, and

APOPTOSIS. Interestingly many of the same pathways uncovered with the CNV data were reported. At the top of this list were cell adhesion molecules which are glycoproteins expressed on the cell surface. In the context of cancer, they play integral roles in tumor invasiveness and metastasis. Recently an integrated methylomics and genomics analyses uncovered an epigenetic signature in cell adhesion molecules which may affect the therapeutic outcome survival in ovarian cancer.[41] At a gene level aberrant expression of *HLA.DRB1* leads to aberrant human leukocyte antigen (HLA) phenotypes associated with ovarian cancer.[42] *ORC1* encodes for a DNA replication initiator involved in chromatin organization, centromere function, and cytokinesis,[43-46] making this gene a key cell cycle regulator.[47,48] *ORC1* inhibition reduces cellular invasion and migration[49] making mutations in this gene or its promotors which lead to increased expression potently oncogenic.[50,51] *CUL4A* encodes a scaffold protein key to the assembly of ligase complexes which degrade cellular proteins.[52] Increased *CUL4A* expression has been associated with the accelerated neoplastic formation in ovarian cancer.[53] *CUL4A* overexpression has also been noted in ovarian cancer, with *CUL4A* knockdown reducing EMT and cell proliferation.[54] *PCNA* is involved in DNA synthesis/repair, and regulation of the cell cycle.[55] Increased expression of *PCNA* with increased WNT and Hedgehog pathway activity leads to tumor cell proliferation.[56] *NFKB1* is a transcription factor involved in numerous core cellular processes, including inflammation, immunity, and apoptosis. Given these cellular roles, it is unsurprising that *NFKB1* has been implicated as both a tumor suppressor and potent oncogene, with altered *NFKB1* activity being linked to cancer through numerous mechanisms, often by augmenting the expression of other genes to reduce apoptosis and increase tumor angiogenesis, inflammation, and EMT.[57-60]

5 | DISCUSSION

In this article, we present Bayes-InGRiD, a Bayesian sparse latent factor model for the simultaneous identification of cancer patient subtypes and key molecular features within a unified framework, based on the integrative analysis of continuous, binary, and count data. Bayes-InGRiD does not only improve the accuracy of patient subgroup and key molecular feature identification, but also improves biological interpretation by using pathway information. The results from the simulation studies revealed the superiority of the pathway-level analyses over gene-level analyses in identifying key molecular features for both separate data analysis and integrative data analysis, especially when the signal-to-noise ratio is low. Additionally, we observed higher sensitivity and specificity in integrated data analysis compared to separate data analysis, demonstrating the benefit of added information through joint data analysis. Bayes-InGRiD outperforms the gene-level approach and it provides a means for us to incorporate additional pathway information into the inference of gene and pathway association. Some limitations of our work must be acknowledged. First, we modified the gene set association of shared genes to solve the pathway overlapping issue. This is a preprocessing step before we apply the latent factor model and could lead to a loss of information where we use pathway information to guide factor definition. Alternatively, one could incorporate overlapping gene information in the latent factor model setting, and we also plan to investigate this direction in future work. Second, we intend to extend the proposed model to allow for the imputation of missing at random observations when the number of individuals is different across different datasets. In addition, we plan to extend our model to include ordinal data type.= In summary, Bayes-InGRiD can be a powerful approach for investigating cancer patient subgroups and their molecular features.

CONFLICT OF INTEREST STATEMENT

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

DATA AVAILABILITY STATEMENT

The R package "INGRID" implementing the proposed approach is currently available in our research group GitHub webpage (https://dongjunchung.github.io/INGRID/). The data sets in the real data analysis section are also available in the R package.

ORCID

Zequn Sun https://orcid.org/0000-0001-9301-6295
Dongjun Chung https://orcid.org/0000-0002-8072-5671
Brian Neelon https://orcid.org/0000-0002-8929-6033
Gary Hardiman https://orcid.org/0000-0003-4558-0400

REFERENCES

1. The Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature*. 2011;474(7353):609-615. http://www.nature.com/articles/nature10166
2. Karl Pearson FRS. LIII. On lines and planes of closest fit to systems of points in space. *Lond Edinb Dublin Philos Mag J Sci*. 1901;2(11):559-572.
3. Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*. 2009;25(22):2906-2912. doi:10.1093/bioinformatics/btp543
4. Yang Z, Michailidis G. A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. *Bioinformatics*. 2015;32:1-8. doi:10.1093/bioinformatics/btv544
5. Lock EF, Hoadley KA, Marron JS, Nobel AB. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *Ann Appl Stat*. 2013;7(1):523-542. http://projecteuclid.org/euclid.aoas/1365527209
6. Shen R, Wang S, Mo Q. Sparse integrative clustering of multiple omics data sets. *Ann Appl Stat*. 2013;7(1):269-294. http://projecteuclid.org/euclid.aoas/1365527199
7. Mo Q, Wang S, Seshan VE, et al. Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proc Natl Acad Sci*. 2013;110(11):4245-4250. doi:10.1073/pnas.1208949110
8. Mo Q, Shen R, Guo C, Vannucci M, Chan KS, Hilsenbeck SG. A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data. *Biostatistics*. 2017;19(1):71-86. doi:10.1093/biostatistics/kxx017
9. Webber JT, Kaushik S, Bandyopadhyay S. Integration of tumor genomic data with cell lines using multi-dimensional network modules improves cancer pharmacogenomics. *Cell Syst*. 2018;7(5):526-536.e6. https://linkinghub.elsevier.com/retrieve/pii/S2405471218303910
10. Ma H, Zhao H. iFad: an integrative factor analysis model for drug-pathway association inference. *Bioinformatics*. 2012;28(14):1911-1918. doi:10.1093/bioinformatics/bts285
11. Ma H, Zhao H. FacPad: Bayesian sparse factor modeling for the inference of pathways responsive to drug treatment. *Bioinformatics*. 2012;28(20):2662-2670. doi:10.1093/bioinformatics/bts502
12. Wei W, Sun Z, da Silveira WA, et al. Semi-supervised identification of cancer subgroups using survival outcomes and overlapping grouping information. *Stat Methods Med Res*. 2019;28(7):2137-2149. doi:10.1177/0962280217752980
13. Chung D, Keles S. Sparse partial least squares classification for high dimensional data. *Stat Appl Genet Mol Biol*. 2010;9(1):17. https://www.degruyter.com/view/j/sagmb.2010.9.1/sagmb.2010.9.1.1492/sagmb.2010.9.1.1492.xml
14. Chib S, Greenberg E. Understanding the Metropolis-Hastings algorithm. *Am Stat*. 1995;49(4):327-335. doi:10.1080/00031305.1995.10476177
15. Pillow JW, Scott J. Fully Bayesian inference for neural models with negative-binomial spiking. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ, eds. *Advances in Neural Information Processing Systems 25*. Red Hook, NY: Curran Associates; 2012:1898-1906. http://papers.nips.cc/paper/4567-fully-bayesian-inference-for-neural-models-with-negative-binomial-spiking.pdf
16. Polson NG, Scott JG, Windle J. Bayesian inference for logistic models using Pólya–Gamma latent variables. *J Am Stat Assoc*. 2013;108(504):1339-1349. doi:10.1080/01621459.2013.829001
17. George EI, McCulloch RE. Approaches for Bayesian variable selection. *Stat Sin*. 1997;7(2):339-373. http://www.jstor.org/stable/24306083
18. Zhou M, Carin L. Negative binomial process count and mixture modeling. *IEEE Trans Pattern Anal Mach Intell*. 2015;37(2):307-320. http://ieeexplore.ieee.org/document/6636308/
19. Zamani Dadaneh S, Zhou M, Qian X. Covariate-dependent negative binomial factor analysis of RNA sequencing data. *Bioinformatics*. 2018;34(13):i61-i69. https://academic.oup.com/bioinformatics/article/34/13/i61/5045747
20. Lopes H, West M. Bayesian model assessment in factor analysis. *Stat Sin*. 2004;01(14):41-67.
21. Gao C, McDowell IC, Zhao S, Brown CD, Engelhardt BE. Context specific and differential gene co-expression networks via Bayesian biclustering. *PLoS Comput Biol*. 2016;12(7):e1004791. doi:10.1371/journal.pcbi.1004791
22. Argelaguet R, Arnol D, Bredikhin D, et al. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol*. 2020;21(1):111. doi:10.1186/s13059-020-02015-1
23. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*. 2011;27(12):1739-1740. doi:10.1093/bioinformatics/btr260
24. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci*. 2005;102(43):15545-15550. https://www.pnas.org/content/102/43/15545
25. Jones S, Zhang X, Parsons DW, et al. Core Signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science*. 2008;321(5897):1801-1806. doi:10.1126/science.1164368
26. Winterhoff BJ, Maile M, Mitra AK, et al. Single cell sequencing reveals heterogeneity within ovarian cancer epithelium and cancer associated stromal cells. *Gynecol Oncol*. 2017;144(3):598-606. https://linkinghub.elsevier.com/retrieve/pii/S0090825817300598
27. Scarbrough P, Weber R, Iversen E, et al. A cross-cancer genetic association analysis of the DNA repair and DNA damage signaling pathways for lung, ovary, prostate, breast, and colorectal cancer. *Cancer Epidemiol Biomarkers Prev*. 2015;25:193-200.
28. Reynolds AP, Richards G, de la Iglesia B, Rayward-Smith VJ. Clustering rules: a comparison of partitioning and hierarchical clustering algorithms. *J Math Model Algorithms*. 2006;5(4):475-504. doi:10.1007/s10852-005-9022-1
29. Thorsen J, Brejnrod A, Mortensen M, et al. Large-scale benchmarking reveals false discoveries and count transformation sensitivity in 16S rRNA gene amplicon data analysis methods used in microbiome studies. *Microbiome*. 2016;4(1):62. doi:10.1186/s40168-016-0208-8

30. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2018;68(6):394-424.
31. Matulonis UA, Sood AK, Fallowfield L, Howitt BE, Sehouli J, Karlan BY. Ovarian cancer. *Nat Rev Dis Primers*. 2016;2:16061.
32. Pećina-Šlaus N, Kafka A, Salamon I, Bukovac A. Mismatch repair pathway, genome stability and cancer. *Front Mol Biosci*. 2020;7:122.
33. Masoumi-Dehghi S, Babashah S, Sadeghizadeh M. microRNA-141-3p-containing small extracellular vesicles derived from epithelial ovarian cancer cells promote endothelial cell angiogenesis through activating the JAK/STAT3 and NF-B signaling pathways. *J Cell Commun Signal*. 2020;14(2):233-244.
34. Yokoyama T, Kohn EC, Brill E, Lee JM. Apoptosis is augmented in high-grade serous ovarian cancer by the combined inhibition of Bcl-2/Bcl-xL and PARP. *Int J Oncol*. 2017;50(4):1064-1074.
35. Rinne N, Christie EL, Ardasheva A, et al. Targeting the PI3K/AKT/mTOR pathway in epithelial ovarian cancer, therapeutic treatment options for platinum-resistant ovarian cancer. *Cancer Drug Resist*. 2021;4(3):573-595.
36. Edelbrock MA, Kaliyaperumal S, Williams KJ. Structural, molecular and cellular functions of MSH2 and MSH6 during DNA mismatch repair, damage signaling and other noncanonical activities. *Mutat Res*. 2013;743-744:53-66.
37. Cederquist K, Emanuelsson M, Wiklund F, Golovleva I, Palmqvist R, Grönberg H. Two Swedish founder MSH6 mutations, one nonsense and one missense, conferring high cumulative risk of lynch syndrome. *Clin Genet*. 2005;68(6):533-541.
38. Razaghi A, Brusselaers N, Björnstedt M, Durand-Dubief M. Copy number alteration of the interferon gene cluster in cancer: individual patient data meta-analysis prospects to personalized immunotherapy. *Neoplasia*. 2021;23(10):1059-1068.
39. Im-aram A, Farrand L, Bae SM, et al. The mTORC2 component rictor contributes to cisplatin resistance in human ovarian cancer cells. *PLoS One*. 2013;8(9):e75455.
40. Yung MMH, Ngan HYS, Chan DW. Targeting AMPK signaling in combating ovarian cancers: opportunities and challenges. *Acta Biochim Biophys Sin*. 2016;48(4):301-317.
41. Chang PY, Liao YP, Wang HC, et al. An epigenetic signature of adhesion molecules predicts poor prognosis of ovarian cancer patients. *Oncotarget*. 2017;8(32):53432-53449.
42. Kübler K, Arndt PF, Wardelmann E, et al. Genetic alterations of HLA-class II in ovarian cancer. *Int J Cancer*. 2008;123(6):1350-1356.
43. Prasanth SG, Shen Z, Prasanth KV, Stillman B. Human origin recognition complex is essential for HP1 binding to chromatin and heterochromatin organization. *Proc Natl Acad Sci USA*. 2010;107(34):15093-15098.
44. Verdaasdonk JS, Bloom K. Centromeres: unique chromatin structures that drive chromosome segregation. *Nat Rev Mol Cell Biol*. 2011;12(5):320-332.
45. Ruiz-Velasco M, Zaugg JB. Structure meets function: how chromatin organisation conveys functionality. *Curr Opin Syst Biol*. 2017;1:129-136. https://linkinghub.elsevier.com/retrieve/pii/S2452310017300173
46. Green RA, Paluch E, Oegema K. Cytokinesis in animal cells. *Annu Rev Cell Dev Biol*. 2012;28(1):29-58. doi:10.1146/annurev-cellbio-101011-155718
47. Faustova I, Loog M. SLiMs in intrinsically disordered protein regions regulate the cell cycle dynamics of ORC1-CDC6 interaction and pre-replicative complex assembly. *Mol Cell*. 2021;81(9):1861-1862.
48. Li S, Zhu A, Ren K, Li S, Chen L. DEFA1B inhibits ZIKV replication and retards cell cycle progression through interaction with ORC1. *Life Sci*. 2020;263:118564.
49. Xiong W, Xie C, Qiu Y, Tu Z, Gong Q. Origin recognition complex subunit 1 regulates cell growth and metastasis in glioma by altering activation of ERK and JNK signaling pathway. *Mol Cell Probes*. 2020;49:101496.
50. Van Zijl F, Krupitza G, Mikulits W. Initial steps of metastasis: cell invasion and endothelial transmigration. *Mutat Res*. 2011;728(1-2):23-34.
51. Bravo-Cordero JJ, Hodgson L, Condeelis J. Directed cell invasion and migration during metastasis. *Curr Opin Cell Biol*. 2012;24(2):277-283.
52. Sharma P, Nag A. CUL4A ubiquitin ligase: a promising drug target for cancer and other human diseases. *Open Biol*. 2014;4:130217.
53. Birner P, Schoppmann A, Schindl M, et al. Human homologue for Caenorhabditis elegans CUL-4 protein overexpression is associated with malignant potential of epithelial ovarian tumours and poor outcome in carcinoma. *J Clin Pathol*. 2012;65(6):507-511.
54. Han X, Fang Z, Wang H, Jiao R, Zhou J, Fang N. CUL4A functions as an oncogene in ovarian cancer and is directly regulated by miR-494. *Biochem Biophys Res Commun*. 2016;480(4):675-681.
55. Strzalka W, Ziemienowicz A. Proliferating cell nuclear antigen (PCNA): a key factor in DNA replication and cell cycle regulation. *Ann Bot*. 2011;107(7):1127-1140.
56. Tripathy A, Thakurela S, Sahu MK, et al. The molecular connection of histopathological heterogeneity in hepatocellular carcinoma: a role of Wnt and hedgehog signaling pathways. *PLoS One*. 2018;13(12):e0208194.
57. Kim GC, Kwon HK, Lee CG, et al. Upregulation of Ets1 expression by NFATc2 and NFKB1/RELA promotes breast cancer cell invasiveness. *Oncogenesis*. 2018;7(11):91.
58. Chen Y, Lu R, Zheng H, et al. The NFKB1 polymorphism (rs4648068) is associated with the cell proliferation and motility in gastric cancer. *BMC Gastroenterol*. 2015;15:21.
59. Concetti J, Wilson CL. NFKB1 and cancer: friend or foe? *Cell*. 2018;7(9):E133.

60. Wu J, Wood GS. Reduction of Fas/CD95 promoter methylation, upregulation of Fas protein, and enhancement of sensitivity to apoptosis in cutaneous T-cell lymphoma. *Arch Dermatol*. 2011;147(4):443-449.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Sun Z, Chung D, Neelon B, et al. A Bayesian framework for pathway-guided identification of cancer subgroups by integrating multiple types of genomic data. *Statistics in Medicine*. 2023;42(28):5266-5284. doi: 10.1002/sim.9911