2021

# Statistical Methods for Integrative Analysis, Subgroup Identification, and Variable Selection Using Cancer Genomic Data

Zequn Sun
*Medical University of South Carolina*

Statistical Methods for Integrative Analysis, Subgroup Identification, and Variable Selection Using Cancer Genomic Data

Zequn Sun

A dissertation submitted to the faculty of the Medical University of South Carolina
in partial fulfillment of the requirements for the degree of Doctor of Philosophy
in the College of Graduate Studies.

Department of Public Health Sciences

2021

Approved by:

Dr. Dongjun Chung, Co-Chair

Dr. Brian Neelon, Co-Chair

Dr. Stephen P. Ethier

Dr. Kristin Wallace

Dr. Feifei Xiao

**TABLE OF CONTENTS**

**ACKNOWLEDGMENTS**

To begin, I would like to thank my advisor Dr. Dongjun Chung of the Department of Biomedical Informatics at The Ohio State University College of Medicine. Dr.Chung was always willing to help me and patiently explained and re-explained biological and statistical concepts. He consistently allowed this paper to be my own work, while providing invaluable guidance and corrections.

I would also like to thank my committee members who rounded out my understanding of this field and who lent insightful suggestions concerning my research specifically: Dr. Brian Neelon, Dr. Stephen P. Ethier, Dr. Kristin Wallace, Dr. Feifei Xiao. I would also like to acknowledge the wonderful June Watson and Paula Talbot. I would not have been able to navigate the paper work and requirements of the department without their help and counsel.

Finally, I would like to express gratitude to my wife and parents. Together, they make an unparalleled support team, without whose love I would not have been able to arrive at this point with such grace.

Thank you!

Zequn Sun

# 1. INTRODUCTION

## 1.1 Overview

Cancer is one of the leading causes of death. In 2018 alone, there was an estimated 1,735,350 new diagnoses and 609,640 cancer-related deaths in the United States [1]. Much work is ongoing to better understand and treat this group of diseases. However, cancer is an extremely complex disease. There are over 100 types of cancers, located in different organs and subtissues and emerging from different cell types [1]. Despite this complexity and variability, more and more targeted drugs and therapies are developed to treat the various types of cancer. The Cancer Genome Atlas (TCGA) is a landmark cancer genomics platform that provides high throughput genomic data for over 20,000 matched cancers and normal samples including 33 different types of cancer [2]. The joint effort between the National Cancer Institute and the National Human Genome Research Institute offers researchers high quality of TCGA molecular data which helps us better understand cancer biology [3]. TCGA have generated tremendous amount of high throughput genomic datasets including somatic mutation, DNA methylation, gene expression, and DNA copy number alterations (CNA) for each patient. This large-scale cancer genomic data provides unprecedented opportunity to investigate cancer subgroups using integrative approaches based on multiple types of genomic data [4]. With increasing availability of high throughput genomic data, a statistical and computational tool for identification of cancer patient subgroups is of critical importance to clinicians and researchers. However, the development of novel integrative methods that aim to integrate different types of data is non-trivial [5] due to (i) the challenge to understand shared and data-specific variations; and (ii) the challenge to integrate different types of data including continuous, binary, count data; and (iii) the challenge to facilitate biological understanding of novel findings, respectively.

To address these challenges, in Aim 1 we propose a novel latent factor model called "Bayes-InGRiD" for the simultaneous identification of cancer subgroups and key molecular features

within a unified framework, based on the joint analysis of continuous, binary and count data. More over, we plan to use Polya-Gamma mixtures of normal for binary and count data to promote an exact and fully automatic posterior sampling. Last but not the least, pathway information will be utilized to improve accuracy of cancer subgroup and key molecular feature identification, and facilitates biological understanding of novel findings.

Moreover, our goal is to develop a comprehensive and interactive software implementing the method developed in Aim 1. In order to improve the computational efficiency and robustness of the Bayes-InGRiD model, we plan to develop an user-friendly function called "BayesInGRiD" and provide it as a part of the R package "InGRiD" [6].

On the other hand, there are growing interests in developing immunotherapies to fight against various types of cancer. With increasing availability of immune cellular fraction data in the compositional data form, it is of our interest to apply appropraite variable selection in compositional data setting and infer key immune subtypes associated. When it comes to Compositional Data Analysis (CoDA), the most common compositional replacement is to covert the data to ratios using the centred log-ratio (clr) [7] transformation. However, the variable selection applied on the clr-transformed variables makes interpretation challenging. Since our goal is to identify key cell types, it is crucial to address the issue of interpretability for variable selection. Greenacre [8] purposed an stepwise log-ratio procedure, where a smaller set of ratios can be chosen to explain as much variability as required to reveal the underlying structure of the data. For the purpose of identifying key immune cell subtypes, we implement the stepwise pairwise log-ratio procedure using cellular fractions data induced from Colorectal Adenocarcinoma TCGA PanCancer study processed by CIBERSORT [9].

While the pairwise log-ratio stepwise approach is efficient variable selection approach for low-dimensional compositional data. it is not applicable for high-dimensional zero-inflated microbiome datasets generated by high throughput sequencing (HTS) technology. With high-dimensional and zero-inflated nature of microbiome data, we purpose an alternative approach by considering the microbiome data as univariate and apply zero-inflated Wilcoxon test [10] for variable selection.

## 1.2 Gaps in the Current Literature

The increasing availability of large heterogeneous data sets (e.g. TCGA) has prompted the development of novel integrative methods. [5]. A popular latent factor model approach in cancer genomic field is iCluster [11]. iCluster overcomes many of challenges by simultaneously identifying cancer subgroups and important genes by integrating multiple continuous data within a unified framework. The model is built on a latent variable model that captures correlations among variables through latent factors. The limitation is that it only allows continuous data.

An extension of the iCluster model, icluster+ allows to account for binary, counts and categorical data [12]. The upgraded icluster+ expands iCluster by using different modeling approaches for different types of data while sharing a common latent factor matrix across different data platforms [13]. A limitation of the icluster+ is that statistical inference is not straightforward due to the computational complexity.

To address the challenge of the icluters+, Mo and others developed a fully Bayesian method for integrative clustering analysis of multitype omics data [14]. This new method significantly improves the icluster+ 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 icluster+ method. It provides researchers a powerful tool for integration of multiple types of data and identification of key cancer subtypes and potential therapeutic targets. In spite of this, the following challenges still remain unsolved. First, gene-level analysis doesn't promote biological understanding and additional gene set enrichment analyses are needed to understand these genes in the context of biological networks. Second, as closed form is missing for some parameter updates, Metropolis-Hasting algorithm needs to be employed and this requires tuning of proposal distributions, which is not always straightforward.

Several other matrix factorization approaches like JIVE [15] and iNMF [16] are purposed in recent years. JIVE and iNMF extends iCluster by adding a data-specific term. This improvement promotes biological understanding of individual structures and help us study how can data-specific variations impact the estimation of the shared structure in partial least squares models

[17]. However, unlike iCluster that provide tools to cluster samples from the latent variables, JIVE and iNMF do not give guideline to generate a final sample clustering. Also, JIVE and iNMF are limited as gene-level approach without pathway information in the model.

Another well known multiple data set integration approach is Bayesian consensus clustering (BCC) [18]. Bayesian consensus clustering is a Bayesian method that represents each data set with a Dirichlet-multinomial model [19] and it uncovers a single common clustering across sources. Still, several key limitation remains to be unsolved for BCC: (1)it allows only continuous data; (2) it doesn't support embedding of pathway information in the model; (3) there is no identification of key genes and pathways for BCC.

As discussed above, most integrative approach do not incorporate pathway-level information, ignoring the fact that Pathway-level can play an importatant role facilitate biological findings. Popular approaches such as iFad and PacFad used Bayesian sparse factor analysis models to jointly analyze the paired gene expression and drug sensitivity datasets [20] [21]. The key innovation of these models is that they represent biological pathways as latent factors. Still, the limitation is excessive number of latent components in the model. In addition, how to speed up the computation process remains to be a challenge since MCMC methods are usually time-consuming when applied to high-dimensional inference. Also, the model allows only continuous data.

To address the problem of low reproducibility and instability of identified cancer subgroups and molecular features for gene-level approaches, Wei and others developed InGRiD (Integrative Genomics Robust iDentification of cancer subgroups) [6]. InGRiD is a multi-step approach where a Sparse Partial Least Squares (SPLS) Cox regression model is used for gene-level analysis and a LASSO-penalized Cox regression model is built on SPLS latent components to select a parsimonious set of pathways. However, this approach only considered the gene expression data. In order to capture more complete picture of cancer landscape and detect signals that are missing in gene expression data, it will be critical to utilize and integrate other genomic data types, such as CNA, DNA methylation, and somatic mutations for predicting patient risk.

With the rapid development of immune therapy for cancer, we are interested in analyz-

ing compositional data of the immune cell composition in clinical tumour samples. In recent years, many appoaches have been purposed to estimate the fraction of immune cells in clinical tumour samples. One of the widely used approaches is CIBERSORT (Robust enumeration of cell subsets from tissue expression profiles), which is developed by Newman and others to estimate cell composition of tumour cells from their gene expression profiles [9]. CIBERSORT implements support vector regression (SVR) to improve deconvolution performance through a combination of variable selection and other optimization techniques. CIBERSORT is validated by many researchers that it is a useful approach for high throughput characterization of cell types [22]. Still, deconvolution is often sensitive to noise, the robustness of the method could be further improved. [23].

When it comes to Compositional Data Analysis (CoDA), the most common compositional replacement is to covert the data to ratios using the centred log-ratio (clr) [7] transformation. A drawback of this method is that the variable selection applied on the clr-transformed variables makes interpretation challenging. Since our goal is to identify key cell types, it is crucial to address the issue of interpretability for variable selection. Hron and others [24] purposed a covariance-based stepwise procedure for variable selection in 2013. In this procedure, variable selection is achieved by eliminating the variable whose variance of the corresponding clr variable is the smallest, calculating normalized variance of transformed variables of the remaining sample space, and repeating the procedure until a purposed test statistics reach a pre-specified threshold. Another variable selection approach is proposed by Greenacre [8] where all pairwise ratios of parts are considered for key marker identification. A smaller set of ratios can be chosen to explain as much variability as required to reveal the underlying structure of the data. For the purpose of identifying key immune cell subtypes.

Although the covariance-based stepwise procedure and pairwise log-ratio stepwise approach are efficient variable selection approach for low-dimensional compositional data. these two approaches are no longer applicable when it comes to high-dimensional zero-inflated microbiome datasets generated by HTS technology. Microbiome datasets generated by HTS are compositional. To deal with the large proportion of zeros in the microbiome data, many imputation

approaches have emerged in recent years. The R package zCompositions [25] provides several methods for the multivariate imputation of zeros and non-detects in compositional data. These approaches are proposed based on an appropriate coordinate representation of the compositional data in the usual Euclidean geometry. The imputation is achieved by using iterative approaches where EM algorithm [26], Markov Chain Monte Carlo (MCMC) [27] or multiple imputation are utilized. However, in some extreme cases, we could face microbiome data where the majority of the data are zeros and the number of variables could be hundreds. The imputation approaches are not applicable given overwhelming amount of zeros in the data. In addition, it is important to realize many assumption of multivariate approach that was developed based on compositional data setting are not fit given the high dimensionality.

### 1.3 Overall Goal and Specific Aims

During the last decade, the increasing availability of large heterogeneous data sets has prompted the development of novel integrative methods that aim to capture weak yet consistent patterns across data types [5]. This task is, however, non-trivial due to (i) the challenge to decipher data-specific from inter-source variations [11]; and (ii) the different types of noise and confounding effects across platforms, resulting in data heterogeneity. [12]; and (iii) the identification of key genes does not directly promote understanding of biological networks and additional downstream analyses are needed to characterize these genes in the context of biological networks. [14]. It has been shown that pathway (gene set) information may not only improve accuracy and robustness of cancer subgroup and key molecular feature identification, but also facilitates biological understanding of novel findings [6]. While various algorithms, such as iClusterBayes [14], have been developed to identify cancer subtypes using information from multiple genomic platforms, it still remains a challenging task to simultaneous identify of cancer subgroups and key molecular features while facilitating biological understanding of novel findings using pathway information within a unified framework.

Moreover, with the rapid development of immune therapy for cancer related disease, many

researchers are interested in analyzing compositional data of the immune cell composition in clinical tumour samples. Given the increasing availability of immune cellular fraction data in the compositional data form, it is of great importance to identify and interpret key immune subtypes associated in the tumor immune microenvironment. When it comes to Compositional Data Analysis (CoDA), the most common compositional replacement is to covert the data to ratios using the centred log-ratio (clr) [7] transformation. However, the variable selection applied on the clr-transformed variables makes interpretation challenging, one must apply inverse transformation for interpretation. Since our goal is to identify key cell types, it is crucial to address the issue of interpretability for variable selection.

Significant advancements have been made in the development and use of targeted drugs for many types of cancer. However, there are few approaches to evaluate the similarities and differences that exist between genomic features in cell lines and patient samples [28]. Because of the complex and multi-factorial nature of the disease, such a comparison must consider the complement of somatic mutations, copy number, gene expression, methylation, and proteomic changes found in tumors as well as the molecular features whose variation across patients is inextricably linked, necessitating a modular analysis [29]. In order to address this, it is of need to integrate multiple types and molecular features and implement system-level analysis for the cell line data.

In addition, microbiome datasets generated by HTS are compositional because the total number of sequenced reads depends on the capacity of the instrument. With the high-dimensional and zero-inflated nature of microbiome data, much more care needs to be devoted to a reasonable coordinate representation and selection of methods to be used in the compositional data setting. To deal with the issue of zero-inflation in the microbiome data, imputation approach such as Model-based ordinary and robust Expectation-Maximisation algorithms (lrEM) [26] has emerged in recent years. However, most imputation approaches are not applicable given overwhelming amount of zeros in some extreme cases. In addition, it is important to realize many assumption of multivariate approach that was developed based on compositional data setting are not fit given the high dimensionality.

**Aim 1: Develop a Bayesian latent factor model for pathway-guided identification of cancer subgroups by integrating multiple types of genomic data.** This statistical method will provide researchers a unified framework to simultaneously identify cancer subgroups (clustering) and key molecular markers (variable selection) based on the joint analysis of continuous, binary and count data. In addition, we plan to use Polya-Gamma mixtures of normal for binary and count data to promote an exact and fully automatic posterior sampling. Moreover, pathway information will be used to improves accuracy and robustness of cancer subgroup and key molecular features identification.

**Aim 2: Develop a comprehensive software implementing the method developed in Aim 1, and apply it to simultaneously identify sample clustering and key features in cancer genomic study.** We aim to develop an user-friendly function called "BayesInGRiD" and provide it as a part of the R package "InGRiD"

**Aim 3: Variable selection in compositional data analysis with application in immunology data and microbiome data.** We aim to to investigate various variable selection approaches in compositional data setting, infer key immune subtypes associated by applying stepwise pairwise log-ratio procedure on immune cellular fractions data, and identify key species in the microbiome data by using zero-inflated Wilcoxon rank sum test for Colorectal Adenocarcinoma.

The proposed methods will be developed and evaluated on three types of cancer genomics datasets: 1) large scale public datasets from the TCGA database, including gene expression, DNA copy number alteration, somatic mutation data; 2) cellular fractions data induced from Colorectal Adenocarcinoma TCGA Pan- Cancer study; 3) high dimensional zero-inflated microbiome data from studies of colorectal cancer.

## 2. STATISTICAL BACKGROUND

### 2.1 Overview of the statistical background

A few strategies have been proposed to integrate heterogeneous omics data to uncover coordinated cellular processes acting across different omic layers. The first group of methods is the iCluster series approaches, including Integrative clustering of multiple genomic data [11], Integrative clustering of multi-type genomic data [12] and a fully Bayesian latent variable model for integrative clustering [14], where a latent factor framework were purposed to integrate multiple datasets. The second category includes introduce Joint and individual variation explained (JIVE) [15] and non-negative matrix factorization (iNMF) [16] for investigating common variations across omics using matrix factorization. Finally, we will review Bayesian and network-based approach specifically tailored for data integration such as Bayesian concensus clustering (BCC) [18].

It is key to realize that existing integrative approaches only focus on gene-level analysis and lack the ability to facilitate biological findings in pathway-level. Some statistical approach have been purposed to improve the robustness in biological findings and the interpretation of subgroups of cancer patients, including integrative factor analysis model (iFad) [20] and Bayesian sparse factor modeling for the inference of pathways responsive to drug treatment (FacPad) [21], where biological pathways are represent as latent factors. To address the challenge of low reproducibility and instability of identified cancer subgroups and molecular features, Integrative Genomics Robust iDentification of cancer subgroups (InGRiD) [6] simultaneously represents the gene expression profiles of pathways and selects key genes from each pathway by constructing pathway-level latent components based on sparse partial least squares (SPLS) Cox regression.

In order to extend the latent factor model framework to the compositional data setting, we look into a well-known deconvolution method, CIBERSORT (Robust enumeration of cell sub-

sets from tissue expression profiles) [9], that describes the gene expression of a sample as the weighted sum of the expression profiles of the cell types. Since we are interested in identifying patient clusters and the key cell type associate, we also introduce a covariance-based variable selection for compositional data [24] that performs covariance-based stepwise procedure to omit variables.

### 2.2 Integrative clustering of multiple genomic data types using a joint latent variable model (iCluster)

iCluster is a Gaussian joint latent variable model that perform clustering on single shared latent factor matrix across $T$ data sets $Y_t$ of dimensions $p_t$ $(t = 1, \cdots, T)$ by $N$ measured on the same $N$ samples [11]. The model is based on an latent variable model that captures correlations among variables through latent factors where the latent variable matrix is shared across $T$ data sets. For $t$th data, the model can be written as:

$$Y_t = \beta_t Z + \epsilon_t$$

$$Z \sim N_q(0, I)$$

where $\beta_t$ is the $p_t$ by $q$ factor loading matrix of data set $t$, $Z$ is the $q$ by $N$ common latent variable matrix, and $\epsilon_t$ is the $p_t$ by $N$ error matrix that has a multivariate Gaussian distribution $N_{p_t}(0, \phi_t)$ with zero mean and diagonal covariance matrix $\phi_t = diag(\sigma_{t_{2,1}}, \cdots, \sigma_{t_{2,p_t}})$. The key is to assume latent variables are shared among all $T$ data sets. An Expectation-Maximization algorithm [30] is performed for parameter estimation on the multivariate normal distribution. Then $k$-means clustering is applied to the posterior expectation of the latent factors $E(Z|Y)$ for patient subgroup clustering. An $l_1$ penalty is imposed on the loading coefficients to perform variable selection. The Least absolute shrinkage and selection operator (LASSO) penalty [30] results in sparse estimates of $\beta$ in which many of the coefficients are shrunken toward zero while some coefficients to exactly zero. As a result, key genes can be identified using loading

matrix and subtype can be clustered on the latent factor $Z$.

The contribution of iCluster is that it serves as a landmark where it integrates multiple continuous data, while simultaneously reducing the dimensionality of the datasets and identifying patient clusters. Still, it is in need to develop approaches that can integrate different types of data including binary, counts and categorical data.

### 2.3 Integrative clustering of multi-type genomic data (iCluster+)

Shen and others extended the iCluster model by creating icluster+, where icluster+ allows to account for binary, counts and categorical data [12]. Specifically, icluster+ expands iCluster by making the assumption of different modeling approaches for different data platforms. iCluster+ fits a latent variable model that integrates diverse data types including binary, continuous, categorical and count data with different modeling assumptions including logistic, normal linear, multilogit, and Poisson distributions [12].

Building on original iCluster model, now if $Y_t$ is binary data, it is modeled with the classical logistic regression; if $Y_t$ is count data, Poisson regression is considered for the model; If $Y_t$ is a multicategory variable, the model would be constructed based on multilogit regression. As for the common latent variable vector $Z$, it still represents the underlying driving factors that can be used for disease subtype clustering. LASSO penalty is introduced to address the sparsity issue in $\beta_{jt}$ [31]. The sparsity-inducing parameter $\lambda_t$ is allowed to take different values for different data types.

The Strength of the iCluster+ is that it promotes simultaneously identification of cancer subgroups and key genes from multiple genomic platforms. And it boosts a joint analysis of binary, counts and categorical data within a unified statistical model. However, iCluster+ has the limitation of challenging parameter tuning.

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

To address the challenges in iCluster+, Mo et al. developed a fully Bayesian latent variable method (called iClusterBayes) that can jointly models omics data of continuous and discrete data types for identification of tumor subtypes and relevant omics features [14]. Specifically, it uses a Bayesian latent factor model to integrate multiple omics data sets and achieve joint dimension reduction. As a result, the tumor samples can be clustered in the latent variable space and key molecular features that drive the sample clustering are identified through Bayesian variable selection.

The Bayesian latent factor model allows one to make assumptions on multiple types of data sets with distinct distributions, as well as on the correlations among data sets. More over, it avoids complicated parameter tuning required when a penalization approach is used. In contrast, the icluster+ 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. And the Bayesian model provides an posterior estimation for each omics feature, which can be used as an uncertainty measurement for feature selection and patient subgroup prediction.

A few limitations still remain for iClusterBayes. First, it is a gene-level approach, and we need a model that embeds pathway information in the model. Second, Metropolis-Hasting approach is used for binary and count data analysis where parameter tuning could make inference less straightforward.

## 2.5 Joint and individual variation explained for integrated analysis of multiple data types (JIVE)

JIVE extends iCluster by adding a data-specific term [15]. Motivated by the biological interest of studying individual structures and also by observing that data-specific variations, the addition of the data-specific term plays an important role in the estimation of the shared structure in partial least squares models. The model can be written as follow:

$$Y_t = \beta_t Z + \beta_t^s Z_t^s + \epsilon_t$$

where $\beta_t^s$ of size $p_t$ by $q_t$ and $Z_t^s$ of size $q_t$ by $N$ are the data-specific loading and latent variable matrices, respectively. Note that $q$ and $q_t$ are not necessarily equal, implying that the joint and individual approximations may be of different dimensions. To solve the identifiability issue of the decomposition, the authors imposed an orthogonality constraint between the joint and individual terms. The parameter estimation is performed by estimating the joint and individual structures via singular value decomposition (SVD). Sparsity is induced during the estimation procedure by an $L1$ penalty on the loading matrices. The level of sparsity is determined using the Bayesian information criterion.

One should notice that JIVE allows only continuous data, and no pathway or gene set information is incorporated in the model. Unlike iCluster series approach that provide tools to cluster samples from the latent variables, no guideline is given to generate a final sample clustering.

### 2.6 A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data (iNMF)

Similar to JIVE we introduced above, iNMF [16] also aims to capture the shared and data specific structures. However, there are two notable differences for iNMF. First, the latent variables are estimated using a non-negativity constraint instead of orthogonality. Second, a common coefficient matrix $\beta_t$ is shared between the data-specific $Z_T^s$ and the common $Z$ basis matrices where the coefficient and basis matrices are the counterparts of the loading and latent variable matrices. iNMF optimizes the following problem with a Euclidean loss function:

$$\min_{\substack{Z, Z_1^s, \cdots Z_T^s \\ \beta_1, \cdots, \beta_T}} \sum_{t=1}^{T} \parallel Y_t - (Z + Z_t^s)\beta_t \parallel^2 + \lambda \sum_{t=1}^{T} \parallel Z_t^s \beta_t \parallel^2$$

While non-negative factorization approaches have a naturally sparse representation [32],

iNMF also applys an $l1$-penalization on the data-specific term to induce sparsity. This constraint imposed on the data-specific effects implies that the parameter $\lambda$ controls for the factorization homogeneity. The authors also propose to apply an $l1$-penalty on the coefficient matrix $\beta_k$ to enforce variable selection. The dimension of the shared and specific structures is chosen through a consensus-based approach that selects the number of latent variables maximizing the clustering stability across multiple iNMF runs.

Similarly to JIVE, no guidelines are provided to obtain a final sample clustering. And iNMF is limited to continuous data without incorporation of pathway information.

### 2.7 Bayesian consensus clustering (BCC)

When it comes to integrative approaches that contructed based on Bayesian framework setting, Bayesian consensus clustering (BCC) [18] is a Bayesian method that represents each data set with a Dirichlet-multinomial allocation mixture model [19]. BCC aims to uncover a single common clustering across sources by relating the source-specific clustering $L_t$ in data set $t$ to a consensus clustering through the following dependence function:

$$P(L_{tn} = l \mid C_n) = \begin{cases} \alpha_t, & \text{if } C_n = 1 \\ \frac{1-\alpha_t}{1-q}, & \text{otherwise} \end{cases}$$

where, for sample $n$, $C_n$ and $L_{tn}$ are the overall and source-specific cluster allocations in data source $t$, $\alpha_t$ is the adherence of data set $t$ to the overall clustering and $q$ is the maximum number of clusters (both shared and source specific). The adherence parameter $\alpha_t$ models how specific and shared clusters are related to each other. The parameter $q$ is chosen so that the mean adherence over the sources is maximized, which results in a smaller number of selected clusters. For BCC model, a Gibbs sampler is used to estimate the posterior distribution of the parameters.

BCC model allows only continuous data, and it doesn't promote embedding of pathway information in the model. Note that the BCC model assumes a simple and general dependency structure between data sources. When an overall clustering is not sought, or when such a clus-

tering does not make sense as an assumption, BCC may be no longer be appropriate. In addition, this two-step procedure of separate clusterings followed by $post\ hoc$ integration limits the power to detect the structure shared between different omics data.

### 2.8 Integrative factor analysis model (iFad) and Bayesian sparse factor modeling for the inference of pathways responsive to drug treatment (FacPad)

In order to study gene-pathway–drug-pathway associations and integrate gene expression and drug sensitivity data, a Bayesian sparse factor analysis model (iFad) was developed to jointly analyze the paired gene expression and drug sensitivity datasets measured for same samples [20]. iFad enables direct incorporation of prior knowledge of gene-pathway and/or drug-pathway associations by using pathway to guide factor definition, where pathways are treats as latent factors. In order to use the prior knowledge of pathway information, spike-and-slab mixture prior is used [33] for the factor loading matrices. Ma and Zhao then used a modified collapsed Gibbs sampling algorithm for statistical inference.

Later on, Ma and Zhao developed FacPad, a Bayesian sparse factor model, for the inference of pathways responsive to drug treatments [21]. Similar to iFad, this approach also represents biological pathways as latent factors. The difference of FacPad compared to iFad is that it calculates the similarity among different drugs and tries to detect the association strength between drugs and pathways through large-scale data mining. It is worthy of note that the FacPad model is developed under the assumption that all the latent factors needed for decomposition of the data matrix are known, where the latent factors are matched to KEGG pathways.

For both iFad and FacPad, the limitation is excessive number of latent components in the model, MCMC approach is time-consuming when applied to high-dimensional inference. Also, it is important to note that the model allows to be applied to only continuous data.

## 2.9 Semi-supervised identification of cancer subgroups using survival outcomes and overlapping grouping information (InGRiD)

Another newly developed pathway-level analysis approach is Integrative Genomics Robust iDentification of cancer subgroups (InGRiD) [6]. Here Wei et al aimed to address the issue of low reproducibility and instability when it comes to identification of cancer subgroups and molecular features. InGRiD integrates pathway information with gene expression data to improve the robustness and accuracy for identification of key molecular feature and gene sets as well as cancer patients subgroups. [6].

InGRiD is a multi-step semi-supervised approach. A Sparse partial least squares (SPLS) Cox regresion model is constructed for gene-level analysis [34]. For each pathway, InGRiD simutaneously summarizes its genes as a latent component (dimension reduction) and selects key genes among those (variable selection). When it comes to pathway-level analysis, a LASSO-penalized Cox regression model is built on SPLS latent components to obtain a parsimonious set of key pathways. InGRiD also provides patient subgroup resulting as low, intermediate and high risk group of patient.

A limitation of InGRiD is that the multi-stage setting is essentially making pathway-level predictions based on prediction result from previous step (gene-level). Gene-level and Pathway-level information are not integrated within a unified model to guide gene selection. One should also notice InGRiD is not an integrative method and it limits to only continuous data.

## 2.10 Robust enumeration of cell subsets from tissue expression profiles (CIBERSORT)

In recent years, many appoaches have been purposed to estimate the immune cell composition in clinical tumour samples. One of the widely used approach is CIBERSORT (Robust enumeration of cell subsets from tissue expression profiles), which was developed by Newman and others to process cell composition of tumour cells from their gene expression profiles [9]. CIBERSORT implements support vector regression (SVR) to improve deconvolution performance through a combination of variable selection and other optimization techniques.

CIBERSORT is validated by many researchers that it is a useful approach for high throughput characterization of cell types [22].

The objective of most gene expression deconvolution algorithms, including CIBERSORT, is to solve the following system of linear equations for $\mathbf{f}$:

$$\mathbf{m} = \mathbf{f} \times \mathbf{B}$$

where $\mathbf{m}$ is defined as a vector of a mixture gene expression profiles, $\mathbf{f}$ is a vector containing the fraction of each cell type in the signature matrix, and $\mathbf{B}$ is a "signature matrix" that includes signature genes for cell subsets of interest. Notice that $\mathbf{m}$ is the input data, $\mathbf{B}$ is the signature information that are known a priori, and $\mathbf{f}$ is unknown.

The difference between CIBERSORT and previous deconvolution methods is in its application of a machine learning technique, $v$-support vector regression ($v$-SVR), to solve for $\mathbf{f}$. More specifically, SVR uses a hyperplane that includes as many data points as possible given pre-specified constraints, and reduces overfitting by only penalizing data points outside support vectors using a loss function. $f$ is determined by orientation of the hyperplane. The parameter $v$ determines the lower bound of support vectors and the upper bound of training errors. In addition, $v$-SVR incorporates $L2$-norm regularization, which minimizes the variance in the weights assigned to highly correlated cell types, thereby mitigating issues owing to multicollinearity.

Despite the various successful applications of computational methodologies, several issues remain to be improved. For example, since deconvolution is sensitive to noise, it is in need to improve the robustness of the method. [23].

### 2.11 Covariance-Based Variable Selection for Compositional Data

To address the growing interest in development of immune therapy of cancer related disease, researchers are often facing the challenge of analyzing compositional data of immune cell composition in clinical tumour samples. In order to perform Compositional Data Analysis (CoDA), an widely used approach is to transform the compositional data to ratios using the centred

log-ratio (clr) [7] transformation. However, one must realize the limitation that the variable selection applied on the clr-transformed variables makes interpretation challenging. For that purpose, Hron and others [24] developed an covariance-based stepwise procedure to omit variables in compositional data analysis.

First, the sample space can be defined as $\boldsymbol{x} = (x_1, \cdots, x_p)^T$, let $\boldsymbol{y} = clr(\boldsymbol{x}) = (y_1, \cdots, y_p)^T$. An important theorem of the normalized variance is that $var(y_i) \geq var(y_j)$ if and only if

$$\sum_{k=1}^{p} var(ln\frac{x_i}{x_k}) \geq \sum_{k=1}^{p} var(ln\frac{x_j}{x_k})$$

Next, if we consider a composition $\boldsymbol{x} = (x_1, \cdots, x_p)^T$ such that $var(y_1) \geq \cdots \geq var(y_p)$, more specifically,

$$\sum_{k=1}^{p} var(ln\frac{x_1}{x_k}) \geq \sum_{k=1}^{p} var(ln\frac{x_2}{x_k}) \geq \cdots \geq \sum_{k=1}^{p} var(ln\frac{x_p}{x_k})$$

The step-wise algorithm can be performed as follow:

1. Eliminate variable $x_p$ whose variance of the corresponding clr variable is the smallest.
2. Perform clr transformation on the induced subcomposition $\boldsymbol{x}_1$, where $\boldsymbol{x}_1 = (x_1, \cdots, x_{p-1})^T$.
3. Calculate the variances of the elements in the transformed subcomposition $\boldsymbol{x}_1$.
4. Repeat step 1 to 3 until $H_0$ is rejected or the number of steps reaches p-2.

Here the null hypothesis is that the total variance of $\boldsymbol{x}_i$ being equal to the total variance of $\boldsymbol{x}_{i-1}$, while the alternative is the total variance of $\boldsymbol{x}_i$ being less than the total variance of $\boldsymbol{x}_{i-1}$, Given these hypotheses, the null is rejected if proposed test statistics $U_i^+ < Z_{0.95}$ where

$$U_i^+ = \frac{\hat{totvar}(\boldsymbol{x}_i) - totvar(\boldsymbol{x}_{i-1})}{\sqrt{\frac{2}{n-1}tr(\hat{\Sigma_i^2})}}$$

And $\hat{\Sigma}_i$ represents the sample covariance matrix of the composition $\boldsymbol{x}_i$ in clr coordinates.

The goal of this procedure is to eleminate components in compositional data based on the total variance until a significant reduction would occur. And the procedure allows us to reach

a subcomposition that still retains the important information contained in the multivariate data structure. The strength of the algorithm is that it make the loss of the information rather negligible when moving from composition to subcomposition.

### 2.12 Stepwise Pairwise Log-ratio Variable Selection for Compositional Data

Another variable selection approach for compositional data is proposed by Greenacre [8] where all pairwise ratios of parts are considered for key marker identification. A smaller set of ratios can be chosen to explain as much variability as required to reveal the underlying structure of the data.

We consider the compositional data that are made up of the relative proportions of a whole and can be represented in the simplex of $d$ parts:

$$S^d := \{x = (x_1, \cdots, x_d) \in R^d | \sum_{i=1}^{d} \theta_i = 1, \theta_i \geq 0, \forall i\}$$

The basic measure of variability of a random composition $x = (x_1, \cdots, x_d)$ is the variation matrix [7], defined as

$$\mathbf{T} = \left\{ var\left( ln\frac{x_i}{x_j} \right) \right\}_{i,j=1}^{D}$$

Each element in the variation matrix defines the variability of the log-ratio $ln\frac{x_i}{x_j}$ : The log-ratio tends to be a constant if the value of the variance is small. Total variance is defined as the sum of the elements of the variation matrix, where

$$totvar(x) = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} var\left( ln\frac{x_i}{x_j} \right)$$

Redundancy analysis (RDA) were used to measure how much of the total variance is explained by a subset of logratios of certain explanatory variables after the logratio variance are calculated. RDA is a form of multivariate regression. If a variable is correlated with many of

the other logratios, the explained variance of the corresponding variable will be high. In addition, Procrustes analysis was applied to decide how close their multivariate structures are, more specifically, it decide how close a configuration based on a subset of logratios is to the configuration based on all the logratios. Procrustes correlation is the measurement for the matching of two configurations.

The stepwise procedure variable selection in the compositional data can proceed as follow:

1. Calculate all the pairwise logratios.

2. Select the one with the highest percentage of variance explained. This ratio is then fixed as the first logratio.

3. The second best logratio in combination with the first is sought, then fixed, and so on.

4. Repeat step 1 to 3 until variance explained reach 100%.

It must take into account that one should choose ratios that are independent of the ones already chosen: for example, if A/B and B/C have already been selected, then $A/C$ is no longer a candidate for selection, since it depends on the others: $A/C = A/B \times B/C$ . On the log scale, $log(A) - log(C)$ is the sum of, and thus linearly dependent on, $log(A) - log(B)$ and $log(B) - log(C)$. Since the dimensionality of an $m$-part compositional data set is $m - 1$, and all the parts will have appeared in at least one logratio after $m - 1$ steps of the above procedure, the variance explained will be 100%.

### 2.13 Zero-inflated Wilcoxon Rank Sum Test (ZIW)

Zero-inflated Wilcoxon test was first proposed in 2010 by Hallstrom [10] and it is further modified by Wang and others. [35]. The theory of the zero-inflated Wilcoxon rank sum test is as follow. We consider $2N$ patients in a randomized study, where $N$ patients are assigned to the treatment group $T_1$ and control group $T_2$, respectively . We define $f_1$ and $f_2$ as the distributions of the non-zero values under $T_1$ and $T_2$. Let $n_i$ be the number of non-zero scores in each group, $n = max(n1, n2)$ and $m = |n1 - n2|$ . Without loss of generosity, we assume there are no ties

among the 2nm non-zero scores. In order to compute the rank-sums, we assign rank 1 to the highest score, rank 2 to the second highest score and so on. Hence, we have 2(Nn)+m zeros tied at the highest rank.

The zero-inflated Wilcoxon rank sum test will be based on the $2n$ observations remaining when $N - n$ observations with zero score have been removed from each group. Let $r$ be the sum of the ranks of the observations in group 1 among all 2n observations. Let r0 be the sum of the ranks of the non-zero scores of group 1. Then

$$r = \begin{cases} r_0 + m\dfrac{(2n - m + 1 + 2n)}{2}, & \text{if } n_1 \leq n_2 \\ r_0, & \text{if } n_1 \geq n_2 \end{cases} \tag{2.1}$$

and under the null $f_1 = f_2$, the Wilcoxon rank-sum statistic, $s = r - N(2N + 1)/2$, satisfies

$$E(s = r - n(2n + 1)/2|n1, n2) = \begin{cases} mn/2, & \text{if } n_1 \leq n_2 \\ -mn/2 & \text{if } n_1 \geq n_2 \end{cases} \tag{2.2}$$

Then

$$\begin{aligned} Var(s|n_1, n_2) &= Var(r|n_1, n_2) \\ &= Var(r_o|n_1, n_2) \\ &= n(nm)(2nm + 1)/12 \\ &= n^3/6 + nm^2/12 - mn^2/4 + n^2/12 - nm/12 \end{aligned} \tag{2.3}$$

Let $\mu_{i,j} = E((n/N)^i(m/N)^j)$. Then

$$E(Var(s|n_1, n_2)) = N^3(\mu_{3,0}/6 + \mu_{1,2}/12 - \mu_{2,1}/4) + N^2(\mu_{2,0} - \mu_{1,1})/12 \tag{2.4}$$

Under the null, it is equally likely that $n_1$ is less than or greater than $n_2$, so $E(s) = 0$ and $Var(E(s|n_1, n_2) = E((mn/2^)2) = N^4\mu_{2,2}/4$. Since $Var(s) = Var(E(s|n_1, n_2)+E(Var(s|n_1, n_2))$,

it follows

$$Var(s) = N^4\mu_{2,2}/4 + N3(\mu_{3,0}/6 + \mu_{1,2}/12 - \mu_{2,1}/4) + N^2(\mu_{2,0} - \mu_{1,1})/12$$

It is defined that the zero-inflated Wilcoxon rank sum test by $W = s/\sqrt{Var(s)}$.

## 3. SPECIFIC AIM 1

For Aim 1, our goal is to develop a Bayesian sparse latent factor model for pathway-guided identification of cancer subgroups by integrating multiple types of genomic data. This statistical method will provide researchers a unified framework to simultaneous identify cancer subgroups (clustering) and key molecular markers (variable selection) based on the joint analysis of continuous, binary and count data. In addition, we plan to use Polya-Gamma mixtures of normal for binary and count data to promote an exact and fully automatic posterior sampling. Moreover, pathway information will be used to improves accuracy and robustness of cancer subgroup and key molecular features identification.

### 3.1 Introduction

In cancer genomics, it is of critical interest to identify cancer patient subgroups as it can facilitate development of personalized medicine. The identification of novel molecular features associated with these patient subgroups can potentially lead to a novel biomarker for prognosis, and diagnosis and novel therapeutic targets. The emergence of comprehensive cancer genomics platform, such as The Cancer Genome Atlas (TCGA) [2], 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 [2]. At the same time, the availability of large-scale high-throughput multi-omics data sets requires 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 form, and hence, there is need for a statistical approach that is also capable of handling various types of data.

Traditionally, principal component analysis (PCA) [36] 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 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, e.g., iCluster [11], iNMF [16] and JIVE [15]. Specifically, iCluster features a joint latent variable model and cancer subgroups can be identified by applying a clustering algorithm on the shared latent factors [11], 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 [16] [15]. Although integrative approaches like JIVE and iNMF promoted 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 [12] [13]. Recently, Mo and others improved iCluster+ and developed a Bayesian sparse latent factor model to integrate multiple types of omics data, called iClusterBayes [14]. 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 us to avoid complicated parameter tuning required when a penalization algorithm is used; and (iii) one could incorporate prior biological expert knowledge. This new method enables a posterior probability estimation for gene selection and improves the iCluster+ method regarding computational speed significantly [14].

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 [29]. iFad [20] and PacFad [21] enable 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 [6] 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 (SPLS) Cox regression model [34], 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 (i.e., continuous, discrete) at the pathway level.

Another limitation of the iClusterBayes approach is the need for Metropolis–Hastings sampling [**?**] 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. Although the Metropolis–Hastings sampling is popularly used, it still can be less computationally efficient compared to the Gibbs sampler and involves parameter tuning, which might not be straightforward in practice. Hence, an alternative posterior sampling strategy without a 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 [37]. Pillow and Scott further extended the model to handle negative binomial (NB) case [38] 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.

### 3.2 Methods

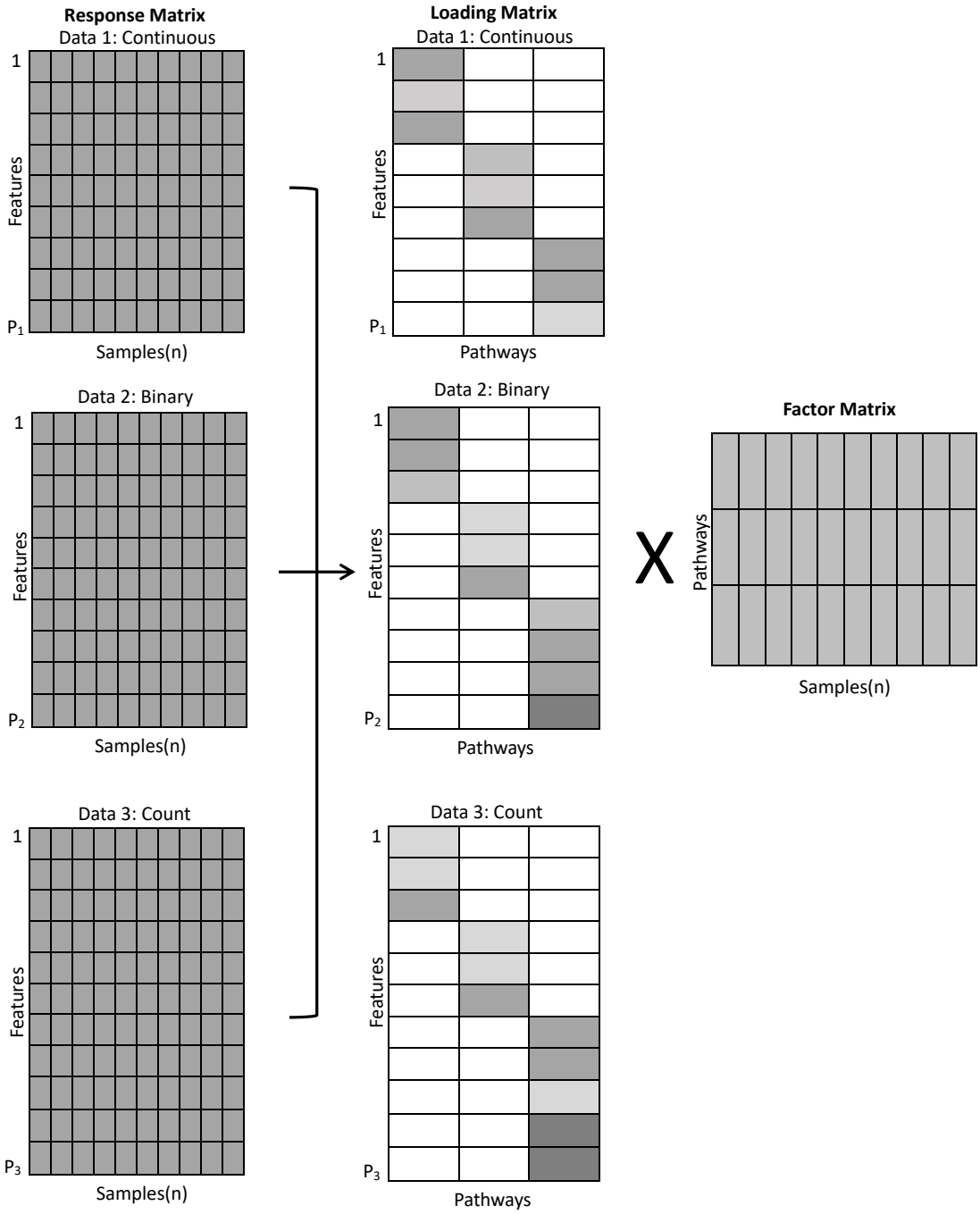
Here our main goals include (i) construction of a natural and unified framework for integrative analysis; (ii) incorporating prior biological knowledge; and (iii) implementing efficient posterior sampling. We developed our model based on a Bayesian sparse latent factor model equipped with the Pólya-Gamma 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 $\mathbf{y}_{it} = (y_{i1t}, \cdots, y_{ip_tt})^T$ to be the data vector, where $y_{ijt}$ denotes genomic measurement for the $j$-th molecular feature $(j = 1, \cdots, p_t)$ of the $i$-th sample $(i = 1, \cdots, n)$ in the $t$-th data type $(t = 1, \cdots, m)$. 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 $\mathbf{Z} = (\mathbf{z}_1, \cdots, \mathbf{z}_n)^T$. We define $\mathbf{z}_i = (z_{i1}, \cdots, z_{ik}), i = 1, \cdots, n$, 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 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.

#### 3.2.1 Bayesian latent factor model

In this section, we will first introduce the Bayesian latent factor model framework, which is motivated by the iClusterBayes [14] approach. If $y_{ijt}$ is a continuous variable, we assume the following model,

$$y_{ijt} = \mathbf{z}_i \mathbf{\Gamma}_{jt} \beta_{jt} + \epsilon_{ijt}, i = 1, \cdots, n, j = 1, \cdots, p_t, t = 1, \cdots, m, \tag{3.1}$$

where $\mathbf{z}_i = (1, z_{i1}, \cdots, z_{ik})$ is a latent factor vector of the $i$th sample; $\mathbf{\Gamma}_{jt} = diag(1, \gamma_{jt}, \cdots, \gamma_{jt})$ is a diagonal matrix serving as an indicator variable, where $\gamma_{jt}$ takes values of either 0 or 1 for variable selection [39]; $\beta_{jt} = (\beta_{0jt}, \beta_{1jt}, \cdots, \beta_{kjt})^T$ denote the coefficient vector of the




[para]

Figure 3.1: Illustration of the Bayes-InGRiD framework

$j$-th feature in the $t$-th data set. We assume $\epsilon_{ijt} \sim N(0, \sigma^2_{jt})$. 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 $\beta_{jt}$ 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.

$$log\Big(\frac{P(y_{ijt} = 1 \mid \mathbf{z}_i)}{1 - P(y_{ijt} = 1 \mid \mathbf{z}_i)}\Big) = \mathbf{z}_i\mathbf{\Gamma}_{jt}\beta_{jt}, i = 1, \cdots, n, j = 1, \cdots, p_t, t = 1, \cdots, m. \tag{3.2}$$

Moreover, if $y_{ijt}$ is a count variable, we assume the following negative binomial (NB) regression model.

$$\begin{aligned} P(y_{ijt} \mid r_{jt}, \mathbf{z}_i) &= \frac{\Gamma(y_{ijt} + r_{jt})}{\Gamma(r_{jt})y_{ijt}!}(1 - \psi_{ijt})^{r_{jt}}\psi_{ijt}^{y_{ijt}}, r_{jt} > 0, j = 1, \cdots, p_t, t = 1, \cdots, m \\ logit(\psi_{ijt}) &= \mathbf{z}_i\mathbf{\Gamma}_{jt}\beta_{jt}, \end{aligned} \tag{3.3}$$

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

$$P(y_{ijt}, \mathbf{z}_i \mid \beta_{jt}, \mathbf{\Gamma}_{jt}) \propto \prod_{t=1}^{m}\prod_{i=1}^{n}\prod_{j=1}^{p_t} P(y_{ijt} \mid \mathbf{z}_i, \beta_{jt}, \mathbf{\Gamma}_{jt})P(\mathbf{z}_i), \tag{3.4}$$

where $\mathbf{z}_i$ follows a standard multivariate normal distribution $MVN(\mathbf{0}, \mathcal{I}_k)$; $P(y_{ijt} \mid \mathbf{z}_i, \beta_{jt}, \mathbf{\Gamma}_{jt})$ is the conditional density function, where the form of distribution of $P(y_{ijt} \mid \mathbf{z}_i, \beta_{jt}, \mathbf{\Gamma}_{jt})$ 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 above models by introducing scale mixtures of normals using Pólya-Gamma in the following two subsections.

### 3.2.2 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 [38]. Assuming a random variable $\omega$ has a Pólya-Gamma distribution, an important property of the $PG(b,0)$ density – is that for $a \in \Re$, $b > 0$ and $\eta \in \Re$,

$$\frac{(e^{\eta})^a}{(1+e^{\eta})^b} = 2^{-b} e^{\kappa\eta} \int_0^{\infty} e^{-\omega\eta^2/2} p(\omega|b,0) d\omega, \tag{3.5}$$

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

$$P(y_{ijt} \mid \mathbf{z_i}) = \frac{(e^{\mathbf{z}_i \mathbf{\Gamma}_{jt} \beta_{jt}})^{y_{ijt}}}{1+e^{\mathbf{z}_i \mathbf{\Gamma}_{jt} \beta_{jt}}} = \frac{(e^{\eta_{ijt}})^{y_{ijt}}}{1+e^{\eta_{ijt}}}, i = 1, \cdots, n, j = 1, \cdots, p_t, t = 1, \cdots, m, \tag{3.6}$$

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

$$P(y_{ijt} \mid r_{jt}, \beta) = e^{\kappa_{ijt}\eta_{ijt}} \int_0^{\infty} e^{-\omega_{ijt}\eta_{ijt}^2/2} p(\omega_{ijt}|r_{jt}+y_{ijt},0) d\omega_{ijt}, \tag{3.7}$$

where $\kappa_{ijt} = y_{ijt} - 1/2$ and the $\omega_{ijt}$ 's are independently distributed according to $PG(1, \eta_{ijt})$. By using to the above properties of the Pólya-Gamma distribution, we can show the full conditional distribution of $\boldsymbol{\beta}$ is

$$P(\beta_{jt} \mid \mathbf{Z}, \mathbf{y}_{jt}, \mathbf{\Omega}_{jt}, \mathbf{\Gamma}_{jt}) \propto P(\beta_{jt}) exp\{\frac{1}{2}(\mathbf{u}_{jt} - \mathbf{Z}\mathbf{\Gamma}_{jt}\beta_{jt})^T \mathbf{\Omega}_{jt} (\mathbf{u}_{jt} - \mathbf{Z}\mathbf{\Gamma}_{jt}\beta_{jt})\}, \tag{3.8}$$

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

### 3.2.3 Bayesian latent factor model for count data

By parameterizing the NB probability parameter $\psi_{ijt}$ 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 lo-

gistic case [37]. Exploiting the earlier property of the Pólya-Gamma distribution with Equation (5), it follows that $\kappa_{ijt} = (y_{ijt} + r_{jt})/2$ and the $\omega_i$ 's are independently distributed according to $PG(y_{ijt} + r_{jt}, \eta_{ijt})$ [37]. 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

$$P(\beta_{jt} \mid \mathbf{Z}, \mathbf{y}_{jt}, \boldsymbol{\Omega}_{jt}, \boldsymbol{\Gamma}_{jt}) \propto P(\beta_{jt}) exp\{\frac{1}{2}(\mathbf{u}_{jt} - Z\boldsymbol{\Gamma}_{jt}\beta_{jt})^T \boldsymbol{\Omega}_{jt}(\mathbf{u}_{jt} - Z\boldsymbol{\Gamma}_{jt}\beta_{jt})\}, \tag{3.9}$$

where $\mathbf{u}_{jt} = (u_{1jt}, \cdots, u_{njt})$ is a length n vector and its $i$-th 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 sampling of $r_{jt}$ [40] [41]. The approach introduces a sample of latent counts, $l_{ijt}$, underlying each observed count $y_{ijt}$. Regarding sampling of over-dispersion parameter, conditional on $y_{ijt}$ and $r_{jt}$, $l_{ijt}$ has a distribution defined by a CRT distribution:

$$\begin{aligned} l_{ijt} &= \sum_{d=1}^{y_{ijt}} \mu_d \\ \mu_d &\sim Bern\Big(\frac{r_{jt}}{r_{jt} + d - 1}\Big). \end{aligned} \tag{3.10}$$

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. By applying the two-step conjugate Gibbs update for $r_{jt}$ [40], we first draw $l_{ijt}$ according to this CRT distribution. Next, NB distribution can be derived from a random convolution of logarithmic random variables. Specially, they note that, conditional on $r_{jt}$ and $\psi_{ijt}$,

$$\begin{aligned} l_{ijt} &\sim Poisson[-r_{jt}\, ln(1 - \psi_{ijt})] \\ \psi_{ijt} &\sim \frac{e^{\mathbf{z}_i \boldsymbol{\Gamma}_{jt} \beta_{jt}}}{1 + e^{\mathbf{z}_i \boldsymbol{\Gamma}_{jt} \beta_{jt}}} \end{aligned} \tag{3.11}$$

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

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

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 $\psi_{jt}$.

### 3.2.4 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 $G = \{G_1, \cdots, G_s\}$ for $s$ pathways. 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 $\beta_{jt}$ of dimension $p$ by $s + 1$, where $p = \sum_i^s G_i$. Then we put constraints on the factor loading matrix, where the genes in $\beta_{jt}$ that belong to certain pathways are free to update, while the remaining elements in $\beta_{jt}$ 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 $\beta$ from $N(0, 1)$, and force the remaining elements of the first latent factor to be zero. To address the issue of identifiability [42], we update the first non-zero elements of each latent factor using the truncated Gaussian distribution $TN(0, 1, 0, \infty)$. In the pathway-level analysis model, we set the number of latent factors $k$ equal to the number of pathways. We cluster the patients into $k$ subgroups using a $k$-means approach [43]. 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 $\mathbf{Z}$, since it is shared across the data-type-specific models. By joint analysis of multiple data, the patient subgroups can be identified using the latent factor matrix $\mathbf{Z}$ and key molecular feature $j$ that drives the sample clustering can be identified in a given set $t$. In

particular, the model for feature $j$ of $i$-th sample is given as

$$\mathbf{u}_{ij} - \beta_{0j} = \mathbf{\Gamma}_j \beta_j z_i + \epsilon_{ijt}, i = 1, \cdots, n, j = 1, \cdots, p_t, t = 1, \cdots, m, \tag{3.12}$$

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

$$u_{ijt} = \begin{cases} y_{ijt}, & \text{if } t\text{-th data type is continuous,} \\ \dfrac{y_{ijt} - \dfrac{1}{2}}{\omega_{ijt}}, & \text{if } t\text{-th data type is binary,} \\ \dfrac{y_{ijt} - r_{jt}}{2\omega_{ijt}}, & \text{if } t\text{-th data type is count.} \end{cases} \tag{3.13}$$

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

$$\epsilon_{ijt} \sim \begin{cases} N(0, \sigma_{jt}^2), & \text{if } t\text{-th data type is continuous,} \\ N(0, \omega_{ijt}^{-1}), & \text{if } t\text{-th data type is binary,} \\ N(0, \omega_{ijt}^{-1}), & \text{if } t\text{-th data type is count.} \end{cases} \tag{3.14}$$

We define $\mathbf{\Sigma} = diag(Var(\epsilon_{ij1}), \cdots, Var(\epsilon_{ijm}))$, where $\Sigma$ is the diagonal variance-covariance matrix whose diagonal components are variance of random errors. As for prior distributions of model parameters, we assume $\beta_{jt} \sim MVN(\beta_{0t}, \mathbf{\Sigma}_{0t})$, $\sigma_{jt}^2 \sim IG(v_0/2, v_0\sigma_0^2)$, and $\gamma_{jt} \sim Bernoulli(q_t)$, where the coefficient vector $\beta_{jt}$, $\sigma_{jt}^2$, the indicator variable $\gamma_{jt}$ 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 $t$-th

data type as the following:

$$P(\sigma_{jt}^2 \mid \mathbf{Z}, \mathbf{y}_{jt}, \beta_{jt}, \mathbf{\Gamma}_{jt}) = IG\Big(\frac{V_0+n}{2}, \frac{V_0\sigma_0^2 + (\mathbf{y}_{jt} - \mathbf{Z}\mathbf{\Gamma}_{jt}\beta_{jt})^T(\mathbf{y}_{jt} - \mathbf{Z}\mathbf{\Gamma}_{jt}\beta_{jt})}{2}\Big),$$

if $t$-th data type is continuous,

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

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

We have the conditional posterior distributions of $\beta_{jt}$ as follows.

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

Next, we define $\beta_j$ is an $m \times k$ matrix in which the $t$-th row is $(\beta_{1jt}, \cdots, \beta_{kjt})$. In word, it is $\beta_{jt}$ 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:

$$\begin{aligned} P(\mathbf{z}_i \mid \beta_j, \mathbf{y}_{ij}, \mathbf{\Sigma}, \mathbf{\Gamma}_j) &\sim MVN(\mu_n, \mathbf{\Sigma}_n), \text{where} \\ \mu_n &= \{\sum_j^p (\mathbf{\Gamma}_j\beta_j)^T\mathbf{\Sigma}^{-1}\mathbf{\Gamma}_j\beta_j + \mathcal{I}\}^{-1}\{\sum_j^p (\mathbf{\Gamma}_j\beta_j)^T\mathbf{\Sigma}^{-1}(\mathbf{y}_{ij} - \beta_{0j})\}, \\ \mathbf{\Sigma}_n &= \{\sum_j^p (\mathbf{\Gamma}_j\beta_j)^T\mathbf{\Sigma}^{-1}\mathbf{\Gamma}_j\beta_j + \mathcal{I}\}^{-1}. \end{aligned}$$

Since there is no closed-form 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 $\widetilde{\mathbf{\Gamma}_{jt}} = diag(1, 1-\gamma_{jt}, \cdots, 1-\gamma_{jt})$ as a $(k+1)\times(k+1)$ diagonal matrix, and $\widetilde{\psi_{ijt}} = e^{\mathbf{z}_i\widetilde{\mathbf{\Gamma}_{jt}}\beta_{jt}}/(1+e^{\mathbf{z}_i\widetilde{\mathbf{\Gamma}_{jt}}\beta_{jt}})$. Finally, we have the conditional posterior distributions of the indicator variable term $\mathbf{\Gamma}_{jt}$ depending on the $t$-th data type as follows.

$$P(\mathbf{\Gamma}_{jt} \mid \beta_{jt}, \mathbf{Z}, \mathbf{y}_{jt}, \sigma^2_{jt}) \propto \frac{exp\{-\frac{1}{2\sigma^2_{jt}}(\mathbf{y}_{jt} - \mathbf{Z}\mathbf{\Gamma}_{jt}\beta_{jt})^T(\mathbf{y}_{jt} - \mathbf{Z}\mathbf{\Gamma}_{jt}\beta_{jt})\}}{exp\{-\frac{1}{2\sigma^2_{jt}}(\mathbf{y}_{jt} - \mathbf{Z}\mathbf{\Gamma}_{jt}\beta_{jt})^T(\mathbf{y}_{jt} - \mathbf{Z}\mathbf{\Gamma}_{jt}\beta_{jt})\}) + exp\{-\frac{1}{2\sigma^2_{jt}}(\mathbf{y}_{jt} - \mathbf{Z}\widetilde{\mathbf{\Gamma}_{jt}}\beta_{jt})^T(\mathbf{y}_{jt} - \mathbf{Z}\widetilde{\mathbf{\Gamma}_{jt}}\beta_{jt})\}},$$

if $t$-th data type is continuous,

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

$$P(\mathbf{\Gamma}_{jt} \mid \beta_{jt}, \mathbf{Z}, \mathbf{y}_{jt}, r_{jt}) \propto \frac{\prod_{i=1}^{n} \frac{\Gamma(y_{ijt}+r_{jt})}{\Gamma(r_{jt})y_{ijt}!}(1-\psi_{ijt})^{r_{jt}}\psi_{ijt}^{y_{ijt}}}{\prod_{i=1}^{n} \frac{\Gamma(y_{ijt}+r_{jt})}{\Gamma(r_{jt})y_{ijt}!}(1-\psi_{ijt})^{r_{jt}}\psi_{ijt}^{y_{ijt}} + \prod_{i=1}^{n} \frac{\Gamma(y_{ijt}+r_{jt})}{\Gamma(r_{jt})y_{ijt}!}(1-\widetilde{\psi_{ijt}})^{r_{jt}}\widetilde{\psi_{ijt}}^{y_{ijt}}},$$ if $t$-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.

### 3.3 Simulation

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 amplified signal. To be specific, features with amplified and reduced signal 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 $\beta$ from $N(0,1)$, while forcing the rest of elements of the factor loading matrix to be zero. To address the issue of identifiability, $TN(0,1,0,\infty)$ 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. We used the uninformative priors for $\sigma^2_{jt}$ and $\gamma_{jt}$, where we set Inverse-gamma(1, 1) for $\sigma^2_{jt}$ and $Bernoulli(0.5)$ for the indicator variable $\gamma_{jt}$. In each simulation, we ran 20,000 MCMC iterations, and the first 10 000 were removed as burn-in.

Figure 3.1 A 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 $k = 2$. For all the gene-level analysis models, we used BIC to determine the optimal choice of $k$.

| Gene level | | |
|---|---|---|
| Signal level | Sensitivity (%) | Specificity (%) |
| $\mu$=0.8 | 30.4 | 96.7 |
| $\mu$=1.0 | 57.1 | 96.7 |
| $\mu$=1.2 | 82.1 | 96.7 |
| $\mu$=1.5 | 100.0 | 98.3 |
| Pathway level | | |
| Signal level | Sensitivity (%) | Specificity (%) |
| $\mu$=0.8 | 60.7 | 89.9 |
| $\mu$=1.0 | 82.1 | 91.5 |
| $\mu$=1.2 | 96.4 | 96.7 |
| $\mu$=1.5 | 96.4 | 96.7 |

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

Table 3.1: Feature selection performance for continuous data

Figures 3.1B and 3.1C present the posterior probabilities that 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 3.1 illustrates when signal level $\mu = 1$, and pathway-level analysis showed 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,

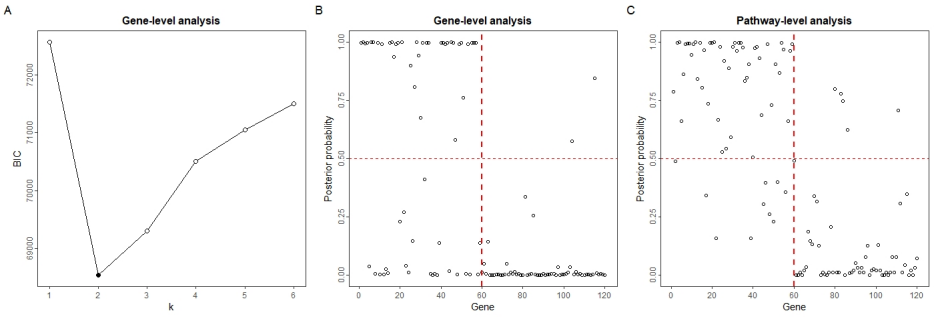



Fig. 3.1. Model and variable selection for the continuous data with $N(1,1)$ and $N(-1,1)$ as signal, $N(0,1)$ as background. (A) The BIC curve as a function of the number of latent components (k). the gene-level analysis model fits the data best when $k=2$. (B): Posterior probabilities of being informative features when $k=2$ for gene-level analysis. Genes with posterior probabilities greater than 0.5 are considered the driver for the patient subgroup clustering. (C): Posterior probabilities of being informative features for pathway-level analysis.

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 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 3.2A shows the BIC values for the gene-level analysis of the binary data with $Bernoulli(0.4)$ as signal and $Bernoulli(0.02)$ as the background. The best model fit was obtained when $k = 2$ based on BIC. Figures 3.2B and 3.2C 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 3.2 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.

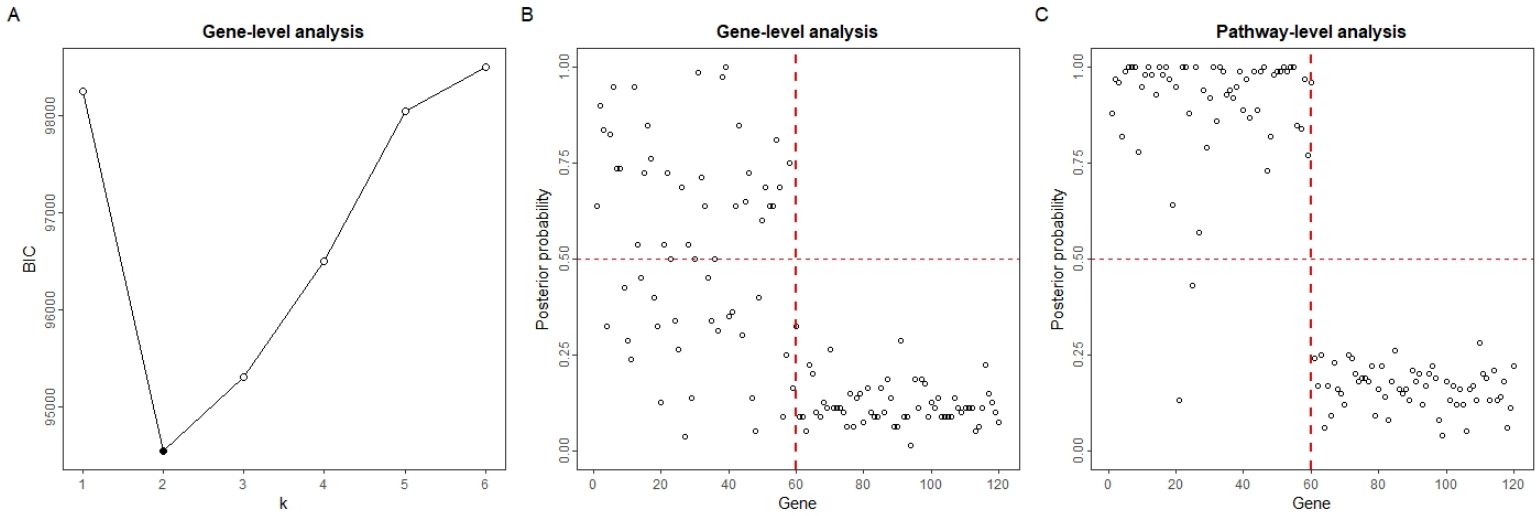



Fig. 3.2 Model and variable selection for the binary data with $Bernoulli(0.4)$ as signal, $Bernoulli(0.02)$ as background. (A) BIC and the gene-level analysis model fits the data best when $k = 2$. (B): Posterior probabilities of being informative features when $k = 2$ for gene-level analysis. Genes with posterior probabilities greater than 0.5 are considered the driver for the patient subgroup clustering. (C): Posterior probabilities of being informative features for pathway-level analysis.

| Gene level | | |
|---|---|---|
| Signal level | Sensitivity (%) | Specificity (%) |
| P=0.4 | 51.8 | 100.0 |
| P=0.5 | 89.4 | 100.0 |
| P=0.6 | 96.4 | 100.0 |
| P=0.7 | 100.0 | 100.0 |
| Pathway level | | |
| Signal level | Sensitivity (%) | Specificity (%) |
| P=0.4 | 96.4 | 100.0 |
| P=0.5 | 100.0 | 100.0 |
| P=0.6 | 100.0 | 100.0 |
| P=0.7 | 100.0 | 100.0 |

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

Table 3.2: Feature selection performance for binary data

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 amplified signal. Specifically, the count data for genes with amplified- and reduced-signal were randomly generated from $NB(\mu = \mu_1)$ and $NB(\mu = 1)$, respectively. We set different signal levels to evaluate model performance, where we set $\mu_1 = 7, 9, 11, 13$. The data for background genes were randomly generated from a $NB(\mu = (\mu_1 + 1)/2)$. Figure 3.3A 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 back-

ground, respectively. The best model fit was obtained when $k = 2$ based on BIC. Figures 3.3B and 3.3C demonstrate better performance for the pathway-level model in detecting true signal genes. Table 3.3 further confirms this observation across different signal-to-noise ratios.

| Gene level | | |
|---|---|---|
| Signal level | Sensitivity (%) | Specificity (%) |
| $\mu_1 = 7, \mu_2 = 1$ | 30.4 | 100.0 |
| $\mu_1 = 9, \mu_2 = 1$ | 48.2 | 100.0 |
| $\mu_1 = 11, \mu_2 = 1$ | 69.6 | 100.0 |
| $\mu_1 = 13, \mu_2 = 1$ | 78.6 | 100.0 |
| Pathway level | | |
| Signal level | Sensitivity (%) | Specificity (%) |
| $\mu_1 = 7, \mu_2 = 1$ | 87.5 | 98.9 |
| $\mu_1 = 9, \mu_2 = 1$ | 89.3 | 100.0 |
| $\mu_1 = 11, \mu_2 = 1$ | 92.9 | 100.0 |
| $\mu_1 = 13, \mu_2 = 1$ | 96.4 | 100.0 |

Sensitivity and specificity for the count data with $NB(\mu = \mu_1)$ and $NB(\mu = 1)$ as signal, $NB(\mu = (\mu_1 + 1)/2)$ as background.

Table 3.3: Feature selection performance for count data

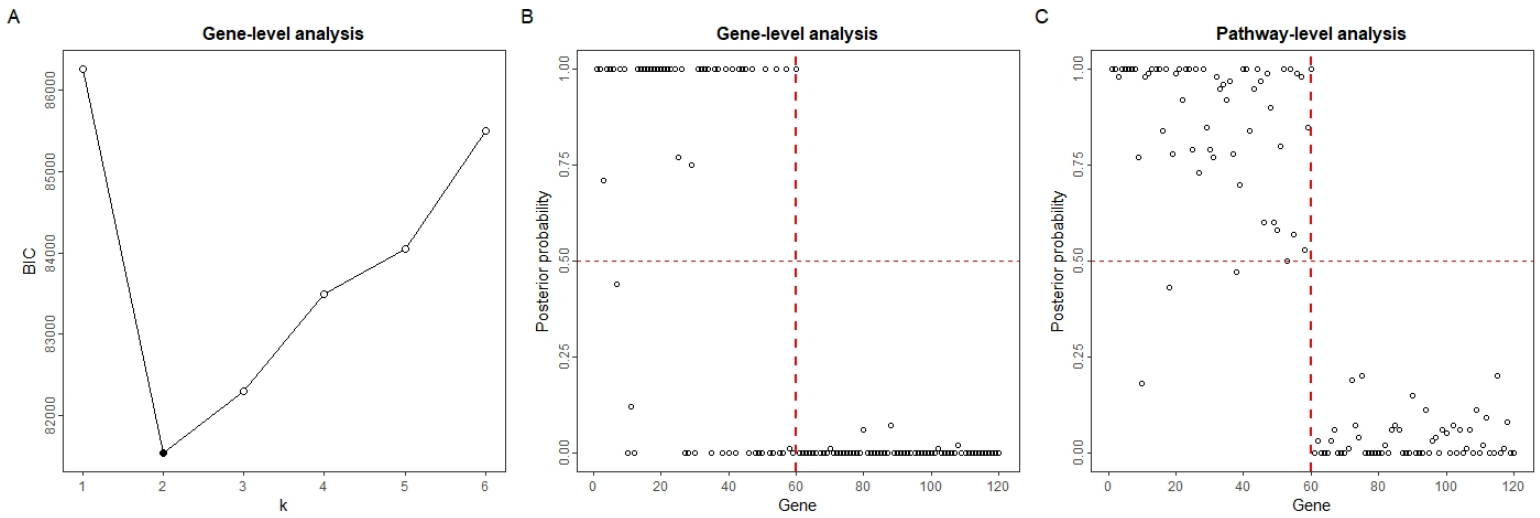



Fig. 3.3 Model and variable selection for the count data with $NB(\mu = 11)$ and $NB(\mu = 1)$ as signal, $NB(\mu = 6)$ as background. (A) BIC and the gene-level analysis model fits the data best when $k = 2$. (B): Posterior probabilities of being informative features when $k = 2$ for gene-level analysis. Genes with posterior probabilities greater than 0.5 are considered the driver for the patient subgroup clustering. (C): Posterior probabilities of being informative features for pathway-level analysis.

Finally, 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 A, the second 20 genes in patient subgroup B, and the third 20 genes in patient subgroup C to be characterized by 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 $k = 4$ as the number of pathways. Table 3.4 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.

| | 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 |

Samples are drawn from $N(0.8, 1)$ and $N(-0.8, 1)$ as signal, $N(0, 1)$ as background for continuous data; $Bernoulli(0.6)$ as signal, $Bernoulli(0.02)$ as background for somatic mutation data; $NB(\mu = 11)$ and $NB(\mu = 1)$ as signal, $NB(\mu = 6)$ as background for gene expression data, respectively. For gene-level analysis, we observe the minimum BIC value when k is equal to 2 for each signal level. For pathway-level analysis, k is set as 4.

Table 3.4: Feature selection performance for integrated data analysis and separate data analysis

### 3.4 Real data analysis

In this section, we used a cohort of high-grade serous ovarian cancer (HGSOC) patients from the TCGA project [2] 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 [**?**] [**?**]. In this analysis, we considered only the 1045 genes from the 15 previously profiled cancer signaling pathways [44], and the importance of these pathways has been discussed in the previous literature [45] [46].

To deal with the issue of overlapping genes among the 15 signaling pathways, we implemented the gene-cluster approach employed by InGRiD [6]. Specifically, if a gene is shared by multiple gene sets, this gene would be reallocated to a new pathway using the Partitioning Around Medoids (PAM) algorithm [47]. 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 Table S1 and S2 of the Supplementary Materials. We used the prior $Bernoulli(0.5)$ for the indicator variable $\Gamma_{jt}, j = 1, ..., p_t, t = 1, ..., m$, and we ran 20,000 MCMC iterations with first 10,000 iterations considered as burn-in. The optimal number of clusters $k$ was the one that maximizes the average silhouette over a range of possible values for $k$. We chose $k = 2$ by comparing $k = 1, \cdots, 10$.

The Bayes-InGRiD results for the mRNA expression data are presented in Table 3.5. 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.

| | 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 are ranked based on 'Pathway coefficient', which are their weighted averages of factor loadings of selected genes. 'Genes selected' refers to the the number of genes selected in each pathway, total number of genes in each pathway are also included within parenthesis in the column 'Genes selected'. Genes that rank top three in coefficient estimates would be shown in column 'Top three genes'

Table 3.5: Top pathways and genes selected for the mRNA expression data

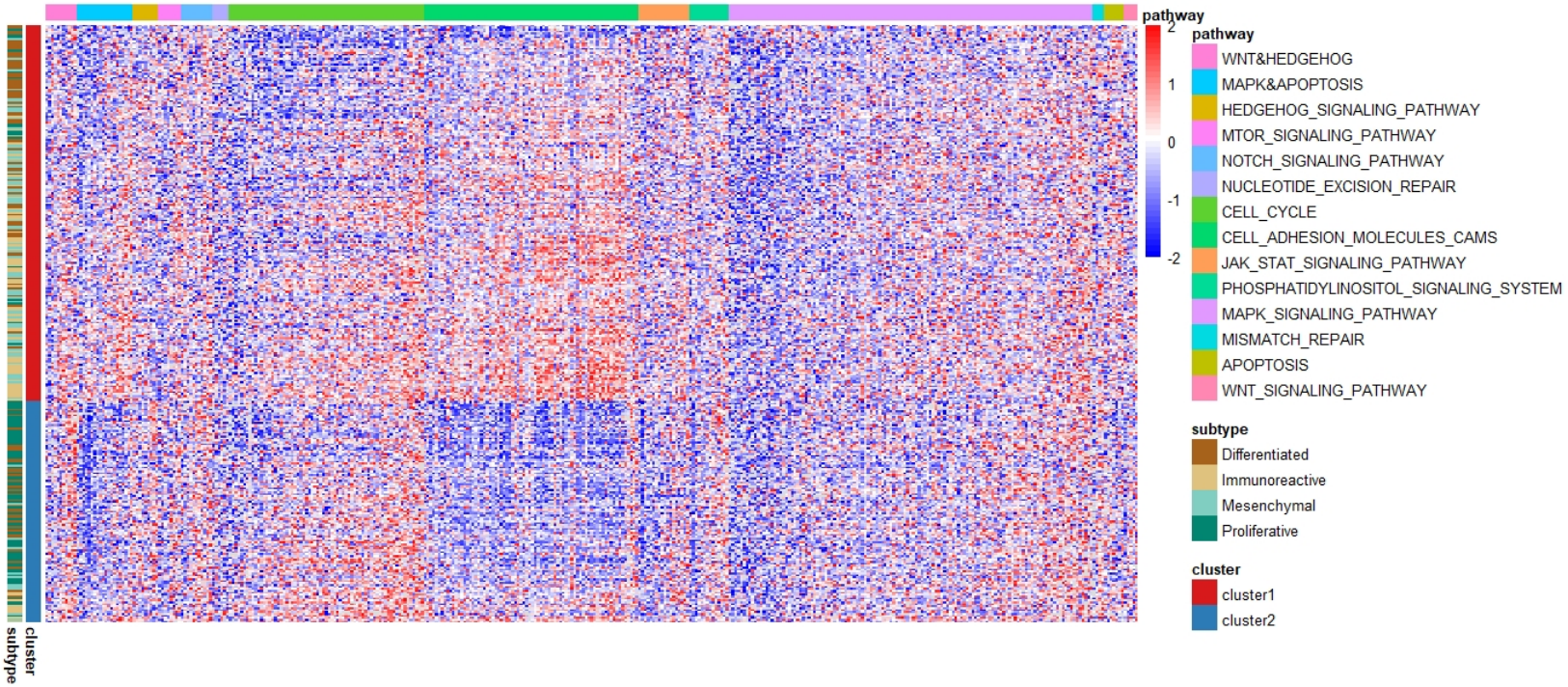



Fig. 3.4 Heatmap of the genes with posterior probability $> 0.5$ for mRNA data. The genomic pattern for gene expression is shown in the heatmap (red, high-level expression; blue, low-level expression). Column color bar labeled 'subtype' contains 4 expression subtypes classified in annotated TCGA subtypes including: differentiated, immunoreactive, mesenchymal and proliferative. Column color bar labeled 'cluster' shows the patient subgroups identified using Bayes-InGRiD. Row color bar shows the 14 pathways with genes selected.

To make sense of the patient subgrouping, we used the 4 expression subtypes classified in annotated TCGA subtypes by Noushmehr and Malta [48], where the patients are clustered into differentiated, immunoreactive, mesenchymal, and proliferative subtypes. Figure 3.4 demonstrates the heatmap of the selected genes for mRNA gene expression data. The integrative cluster 1 is highly correlated with immunoreactive and mesenchymal expression subtypes based on the overlapping color bars in Figure 3.4. The integrative cluster 2 is strongly correlated with the expression subtypes differentiated and proliferative. Figure 3.4 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 3.5 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.

| | 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 | NUMBL |
| 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 'Pathway coefficient', which are their weighted averages of factor loadings of selected genes. 'Genes selected' refers to the the number of genes selected in each pathway, total number of genes in each pathway are also included within parenthesis in the column 'Genes selected'. Genes that rank top three in coefficient estimates would be shown in column 'Top three genes'

Table 3.6: Top pathways and genes selected for the copy number data

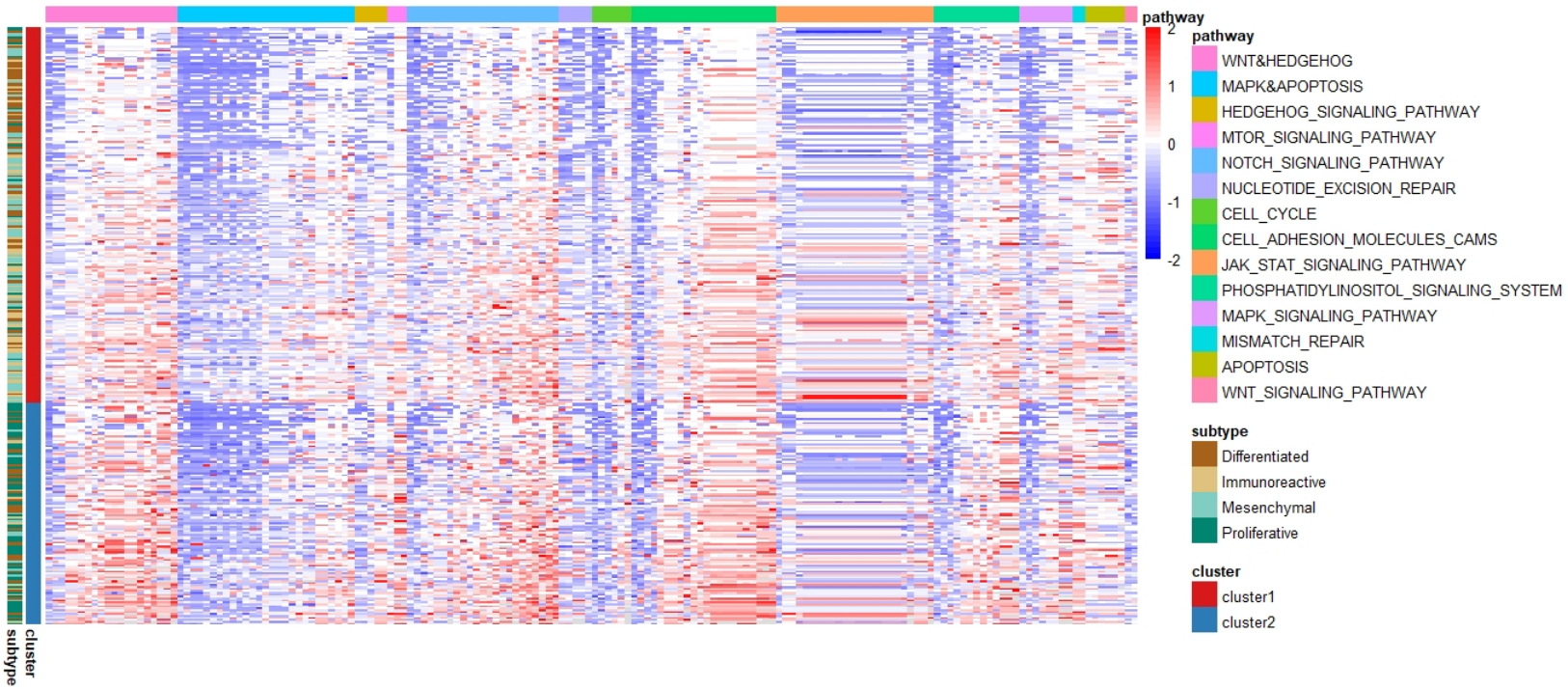



Fig. 3.5 Heatmap of the selected genes for the copy number data. The genomic pattern for gene expression is shown in the heatmap (red, high-level expression; blue, low-level expression). Column color bar labeled 'subtype' contains 4 expression subtypes classified in annotated TCGA subtypes including: differentiated, immunoreactive, mesenchymal and proliferative. Column color bar labeled 'cluster' shows the patient subgroups identified using Bayes-InGRiD. Row color bar shows the 14 pathways with genes selected.

## 3.5 Conclusions

In this aim, 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. In summary, Bayes-InGRiD can be a powerful approach for investigating cancer patient subgroups and their molecular features.

## 4. SPECIFIC AIM 2

For Aim 2, Out goal is to develop a comprehensive software implementing the method developed in Aim 1, and apply it to simultaneously identify sample clustering and key features in cancer genomic study. We aim to develop an user-friendly function called "BayesInGRiD" and provide it as a part of the R package "InGRiD".

### 4.1 Software Development

We have previously developed a unified statistical framework called "InGRiD" (Semi-supervised Identification of Cancer Subgroups using Survival Outcomes and Overlapping Grouping Information) [6] for pathway-guided identification of cancer patient subgroups. InGRiD has advantages of utilizing survival outcomes and addressing the issue of overlapping grouping information. I helped developing the R package called "InGRiD" in 2018 and it is currently publicly available in our research group GitHub webpage (https://dongjunchung.github.io/INGRID/).

The following document is the vignette for the R package "InGRiD":

# 1 Overview

This vignette provides basic information about the package for the pathway-guided identification of cancer subtypes. The proposed approach improves identification of molecularly-defined subgroups of cancer patients by utilizing information from pathway databases in the following four aspects.

(1) Integration of genomic data at the pathway-level improves robustness and stability in identification of cancer subgroups and driver molecular features;

(2) Summarizing multiple genes and genomic platforms at the pathway-level can potentially improve statistical power to identify important driver pathways because moderate signals in multiple genes can be aggregated;

(3) In INGRID, we consider the ''cooperation'' or ''interaction'' between pathways, instead assuming that each pathway operates independently during the cancer progression, which may be unrealistic;

(4) INGRID allows simultaneous inference in multiple biological layers within a statistically rigorous and unified framework without any additional laborious downstream analysis.

The package can be loaded with the command:

```
library(INGRID)
```

# 2 Input Data

The package requires that the response consist of 4 components: (1) gene expression z-scores in the form of a either data frame or matrix; (2) survival time and censoring indicator in the form of vectors; (3) pathway information in the form of a list, where each element is a vector of the names of gene belonging to the pathway.

In this vignette, a small subset of the Cancer Genome Atlas (TCGA) data is used to illustrate the 'INGRID' package. Specifically, we consider z-scores for the mRNA expression data of 389 genes for 50 randomly selected high-grade serous ovarian cancer patients, along with their survival times and censoring statuses. This dataset is included as an example data 'TCGA_full' in the '' package. This TCGA data was originally downloaded from the cBio Portal (http://www.cbioportal.org/) using the R package 'cgdsr' and here we consider z-scores for the mRNA expression data. The 'TCGA_full' is a list object with four elements, including the 'geneexpr' data frame of z-scores for the mRNA expression, the 't' vector of the survival time, the 'd' vector of the censoring status indicator, and the 'pathList' list of

the pathway information. The 'pathList' has 15 elements, each of which contains names of genes belonging to each pathway.

`This dataset can be loaded as follows:

```
library(INGRID)
data(TCGA_full)
TCGA_full$geneexpr[1:5,1:5]
```

```
##        ACSS2        GCK       PGK2       PDHB      PDHA2
## 1 -0.6521542 -1.1418104  0.5335282 -1.8419419 -0.1576928
## 2 -0.4066970 -0.8228400 -0.9112831  0.2184275 -1.1624208
## 3 -0.1197093 -0.6620549  0.5464014 -3.0176011 -1.0981827
## 4  1.5466998 -0.5097975 -0.6933681  0.4179899  0.8276257
## 5  0.1488380 -0.6588924  0.2642286 -1.3683970 -0.9166555
```

```
TCGA_full$t[1:5]
```

```
## [1] 43.89 40.97 49.12  2.00 46.59
```

```
TCGA_full$d[1:5]
```

```
## [1] 1 1 0 1 0
```

```
TCGA_full$pathList[1]
```

```
## $KEGG_HEDGEHOG_SIGNALING_PATHWAY
##  [1] "CSNK1A1L" "HHIP"     "PTCH2"    "GAS1"     "WNT3A"    "ZIC2"
##  [7] "WNT9B"    "WNT9A"    "LRP2"     "CSNK1G1"  "WNT2B"    "WNT11"
## [13] "WNT10B"   "IHH"      "SMO"      "WNT10A"   "WNT4"     "CSNK1G3"
## [19] "SHH"      "WNT1"     "CSNK1D"   "RAB23"    "CSNK1A1"  "CSNK1G2"
## [25] "CSNK1E"   "BMP8A"    "GSK3B"    "WNT7A"    "BTRC"     "WNT7B"
## [31] "WNT8A"    "WNT8B"    "WNT2"     "WNT3"     "PRKX"     "WNT5A"
## [37] "WNT6"     "FBXW11"   "STK36"    "WNT5B"    "GLI1"     "DHH"
## [43] "PRKACA"   "PRKACB"   "SUFU"     "BMP4"     "PRKACG"   "BMP2"
## [49] "GLI2"     "BMP7"     "GLI3"     "PTCH1"    "BMP8B"    "WNT16"
## [55] "BMP5"     "BMP6"
```

## 3 Gene Regrouping

Gene Regrouping step is to redefine the gene set membership of genes. First, the gene remains to be as a member if it is a core member of the gene set. A gene is defined as a "core member" of a gene set if it belongs to only that gene set. Second, if the gene maps to more than one gene sets, then this gene is re-assigned to one of the gene sets based on the k-medoids algorithm minimizing the binary distance between genes within cluster distance.

```
geneRegroup.results <- geneRegroup(TCGA_full$pathList)
geneRegroup.results

## Summary: Gene regrouping results (class: RegroupGene)
## ------------------------------------------------
## Gene sets before the gene regrouping
## List of 15
##  $ KEGG_HEDGEHOG_SIGNALING_PATHWAY           : chr [1:56] "CSNK1A1L" "HHIP
" "PTCH2" "GAS1" ...
##  $ KEGG_MTOR_SIGNALING_PATHWAY               : chr [1:52] "TSC2" "IGF1" "R
PS6KA6" "MTOR" ...
##  $ KEGG_NOTCH_SIGNALING_PATHWAY              : chr [1:47] "HES5" "DTX3" "N
OTCH4" "DTX3L" ...
##  $ KEGG_NUCLEOTIDE_EXCISION_REPAIR           : chr [1:44] "MNAT1" "POLE4"
"ERCC4" "POLE3" ...
##  $ KEGG_CELL_CYCLE                           : chr [1:128] "CDC16" "CDC7"
"CDC45" "GADD45B" ...
##  $ KEGG_CELL_ADHESION_MOLECULES_CAMS         : chr [1:134] "CDH5" "JAM3" "
CDH3" "NLGN3" ...
##  $ KEGG_JAK_STAT_SIGNALING_PATHWAY           : chr [1:155] "STAT3" "STAT4"
"STAT1" "STAT2" ...
##  $ KEGG_PHOSPHATIDYLINOSITOL_SIGNALING_SYSTEM: chr [1:76] "PLCB2" "CALM2"
"INPP1" "PLCB1" ...
##  $ KEGG_MAPK_SIGNALING_PATHWAY               : chr [1:267] "JUN" "MEF2C" "
ELK4" "ELK1" ...
##  $ KEGG_MISMATCH_REPAIR                      : chr [1:23] "MLH3" "POLD1" "
MLH1" "POLD2" ...
##  $ KEGG_APOPTOSIS                            : chr [1:88] "CASP10" "CASP9"
"CASP8" "CASP7" ...
##  $ KEGG_WNT_SIGNALING_PATHWAY                : chr [1:151] "JUN" "LRP5" "L
RP6" "PPP3R2" ...
##  $ KEGG_BASE_EXCISION_REPAIR                 : chr [1:35] "NEIL2" "MPG" "S
MUG1" "XRCC1" ...
##  $ KEGG_NON_HOMOLOGOUS_END_JOINING           : chr [1:14] "XRCC4" "MRE11A"
"POLL" "POLM" ...
##  $ KEGG_TGF_BETA_SIGNALING_PATHWAY           : chr [1:86] "TFDP1" "NOG" "T
NF" "GDF7" ...
## ------------------------------------------------
## Gene sets after the gene regrouping
## List of 17
##  $ gene_set_16                               : chr [1:97] "CSNK1A1L" "WNT3
A" "WNT9B" "WNT9A" ...
##  $ gene_set_17                               : chr [1:80] "PRKX" "PRKACA"
"PRKACB" "PRKACG" ...
##  $ KEGG_HEDGEHOG_SIGNALING_PATHWAY           : chr [1:20] "HHIP" "PTCH2" "
GAS1" "ZIC2" ...
##  $ KEGG_MTOR_SIGNALING_PATHWAY               : chr [1:32] "TSC2" "IGF1" "M
TOR" "EIF4B" ...
##  $ KEGG_NOTCH_SIGNALING_PATHWAY              : chr [1:37] "HES5" "DTX3" "N
OTCH4" "DTX3L" ...
```

```
##  $ KEGG_NUCLEOTIDE_EXCISION_REPAIR           : chr [1:22] "MNAT1" "ERCC4"
"ERCC3" "ERCC6" ...
##  $ KEGG_CELL_CYCLE                           : chr [1:91] "CDC16" "CDC7" "
CDC45" "DBF4" ...
##  $ KEGG_CELL_ADHESION_MOLECULES_CAMS         : chr [1:134] "CDH5" "JAM3" "
CDH3" "NLGN3" ...
##  $ KEGG_JAK_STAT_SIGNALING_PATHWAY           : chr [1:130] "STAT3" "STAT4"
"STAT1" "STAT2" ...
##  $ KEGG_PHOSPHATIDYLINOSITOL_SIGNALING_SYSTEM: chr [1:61] "CALM2" "INPP1"
"PLCD1" "CALM1" ...
##  $ KEGG_MAPK_SIGNALING_PATHWAY               : chr [1:201] "MEF2C" "ELK4"
"ELK1" "JUND" ...
##  $ KEGG_MISMATCH_REPAIR                      : chr [1:8] "MLH3" "MLH1" "MS
H2" "MSH3" ...
##  $ KEGG_APOPTOSIS                            : chr [1:42] "CASP10" "CASP9"
"CASP8" "CASP7" ...
##  $ KEGG_WNT_SIGNALING_PATHWAY                : chr [1:70] "LRP5" "LRP6" "S
FRP2" "SFRP1" ...
##  $ KEGG_BASE_EXCISION_REPAIR                 : chr [1:23] "NEIL2" "MPG" "S
MUG1" "XRCC1" ...
##  $ KEGG_NON_HOMOLOGOUS_END_JOINING           : chr [1:10] "XRCC4" "MRE11A"
"POLM" "NHEJ1" ...
##  $ KEGG_TGF_BETA_SIGNALING_PATHWAY           : chr [1:45] "NOG" "GDF7" "IN
HBB" "INHBC" ...
## ------------------------------------------------
## Comparison of the gene set before and after gene regrouping
##             s1 s2 s3 s4 s5 s6 s7 s8 s9 s10 s11 s12 s13 s14 s15
## gene_set_16 32  2 10 22 26  0  6  4  0  15   0  54  12   4  32
## gene_set_17  4 18  0  0 11  0 19 11 66   0  46  27   0   0   9
##
## where
##                                  gene_set_name
## s1              KEGG_HEDGEHOG_SIGNALING_PATHWAY
## s2                  KEGG_MTOR_SIGNALING_PATHWAY
## s3                 KEGG_NOTCH_SIGNALING_PATHWAY
## s4              KEGG_NUCLEOTIDE_EXCISION_REPAIR
## s5                               KEGG_CELL_CYCLE
## s6            KEGG_CELL_ADHESION_MOLECULES_CAMS
## s7              KEGG_JAK_STAT_SIGNALING_PATHWAY
## s8  KEGG_PHOSPHATIDYLINOSITOL_SIGNALING_SYSTEM
## s9                  KEGG_MAPK_SIGNALING_PATHWAY
## s10                        KEGG_MISMATCH_REPAIR
## s11                              KEGG_APOPTOSIS
## s12                  KEGG_WNT_SIGNALING_PATHWAY
## s13                   KEGG_BASE_EXCISION_REPAIR
## s14             KEGG_NON_HOMOLOGOUS_END_JOINING
## s15             KEGG_TGF_BETA_SIGNALING_PATHWAY
```

The table above shows the number of genes in a new gene set that is from each of the original gene sets. For instance, the new gene set “gene_set_16” has 32 gene from s1, which

is KEGG_HEDGEHOG_SIGNALING_PATHWAY, 2 genes from s2, which is KEGG_MTOR_SIGNALING_PATHWAY, and so on.

To refine the candidate set of genes, we first conduct a supervised pre-filtering by fitting a Cox regression model of the mRNA expression measure of each gene on the patient survival. Only the gene expressions associated with patient survival at p-values smaller than a pre-specified cut-off are included in the subsequent analysis. By default, p = 0.5 is used as cut-off point.

```
prefilter.results <- prefilter( data=TCGA_full$geneexpr
                                time=TCGA_full$t,
                                status=TCGA_full$d,
                                plist=geneRegroup.results@gset )
prefilter.results
```

```
## Summary: Pre-filtering results (class: Prefiltered)
## ---------------------------------------------------
## Number of genes before prefiltering: 4944
## Number of genes after prefiltering: 588
## ---------------------------------------------------
```

# 4 Gene Selection

In order to select key genes associated with patient survivals and effectively summarize them by taking into account correlation among them, we fit a sparse partial least squares (SPLS) Cox regression model of patient survivals on gene expression measurements for each pathway.

Using the object 'prefilter.results', gene-level analysis result can be generated with 'selectGene' function as follows.

```
gene.results <- selectGene(prefilter.results)
gene.results
```

```
## Summary: Gene-level analysis results (class: FitGene)
## ---------------------------------------------------
## Number of prefiltered genes: 588
## Number of selected genes: 58
## ---------------------------------------------------
```

```
coef(gene.results)
```

```
## $gene_set_16
##   gene_set coefficient_estimate
## 1    POLD3          -0.07362205
## 2    POLD2          -0.08861809
## 3     FEN1          -0.07550862
```

```
## 4  RPS6KB2          -0.07542887
## 5     BTRC           0.08027315
## 6     BMP4           0.07546504
##
## $gene_set_17
##   gene_set coefficient_estimate
## 1  RPS6KA2            0.2275404
## 2   PPP3CA            0.2302891
##
## $KEGG_HEDGEHOG_SIGNALING_PATHWAY
##   gene_set coefficient_estimate
## 1     GAS1            0.2149472
## 2  CSNK1G1            0.1606427
## 3  CSNK1G3            0.1681516
## 4   CSNK1D           -0.1455008
##
## $KEGG_MTOR_SIGNALING_PATHWAY
##   gene_set coefficient_estimate
## 1    PDPK1           0.09866496
## 2    VEGFA          -0.10331126
## 3    CAB39           0.09485034
##
## $KEGG_NOTCH_SIGNALING_PATHWAY
##   gene_set coefficient_estimate
## 1   NOTCH4           -0.1959747
##
## $KEGG_NUCLEOTIDE_EXCISION_REPAIR
##   gene_set coefficient_estimate
## 1   GTF2H4           -0.1623209
## 2     DDB2           -0.1579492
##
## $KEGG_CELL_CYCLE
##   gene_set coefficient_estimate
## 1     MCM3           -0.1615686
## 2  ANAPC11           -0.1475621
##
## $KEGG_CELL_ADHESION_MOLECULES_CAMS
##   gene_set coefficient_estimate
## 1  HLA-DOB           -0.2238179
##
## $KEGG_JAK_STAT_SIGNALING_PATHWAY
##   gene_set coefficient_estimate
## 1    IL21R           -0.1412788
## 2    SOCS5            0.1563292
##
## $KEGG_PHOSPHATIDYLINOSITOL_SIGNALING_SYSTEM
##   gene_set coefficient_estimate
## 1    PLCG1            0.1200187
##
## $KEGG_MAPK_SIGNALING_PATHWAY
```

```
##   gene_set coefficient_estimate
## 1  PLA2G2D           -0.2530461
##
## $KEGG_MISMATCH_REPAIR
##   gene_set coefficient_estimate
## 1    SSBP1           -0.1836033
## 2     MLH3           -0.1172121
##
## $KEGG_APOPTOSIS
##   gene_set coefficient_estimate
## 1    APAF1           0.08014538
## 2    IRAK2          -0.08003993
##
## $KEGG_WNT_SIGNALING_PATHWAY
##   gene_set coefficient_estimate
## 1      APC            0.2507789
##
## $KEGG_BASE_EXCISION_REPAIR
##   gene_set coefficient_estimate
## 1      UNG          -0.09162912
## 2    PARP4           0.08610522
## 3    APEX1          -0.09058163
## 4    MUTYH          -0.09184070
##
## $KEGG_NON_HOMOLOGOUS_END_JOINING
##   gene_set coefficient_estimate
## 1    RAD50            0.1191509
##
## $KEGG_TGF_BETA_SIGNALING_PATHWAY
##    gene_set coefficient_estimate
## 1     INHBC           0.01915876
## 2     INHBA           0.02819471
## 3    ACVR2A           0.02383607
## 4     AMHR2          -0.02285503
## 5    BMPR1A           0.03172572
## 6     THBS1           0.02923280
## 7    SMURF1           0.01971317
## 8      COMP           0.04560903
## 9     THBS4           0.02996526
## 10      ID1           0.03262260
## 11      ID4          -0.03161342
## 12    NODAL          -0.02837672
## 13     CHRD          -0.01828115
## 14      ID3           0.03327949
## 15    LTBP1           0.04072656
## 16   LEFTY1          -0.03924803
## 17   LEFTY2           0.02425622
## 18     GDF6           0.04489707
## 19      SP1           0.02523787
## 20   ZFYVE9           0.02635107
```

```
## 21  ZFYVE16           0.01842494
## 22    THBS2           0.03912128
## 23      DCN           0.03840006
```

The list of the SPLS regression coefficients of cancer-related genes can be generated using the function .

The function 'selectGene' has two main tuning parameters: 'eta' represents the sparsity tuning parameter and 'K' is the number of hidden (latent) components. Parameters can be chosen by ($v$-fold) cross-validation. Users can search the range for these parameters and the cross-validation procedure searches within these ranges. Note that '' should have a value between 0 and 1 while '' is integer-valued and can range between 1 and min{p, (v-1) n / v} , where $p$ is the number of genes and $n$ is the sample size. For example, if 10-fold cross-validation is used (default), 'K' should be smaller than min{p, 0.9n}. For the TCGA data, we set the number of folds as 5, 'K' as 5, and search the optimal '' in the range between 0.1 and 0.9.

# 5 Pathway Selection

Next, in order to identify a parsimonious set of pathways associated with patient survivals, we fit a LASSO-penalized Cox regression on latent components derived from all the pathways. Specifically, a pathway is selected if at least one of its latent components has non-zero LASSO coefficient estimate.

This approach has the following two strengths: First, the latent components generated from the SPLS step preserve pathway structure and also reflect correlation among genes and their association with survival outcomes. Second, this approach can potentially improve the stability of estimation in the subsequent analysis.

Using the 'gene.results', pathway-level analysis result can be generated with 'selectPath' function.

```
path.results <- selectPath(gene.results)
path.results
```

```
## Summary: Pathway-level analysis results (class: FitPath)
## -------------------------------------------------
## Number of all pathways: 17
## Number of selected pathways: 8
##
## List of selected pathways:
##  gene_set_16
##  gene_set_17
##  KEGG_HEDGEHOG_SIGNALING_PATHWAY
##  KEGG_NUCLEOTIDE_EXCISION_REPAIR
##  KEGG_CELL_CYCLE
```

```
##  KEGG_CELL_ADHESION_MOLECULES_CAMS
##  KEGG_MAPK_SIGNALING_PATHWAY
##  KEGG_MISMATCH_REPAIR
## ------------------------------------------------

coef(path.results)

##                                       gene_set coefficient_estimate
## 1                                   gene_set_16          0.007902704
## 2                                   gene_set_17          0.084730150
## 3              KEGG_HEDGEHOG_SIGNALING_PATHWAY          0.118576527
## 4              KEGG_HEDGEHOG_SIGNALING_PATHWAY          0.000000000
## 5                  KEGG_MTOR_SIGNALING_PATHWAY          0.000000000
## 6                 KEGG_NOTCH_SIGNALING_PATHWAY          0.000000000
## 7              KEGG_NUCLEOTIDE_EXCISION_REPAIR          0.025337081
## 8                               KEGG_CELL_CYCLE          0.013060080
## 9            KEGG_CELL_ADHESION_MOLECULES_CAMS          0.027034990
## 10             KEGG_JAK_STAT_SIGNALING_PATHWAY          0.000000000
## 11 KEGG_PHOSPHATIDYLINOSITOL_SIGNALING_SYSTEM          0.000000000
## 12                 KEGG_MAPK_SIGNALING_PATHWAY          0.048544818
## 13                        KEGG_MISMATCH_REPAIR          0.022330896
## 14                        KEGG_MISMATCH_REPAIR          0.000000000
## 15                             KEGG_APOPTOSIS          0.000000000
## 16                  KEGG_WNT_SIGNALING_PATHWAY          0.000000000
## 17                   KEGG_BASE_EXCISION_REPAIR          0.000000000
## 18             KEGG_NON_HOMOLOGOUS_END_JOINING          0.000000000
## 19             KEGG_TGF_BETA_SIGNALING_PATHWAY          0.000000000
```

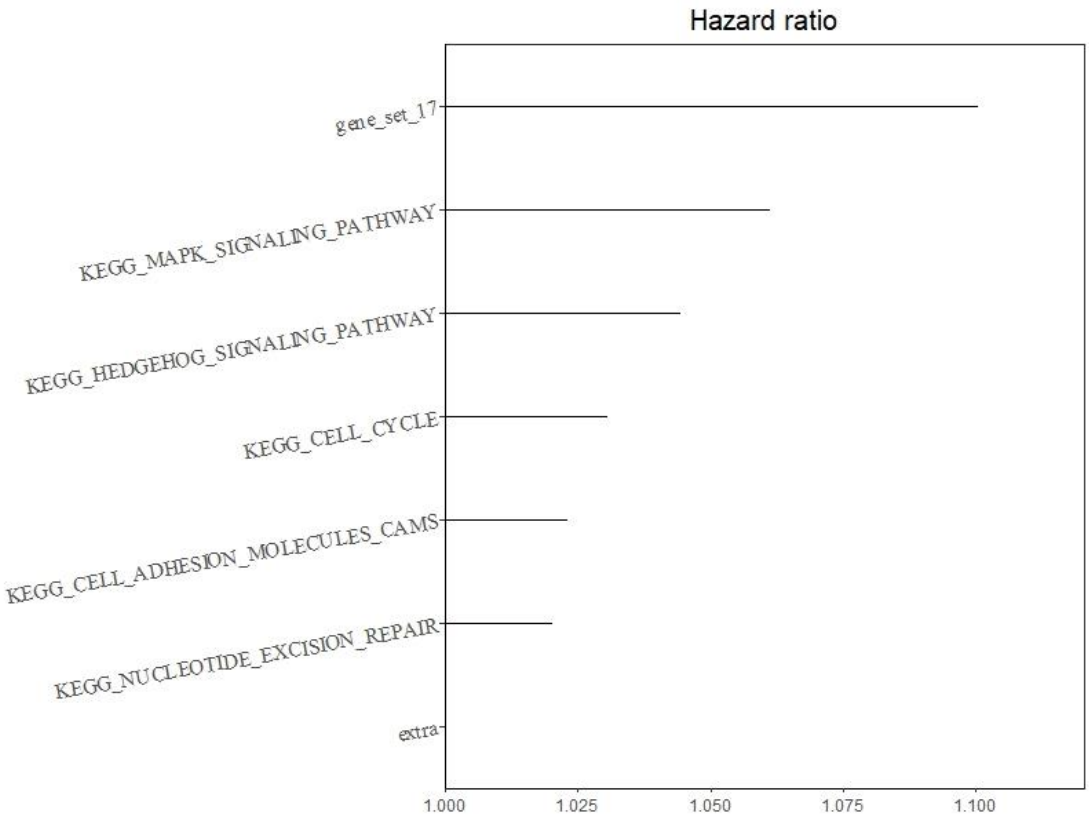
LASSO regression coefficients of cancer-related pathways can be generated using the function .

```
head(coef(path.results))

##                           gene_set coefficient_estimate
## 1                       gene_set_16          0.007902704
## 2                       gene_set_17          0.084730150
## 3 KEGG_HEDGEHOG_SIGNALING_PATHWAY          0.118576527
## 4 KEGG_HEDGEHOG_SIGNALING_PATHWAY          0.000000000
## 5     KEGG_MTOR_SIGNALING_PATHWAY          0.000000000
## 6    KEGG_NOTCH_SIGNALING_PATHWAY          0.000000000
```

Hazard ratio plot associated with each latent component in the selected pathways can be generated using the function  with the argument .

Figure below shows the hazard ratio (HR) associated with each latent component in the pathways selected by the INGRID. Based on the TCGA data, pathways with the largest effect on survival are `gene_set_17` and `KEGG_MAPK_SIGNALING_PATHWAY` gene sets.

```
plot(path.results, type="HR")
```




# 6 Risk Group Prediction

Risk group predictions can be made using the function

```
predicted <- predict(path.results)
```

The function '' returns a list with the following three elements: (1) risk.index : number of pathways with elevated activity for each patient; (2) riskcat: risk group prediction for each patient; (3) cuts: cut off to determine low, intermediate and high risk groups.

```
predicted
```

```
## $risk.index
##   [1] 4 8 2 7 3 3 7 2 5 2 8 3 5 3 4 6 1 3 5 8 5 3 2 3 4 4 6 6 2 6 5 7 4 7
6 5 6
##  [38] 6 2 1 1 1 2 2 6 8 5 2 4 1 4 2 1 0 4 2 7 6 4 6 3 4 3 7 4 1 6 5 3 4 4
5 3 6
##  [75] 6 6 3 3 4 1 2 4 1 2 4 5 3 2 2 3 1 0 2 3 2 3 3 2 4 3 4 5 0 2

##
```

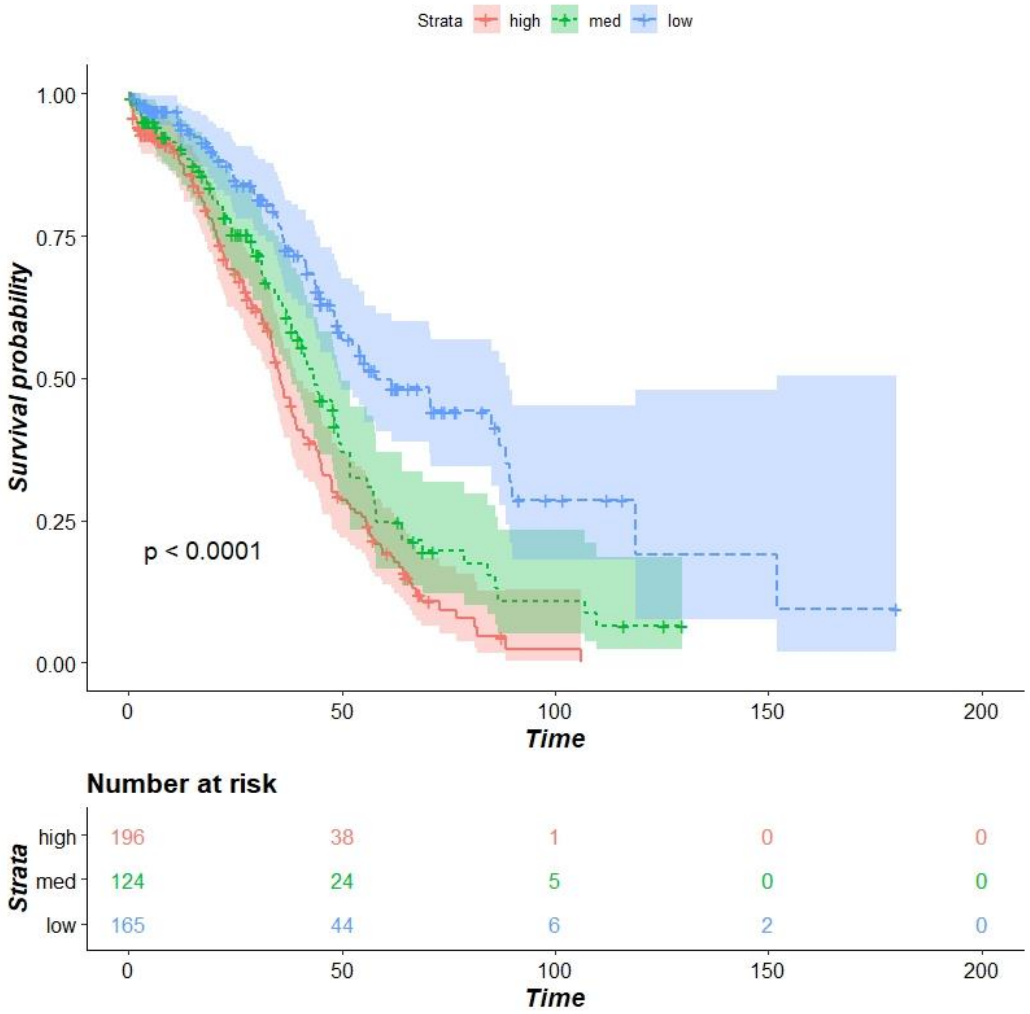
```
## $riskcat
##   [1] "med"  "high" "low"  "high" "low"  "low"  "high" "low"  "med"  "low"
##  [11] "high" "low"  "med"  "low"  "med"  "high" "low"  "low"  "med"  "high
"
##  [21] "med"  "low"  "low"  "low"  "med"  "med"  "high" "high" "low"  "high
"
##  [31] "med"  "high" "med"  "high" "high" "med"  "high" "high" "low"  "low"
##  [41] "low"  "low"  "low"  "low"  "high" "high" "med"  "low"  "med"  "low"
##  [51] "med"  "low"  "low"  "low"  "med"  "low"  "high" "high" "med"  "high
"
##  [61] "low"  "med"  "low"  "high" "med"  "low"  "high" "med"  "low"  "med"
##  [71] "med"  "med"  "low"  "high" "high" "high" "low"  "low"  "med"  "low"
##  [81] "low"  "med"  "low"  "low"  "med"  "med"  "low"  "low"  "low"  "low"
##  [91] "low"  "low"  "low"  "low"  "low"  "low"  "low"  "low"  "med"  "low"
## [101] "med"  "med"  "low"  "low"  "med"  "high" "low"  "med"  "low"  "med"
##
## $cuts
## [1] 3 6
##
## $time
##   [1]  43.89  40.97  49.12   2.00  46.59  18.50  11.86  65.44  63.01  48.7
2
##  [11]  21.55  63.93  49.25  32.49  44.29  33.64  76.48  71.42  56.51  19.9
7
##  [21]  38.34 118.99  29.04  84.14  55.36  27.86  28.62  64.59  36.20  31.8
3
##  [31]  47.67  45.30  29.14  57.69  17.94  35.51   9.95  38.41  89.26  41.5
6
##  [41]  57.53  61.83  55.33  30.46  17.97   3.02   8.54   6.11  35.12   4.5
0
##  [51] 112.29  71.06  63.05  39.72  20.89  12.45  25.89  81.80  18.46  34.1
7
##  [61]  32.85  14.65   6.67  35.38  36.89  19.97  20.89   6.21  24.57  44.5
2
##  [71]  39.56   2.73  12.39  31.14  15.18  17.77  55.13   2.46  11.17  11.5
3
##  [81]  34.04  32.23  31.01  28.55  21.39   8.41  38.90  36.70  28.68  30.3
2
##  [91]  19.55   8.71   7.79   8.11   8.28   8.18  43.79  35.25  35.28  13.6
0
## [101]  68.86 107.10  91.33  74.22  66.59  70.34  61.76  67.94  68.96  43.3
3
##
## $status
##   [1] 1 1 0 1 0 1 1 0 0 1 1 1 1 0 1 1 0 0 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1
1 1 1
##  [38] 1 1 1 0 0 0 0 0 0 0 0 1 1 0 0 0 0 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1
##  [75] 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1
```

# 7 Survival ROC

The predictive performance of "INGRID" method can be presented by Kaplan-Meier curves. Kaplan-Meier curves of predicted patient subgroups can be generated with plot() function with argument `type="KM".`

Figure below shows the Kaplan-Meier curves of predicted patient subgroups and indicates that the INGRID approach successfully separates out high, intermediate and low risk groups.

```
plot(path.results, type="KM")
```



# 8 Survival ROC

The predictive performance of INGRID method can be further evaluated based on area under the time dependent receiver operating curve (ROC). ROC plot can be generated using plot() function with argument `type="ROC"`.

```
plot(path.results, type="ROC")
```

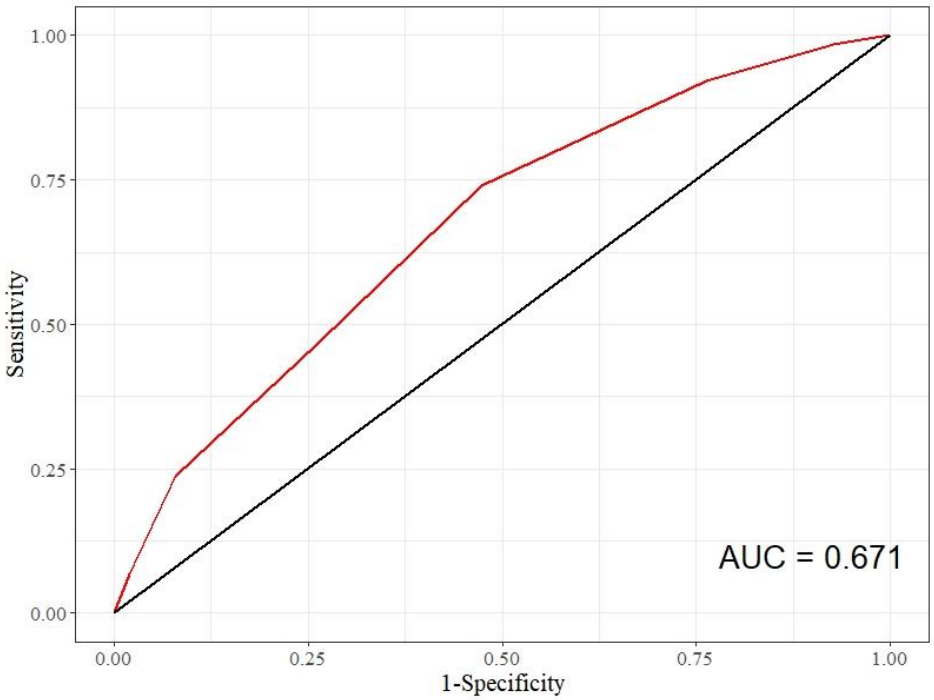



Figure above shows the ROC curves for survival, and for the TCGA data, the area under curve (AUC) associated with the INGRID approach was 0.671.

# 9 Bayes-InGRiD

The InGRiD approach supplies statistically rigorous and biologically interpretable inference tools for molecularly-defined cancer subgroup identification. Unfortunately, InGRiD is limited to only continuous data and lacks the ability of integrating multiple data types. To this end, a novel approach has been developed in AIM 1, namely Bayes-InGRiD, a Bayesian sparse latent factor model for the simultaneous identification of cancer subgroups and key molecular markers within a unified framework, based on the joint analysis of continuous, binary and count data. We developed an user-friendly function called "`bayesIngrid`" and provide it as a part of the R package "`INGRID`".

In this section, we used a cohort of high-grade serous ovarian cancer (HGSOC) patients from the TCGA project 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/).

Using the gene expression and copy number alteration measurements are the data input, these two data can be loaded as follow

```
data(TCGA_bayes)
list(data_cnv[1:5,1:5], data_mrna[1:5,1:5])
```

```
## [[1]]
##                    BMP2   BMP4  BMP5   BMP6   BMP7
## TCGA.13.1510.01 -0.016 -0.800 0.748  0.946 -0.037
## TCGA.13.1404.01 -0.944  0.079 0.074  0.968 -0.118
## TCGA.13.1506.01  0.536 -0.097 0.019 -0.609  0.799
## TCGA.23.1116.01  0.241 -0.287 0.276  0.286  0.358
## TCGA.13.1505.01  0.600 -0.288 0.505 -0.349  0.331
##
## [[2]]
##                    BMP2   BMP4   BMP5   BMP6   BMP7
## TCGA.13.1510.01 -0.570 -1.069 -0.709 -0.197 -0.402
## TCGA.13.1404.01 -0.459 -0.024 -0.343 -0.031 -0.835
## TCGA.13.1506.01 -0.417 -0.807  0.148 -0.240  1.138
## TCGA.23.1116.01 -0.159  0.896 -0.155 -0.143  0.490
## TCGA.13.1505.01  0.174  1.360 -0.230  1.466  0.132
```

In addition, the pathway information after "gene_regroup" step is in the data input, and it can be shown as follow

```
plist_regroup[1]
```

```
## $gene_set_16
##  [1] "BMP2"     "BMP4"     "BMP5"     "BMP6"     "BMP7"     "BMP8A"
##  [7] "BMP8B"    "BTRC"     "CSNK1A1"  "CSNK1A1L" "CSNK1E"   "FBXW11"
## [13] "GSK3B"    "WNT1"     "WNT10A"   "WNT10B"   "WNT11"    "WNT16"
## [19] "WNT2"     "WNT2B"    "WNT3"     "WNT3A"    "WNT4"     "WNT5A"
## [25] "WNT5B"    "WNT6"     "WNT7A"    "WNT7B"    "WNT8A"    "WNT8B"
## [31] "WNT9A"    "WNT9B"    "RPS6KB1"  "RPS6KB2"  "CREBBP"   "CTBP1"
## [37] "CTBP2"    "DVL1"     "DVL2"     "DVL3"     "EP300"    "HDAC1"
## [43] "HDAC2"    "PSEN1"    "CCNH"     "CDK7"     "LIG1"     "PCNA"
## [49] "POLD1"    "POLD2"    "POLD3"    "POLD4"    "POLE"     "POLE2"
## [55] "POLE3"    "POLE4"    "RBX1"     "RFC1"     "RFC2"     "RFC3"
## [61] "RFC4"     "RFC5"     "RPA1"     "RPA2"     "RPA3"     "CCND1"
## [67] "CCND2"    "CCND3"    "CDKN2B"   "CUL1"     "E2F4"     "E2F5"
## [73] "PRKDC"    "RBL1"     "RBL2"     "SKP1"     "SMAD2"    "SMAD3"
## [79] "SMAD4"    "TFDP1"    "IFNG"     "PLCB1"    "PLCB2"    "PLCB3"
## [85] "PLCB4"
```

```
str(plist_regroup)
```

```
## List of 14
##  $ gene_set_16                                : chr [1:85] "BMP2" "BMP4" "BMP5"
"BMP6" ...
##  $ gene_set_17                                : chr [1:74] "PRKACA" "PRKACB"
"PRKACG" "AKT1" ...
##  $ KEGG_HEDGEHOG_SIGNALING_PATHWAY            : chr [1:20] "CSNK1D" "CSNK1G1"
"CSNK1G2" "CSNK1G3" ...
##  $ KEGG_MTOR_SIGNALING_PATHWAY                : chr [1:31] "CAB39" "CAB39L" "DDIT4"
"EIF4B" ...
##  $ KEGG_NOTCH_SIGNALING_PATHWAY               : chr [1:37] "ADAM17" "APH1A" "CIR1"
"DLL1" ...
##  $ KEGG_NUCLEOTIDE_EXCISION_REPAIR            : chr [1:20] "CUL4A" "DDB1" "DDB2"
"ERCC1" ...
##  $ KEGG_CELL_CYCLE                            : chr [1:86] "ABL1" "ANAPC1"
"ANAPC10" "ANAPC11" ...
##  $ KEGG_CELL_ADHESION_MOLECULES_CAMS          : chr [1:122] "ALCAM" "CADM1" "CADM3"
"CD2" ...
##  $ KEGG_JAK_STAT_SIGNALING_PATHWAY            : chr [1:121] "CBL" "CBLB" "CBLC"
"CISH" ...
##  $ KEGG_PHOSPHATIDYLINOSITOL_SIGNALING_SYSTEM: chr [1:59] "CALM1" "CALM2" "CALM
```

```
3" "CALML3" ...
##  $ KEGG_MAPK_SIGNALING_PATHWAY               : chr [1:192] "ARRB1" "ARRB2" "ATF2"
"ATF4" ...
##  $ KEGG_MISMATCH_REPAIR                      : chr [1:8] "EXO1" "MLH1" "MLH3" "MSH
2" ...
##  $ KEGG_APOPTOSIS                            : chr [1:39] "APAF1" "BAD" "BAX" "BCL
2" ...
##  $ KEGG_WNT_SIGNALING_PATHWAY                : chr [1:21] "APC" "APC2" "AXIN1" "AX
IN2" ...
```

In the "`bayesIngrid`" function, user need to specify the number of data, data input, types of the data, and the type of the analysis, pathway level data if pathway analysis is chosen, the latent component number k, the number of MCMC iterations N, and the number of burnin samples in the MCMC.

An example running the "`bayesIngrid`" function can be shown as follow

```
bayesingrid.result=bayesIngrid( ndata = 2,
                                data = list(data_cnv,data_mrna),
                                datatype =c("continuous","continuous"),
                                analysistype = "pathway",
                                pathwaydata = plist_regroup,
                                k = 14,
                                N = 10000,
                                burin = 5000)
```

Note that user must specify the parameters as follow.

| | |
|---|---|
| n | = data number of datas |
| data | = list of N by P dataframes, where N is number of observations, P is number of genes |
| datatype | = type of the datas for the input |
| analysistype | = if analysistype is "pathway", pathway-level analysis,<br>if analysistype is "gene", gene-level analysis |
| pathwaydata | = list of pathway information if analysistype is "pathway" |
| k | = number of latent components. k is the number of pathways, if analysistype = "pathway". |
| N | = number of iterations for MCMC. |
| burnin | = number of iterations discarded as burnin in the MCMC. |

The following warning would show if the input is not correct.

| | |
|---|---|
| if data is not a list | : print "datasets must be a list" |
| if there is missing value in data | : print "data cannot have NAs. please exclude or impute missing values." |
| if k is missing in the input | : print "must specify k (gene-level: # of sample clusters; pathway-level: # of pathways )" |
| if datatype is missing in the input | : print "must specify datatype vector of (continuous, binary or count)" |
| if analysistype is missing in the input | : print "must specify analysistype: "gene" or "pathway" |

User can use "`fitBayes`" function to show the "`bayesIngrid`" function result table

Pathways are ranked based on 'Pathway coefficient', which are their weighted averages of factor loadings of selected genes. 'selected' refers to the number of genes selected in each pathway, total number of genes in each pathway are also included within parenthesis in the column 'selected'. Genes that rank top three in coefficient estimates would be shown in column 'Top three genes'

```
bayes.results = fitBayes( data = Bayes_result )
```

```
## Summary: Bayes-InGRiD Pathway-level analysis results (class: FitPath)
## -------------------------------------------------
## data1 result:
##                                       select selected_average top_3_genes_1
## CELL_ADHESION_MOLECULES_CAMS          76(122)            0.524      HLA.DRB1
## CELL_CYCLE                             69(86)            0.506          ORC1
## NUCLEOTIDE_EXCISION_REPAIR              6(20)            0.493         CUL4A
## MAPK_SIGNALING_PATHWAY                129(192)            0.431          FGF4
## MISMATCH_REPAIR                         4(8)            0.367          MSH6
## APOPTOSIS                              7(39)            0.353         BIRC2
## WNT_SIGNALING_PATHWAY                  5(21)            0.321           APC
## MTOR_SIGNALING_PATHWAY                 8(31)            0.301        PRKAA1
## NOTCH_SIGNALING_PATHWAY               11(37)            0.280          DLL3
## WNT&HEDGEHOG                          11(85)            0.237          PCNA
## MAPK&APOPTOSIS                        20(74)            0.224        PPP3CA
## JAK_STAT_SIGNALING_PATHWAY            18(121)            0.214          JAK2
## PHOSPHATIDYLINOSITOL_SIGNALING_SYSTEM  14(59)            0.196         ITPR2
## HEDGEHOG_SIGNALING_PATHWAY             9(20)            0.187          SUFU
## BASE_EXCISION_REPAIR                   0(25)               NA          <NA>
## NON_HOMOLOGOUS_END_JOINING             0(13)               NA          <NA>
## TGF_BETA_SIGNALING                     0(51)               NA          <NA>
##                                     top_3_genes_2 top_3_genes_3
## CELL_ADHESION_MOLECULES_CAMS             HLA.DPB1      HLA.DPA1
## CELL_CYCLE                                 CDC25C          PLK1
## NUCLEOTIDE_EXCISION_REPAIR                  ERCC5         ERCC1
## MAPK_SIGNALING_PATHWAY                    PLA2G12B        CACNG3
## MISMATCH_REPAIR                               MSH2          MSH3
## APOPTOSIS                                  ENDOD1         BIRC3
## WNT_SIGNALING_PATHWAY                     CTNNBIP1        CSNK2B
## MTOR_SIGNALING_PATHWAY                     RICTOR         RPTOR
## NOTCH_SIGNALING_PATHWAY                    PSENEN         PSEN2
## WNT&HEDGEHOG                                PLCB1         PLCB4
## MAPK&APOPTOSIS                              NFKB1         CASP3
## JAK_STAT_SIGNALING_PATHWAY                  IFNB1          IFNE
## PHOSPHATIDYLINOSITOL_SIGNALING_SYSTEM        DGKH       PIP5K1A
## HEDGEHOG_SIGNALING_PATHWAY                   GAS1         STK36
```

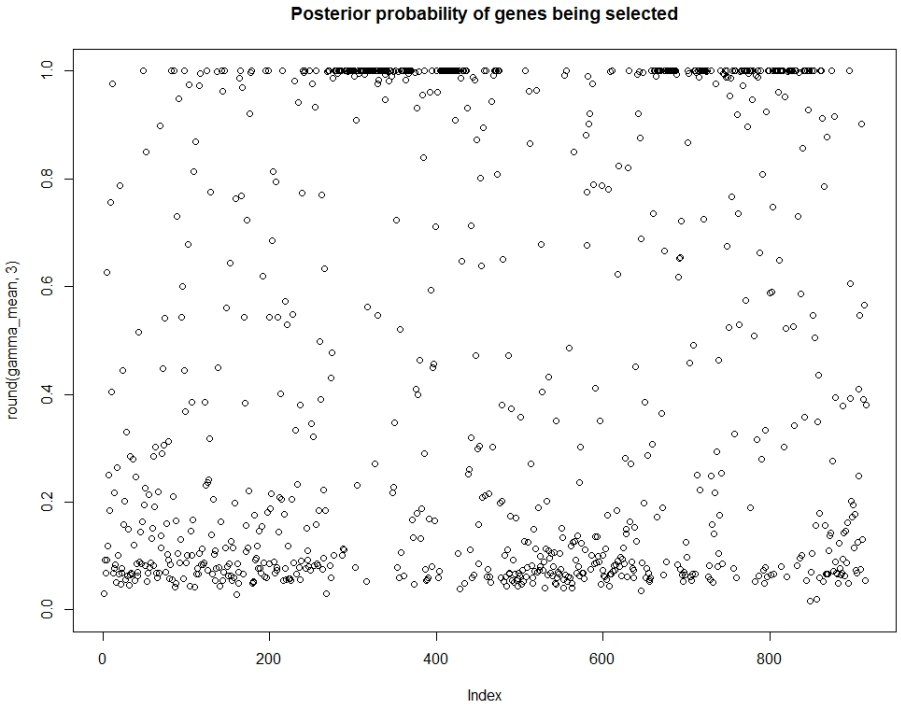
```
## BASE_EXCISION_REPAIR                              <NA>          <NA>
## NON_HOMOLOGOUS_END_JOINING                        <NA>          <NA>
## TGF_BETA_SIGNALING                                <NA>          <NA>
## ------------------------------------------------
## data2 result:
##                                          select selected_average top_3_genes_1
## MISMATCH_REPAIR                            2(8)            0.574          MSH2
## JAK_STAT_SIGNALING_PATHWAY               24(121)            0.438         IFNA6
## APOPTOSIS                                  6(39)            0.354         BIRC2
## MTOR_SIGNALING_PATHWAY                     3(31)            0.352        RICTOR
## WNT_SIGNALING_PATHWAY                      2(21)            0.284           APC
## PHOSPHATIDYLINOSITOL_SIGNALING_SYSTEM    13(59)             0.250         PLCZ1
## NUCLEOTIDE_EXCISION_REPAIR                 5(20)            0.209         CUL4A
## WNT&HEDGEHOG                              20(85)            0.205         PLCB4
## HEDGEHOG_SIGNALING_PATHWAY                 5(20)            0.191         PTCH1
## NOTCH_SIGNALING_PATHWAY                   23(37)            0.182          DLL3
## MAPK&APOPTOSIS                            27(74)            0.127         NFKB1
## MAPK_SIGNALING_PATHWAY                    8(192)            0.126       RASGRP4
## CELL_CYCLE                                 6(86)            0.103         CCNE1
## CELL_ADHESION_MOLECULES_CAMS             22(122)            0.079          CDH4
## BASE_EXCISION_REPAIR                       0(25)               NA          <NA>
## NON_HOMOLOGOUS_END_JOINING                 0(13)               NA          <NA>
## TGF_BETA_SIGNALING                         0(51)               NA          <NA>
##                                         top_3_genes_2 top_3_genes_3
## MISMATCH_REPAIR                              MSH6          <NA>
## JAK_STAT_SIGNALING_PATHWAY                  IFNA2          IFNE
## APOPTOSIS                                   BIRC3        ENDOD1
## MTOR_SIGNALING_PATHWAY                     PRKAA1          MTOR
## WNT_SIGNALING_PATHWAY                      CAMK2A          <NA>
## PHOSPHATIDYLINOSITOL_SIGNALING_SYSTEM     PIK3C2G         ITPR2
## NUCLEOTIDE_EXCISION_REPAIR                  ERCC5         ERCC2
## WNT&HEDGEHOG                                PLCB1          BMP2
## HEDGEHOG_SIGNALING_PATHWAY                   GAS1          SUFU
## NOTCH_SIGNALING_PATHWAY                    PSENEN         NUMBL
## MAPK&APOPTOSIS                             PPP3CA        MAPK10
## MAPK_SIGNALING_PATHWAY                     MAP4K1         PTPRR
## CELL_CYCLE                                   CDK1          FZR1
## CELL_ADHESION_MOLECULES_CAMS               CLDN23         ICAM1
## BASE_EXCISION_REPAIR                         <NA>          <NA>
## NON_HOMOLOGOUS_END_JOINING                   <NA>          <NA>
## TGF_BETA_SIGNALING                           <NA>          <NA>
```

User can use extract the posterior probability of a gene being selected by calling the parameter `gamma` of the "`fitbayes`" output. In addition, the latent factor score can also be acquired from the same output.

```
bayes.results$gamma[[1]][1:20]

##  [1] 1.00000000 0.03448565 0.10406705 0.11692439 1.00000000 0.04819840
##  [7] 0.06099534 0.05894284 0.17149143 0.17808520 0.53551773 0.04860600
## [13] 0.85819443 0.01922394 0.89441458 0.02019641 0.05326754 0.24004749
## [19] 0.52111706 0.01709189

bayes.results$gamma[[2]][1:20]

##  [1] 1.00000000 0.03448565 0.10406705 0.11692439 1.00000000 0.04819840
##  [7] 0.06099534 0.05894284 0.17149143 0.17808520 0.53551773 0.04860600
## [13] 0.85819443 0.01922394 0.89441458 0.02019641 0.05326754 0.24004749
## [19] 0.52111706 0.01709189

bayes.results$score[1:5,1:5]

##            [,1]        [,2]       [,3]       [,4]       [,5]
## [1,] -0.5140492 -0.95313526 -0.5133698 -0.6521319 -1.3382263
## [2,] -1.9191046 -1.41925462 -1.1517966  1.4444156 -1.8951507
## [3,]  0.3793968  0.82842823 -0.2786184  1.0162063  2.3592605
## [4,] -0.1454551  0.43700823 -0.1865393  0.2094766  1.8979575
## [5,]  0.4645647 -0.05658318 -0.5068286  0.1467986 -0.1980622
```

Lastly, the plot of the posterior probability of a gene being selected can be shown using the `plot()` function with `type = "bayes"`

```
plot(bayes.results, type = "bayes")
```


Posterior probability of genes being selected
round(gamma_mean, 3)
0.0
0.2
0.4
0.6
0.8
1.0
0
200
400
600
800
Index

# 5. SPECIFIC AIM 3

For Aim 3, our goal is to investigate various variable selection approaches in compositional data setting, infer key immune subtypes associated by applying stepwise pairwise log-ratio procedure on immune cellular fractions data, and identify key species in the microbiome data by using zero-inflated Wilcoxon rank sum test for Colorectal Adenocarcinoma.

## 5.1 Introduction

Nowadays, in order to find ways to advance research on immunotherapy for cancer treatment, it is of crucial importance to understand the tumor-immune interactions [49]. Many researchers are dedicated to study infiltration of tumour-associated immune cells and try to identify key immune subtypes and make meaningful biological interpretation of the results. For example, the Immune Landscape of Cancer [50] is a landmark that carried out a large-scale immunogenomic analysis of more than 10,000 tumors of 33 various cancer types based on TCGA data. Emergence of such new data type motivates development of an approach that identifies patient subgroups and key immune cell types simultaneously in the compositional data analysis framework setting. In the last decade, many studies have attempted to deconvolve gene expres-

sion data into their constituent cellular fractions. More specifically, these approaches solve the problem as a system of equations that describe the gene expression of a sample as the weighted sum of the expression profiles of the cell types. For example, CIBERSORT estimates proportions of cell types from gene expression profiles using Support Vector Regression (SVR) [51]. There are some other methods focusing on deconvolution of microarray data obtained from normal tissue into cell-type-specific profiles, by calculating enrichment score. [52] [53] [54] [55]. These methods take advantage of the differences in transcriptome properties of distinct cell types [56]. In this aim, one of our interest is to identify key immune subtypes using cellular fractions data induced from Colorectal Adenocarcinoma TCGA PanCancer study processed by CIBERSORT [9].

When it comes to Compositional Data Analysis (CoDA), the most common compositional replacement is to covert the data to ratios using the centred log-ratio (clr) [7] transformation. A drawback of this method is that the variable selection applied on the clr-transformed variables makes interpretation challenging. Since our goal is to identify key cell types, it is crucial to address the issue of interpretability for variable selection. Hron and others [24] purposed a covariance-based stepwise procedure for variable selection in 2013. In this procedure, variable selection is achieved by eliminating the variable whose variance of the corresponding clr variable is the smallest, calculating normalized variance of transformed variables of the remaining

sample space, and repeating the procedure until a purposed test statistics reach a pre-specified threshold. Another variable selection approach is proposed by Greenacre [8] where all pairwise ratios of parts are considered for key marker identification. A smaller set of ratios can be chosen to explain as much variability as required to reveal the underlying structure of the data. For the purpose of identifying key immune cell subtypes, we implement the stepwise pairwise log-ratio developed by Greenacre [8] for variable selection.

While the covariance-based stepwise procedure and pairwise log-ratio stepwise approach are efficient variable selection approach for low-dimensional compositional data. However, when it comes to high-dimensional zero-inflated microbiome datasets generated by high throughput sequencing (HTS) technology, these two approach are no longer applicable. Microbiome datasets generated by HTS are compositional because the total number of sequenced reads depends on the capacity of the instrument. It is crucial to recognize several key features for Microbiome data: 1) data are strictly positive or zero, never negative; 2) each count is not compositional itself, but the share out of counts is; 3) data often present excessive zeros, which may be due to undersampling, high heterogeneity, or real absence.

With high-dimensional and zero-inflated nature of microbiome data, much more care needs to be devoted to a reasonable coordinate representation and selection of methods to be used in the compositional data analysis. To deal with the large proportion of zeros in the microbiome data, many im-

putation approaches have emerged in recent years. The R package zCompositions [25] provides several methods for the multivariate imputation of zeros and non-detects in compositional data. These approaches are proposed based on an appropriate coordinate representation of the compositional data in the usual Euclidean geometry. The imputation is achieved by using iterative approaches where EM algorithm [26], Markov Chain Monte Carlo (MCMC) [27] or multiple imputation are utilized. However, in some extreme cases, we could face microbiome data where the majority of the data are zeros and the number of variables could be hundreds. The imputation approaches are not applicable given overwhelming amount of zeros in the data. In addition, it is important to realize many assumption of multivariate approach that was developed based on compositional data setting are not fit given the high dimensionality. Alternatively, we could consider the data as univariate and apply zero-inflated Wilcoxon test [10] for variable selection.

To investigate variable selection in compositional data analysis in immunology data and microbiome data, we will apply stepwise pairwise logratio for key cell type identification using cellular fractions data induced from Colorectal Adenocarcinoma TCGA Pan-Cancer study processed by CIBERSORT. As for the microbiome data, we take into account key aspects of the data including large proportion of zeros and high dimensionality and apply zero-inflated Wilcoxon test to identify key species in the metagenomic data of six cross-sectional studies of colorectal cancer.

## 5.2 Methods

In the methods section, we first look at the variable selection approach that can be applied to low-dimensional immunology data in the compositional data analysis. Next, we will discuss the variable selection for high-dimensional zero-inflated microbiome data.

For low-dimensional immunology data in the compositional data setting, we apply the stepwise pairwise log-ratio approach [8] for key cell type identification using cellular fractions data. We consider the compositional data that are made up of the relative proportions of a whole and can be represented in the simplex of $d$ parts:

$$S^m := \{x = (x_1, \cdots, x_m) \in R^m | \sum_{i=1}^{m} x_i = 1, x_i > 0, \forall i\}$$

The basic measure of variability of a random composition $x = (x_1, \cdots, x_m)$ is the variation matrix [7], defined as

$$\mathbf{T} = \left\{ var\left( ln\frac{x_i}{x_j} \right) \right\}_{i,j=1}^{M}$$

Each element in the variation matrix defines the variability of the log-ratio $ln\frac{x_i}{x_j}$ : The log-ratio tends to be a constant if the value of the variance is small. Total variance is defined as the sum of the elements of the variation matrix,

where

$$totvar(x) = \frac{1}{2m}\sum_{i=1}^{M}\sum_{j=1}^{M} var\left(ln\frac{x_i}{x_j}\right)$$

Total variance presents the total variability of the compositional data set. It can also be written as

$$totvar(x) = \sum_{j<j'}\frac{1}{n}\sum_{i}(z_{i,jj'} - \bar{z}_{jj'})^2 \tag{5.1}$$

where $z_{i,jj'} = log\frac{x_{ij}}{x_{ij'}}$ and $\bar{z}_{jj'} = \frac{1}{n}\sum_i z_{i,jj'}$, the notation $\sum_{j<j'}$ indicates the double summation over all $\frac{1}{2}m(m-1)$ unique pairs of the index. It is important to note that the sum of all the $\frac{1}{2}m(m-1)$ logratio variances, which is the half triangle of this matrix, can measure of total variability for the composition.

Furthermore, the total variance can derived that in terms of all the pairwise-squared differences between the logratios.

$$\begin{aligned} totvar(x) &= \frac{1}{n^2}\sum_{i<i'}\sum_{j<j'}(z_{i,jj'} - z_{i',jj'})^2 \\ &= \frac{1}{n^2}\sum_{i<i'}\sum_{j<j'}(log\frac{x_{ij}}{x_{ij'}} - log\frac{x_{i'j}}{x_{i'j'}})^2 \\ &= \frac{1}{n^2}\sum_{i<i'}\sum_{j<j'}(log\frac{x_{ij}}{x_{ij'}}\frac{x_{i'j}}{x_{i'j'}})^2 \end{aligned} \tag{5.2}$$

We define ALR:k as a set of ALR logratios with respect to a specified part k,

where the ith sample contains $(m-1)$ values. Specifically,

$$ALR: k(i) = log\frac{x_{ij}}{x_{ik}}, j = 1, \cdots, m, j \neq k.$$

For compositional data, it sums to 1 for each row, where the row weights are $r_i = 1/n$, and the weights are constant for all samples. We define the column weights as $c_1, c_2, \cdots, c_m$, where $c_j$ is $j$th part mean, notice that the column weightes also sum to 1. As a result, variables with relatively low means have high variance in the corresponding logratios, and these variables will be down-weighted. It follows

$$totvar(x) = \sum_{j<j'} c_j c_{j'} \sum_i r_i (z_{i,jj'} - \bar{z}_{jj'})^2 \tag{5.3}$$

where $z_{i,jj'} = log\frac{x_{ij}}{x_{ij'}}$ and $\bar{z}_{jj'} = \sum_i r_i z_{i,jj'}$

$$\begin{aligned} totvar(x) &= \sum_{i<i'} r_i r_{i'} \sum_{j<j'} c_j c_{j'} (z_{i,jj'} - z_{i',jj'})^2 \\ &= \sum_{i<i'} \sum_{j<j'} r_i r_{i'} c_j c_{j'} (log\frac{x_{ij}}{x_{ij'}} - log\frac{x_{i'j}}{x_{i'j'}})^2 \\ &= \sum_{i<i'} \sum_{j<j'} r_i r_{i'} c_j c_{j'} (log\frac{x_{ij}}{x_{ij'}}\frac{x_{i'j}}{x_{i'j'}})^2 \end{aligned} \tag{5.4}$$

The weighted structure presents a perfect symmetric formulation in terms of rows and columns. If $c_j = 1/m$ for all $j$, the total variance can be defined as

"unweighted". Meanwhile, if $r_i = 1/n$ for all $i$, it would be same as equation 5.1 divided by $m^2$. The logratio distance between two samples $i$ and $i'$ is

$$\begin{aligned} d_{ii'} &= \sqrt{\sum_{j<j'} c_j c_{j'} (log\frac{x_{ij}}{x_{ij'}} - log\frac{x_{i'j}}{x_{i'j'}})^2} \\ &= \sqrt{\sum_{j<j'} c_j c_{j'} (log\frac{x_{ij}}{x_{ij'}}\frac{x_{i'j}}{x_{i'j'}})^2} \end{aligned} \tag{5.5}$$

so that the logratio variance can also be written as the weighted sum of squares of all the inter-sample distances.

$$totvar(x) = \sum_{i<i'} r_i r_{i'} d_{ii'}^2 \tag{5.6}$$

After the logratio variance are calculated, Redundancy analysis (RDA) [57] were used to measure how much of the total variance is explained by a subset of logratios of certain explanatory variables. RDA is a form of multivariate regression. If a variable is correlated with many of the other logratios, the explained variance of the corresponding variable will be high. In addition, Procrustes analysis [58] was applied to decide how close their multivariate structures are, more specifically, it decide how close a configuration based on a subset of logratios is to the configuration based on all the logratios. Procrustes correlation is the measurement for the matching of two configurations, and it was achieved by matching one to the other by rotation, translation and rescaling.

The stepwise procedure variable selection in the compositional data can proceed as follow:

1. Calculate all the pairwise logratios.

2. Select the one with the highest percentage of variance explained. This ratio is then fixed as the first logratio.

3. The second best logratio in combination with the first is sought, then fixed, and so on.

4. Repeat step 1 to 3 until variance explained reach 100%.

It must take into account that one should choose ratios that are independent of the ones already chosen: for example, if A/B and B/C have already been selected, then $A/C$ is no longer a candidate for selection, since it depends on the others: $A/C = A/B \times B/C$ . On the log scale, $log(A) - log(C)$ is the sum of, and thus linearly dependent on, $log(A) - log(B)$ and $log(B) - log(C)$. Since the dimensionality of an $m$-part compositional data set is $m-1$, and all the parts will have appeared in at least one logratio after $m - 1$ steps of the above procedure, the variance explained will be 100%.

As for microbiome data, we must take into account key issues of data including the high dimensionality and the large proportion of zeros. First, it is not a good fit to apply multivariate approach on compositional data setting, since the assumption of multinomial distribution is often violated due to the high dimensionality. In alternative, to address the challenge of zero inflation, we could consider the data as univariate and apply zero-inflated

Wilcoxon test [10] for variable selection.

Zero-inflated Wilcoxon test was first proposed in 2010 by Hallstrom [10] and it is further modified by Wang and others. [35]. The theory of the zero-inflated Wilcoxon rank sum test is as follow. We consider $2N$ patients in a randomized study, where $N$ patients are assigned to the treatment group $T_1$ and control group $T_2$, respectively . We define $f_1$ and $f_2$ as the distributions of the non-zero values under $T_1$ and $T_2$. Let $n_i$ be the number of non-zero scores in each group, $n = max(n1, n2)$ and $m = |n1 - n2|$ . Without loss of generosity, we assume there are no ties among the 2nm non-zero scores. In order to compute the rank-sums, we assign rank 1 to the highest score, rank 2 to the second highest score and so on. Hence, we have 2(Nn)+m zeros tied at the highest rank.

The zero-inflated Wilcoxon rank sum test will be based on the $2n$ observations remaining when $N-n$ observations with zero score have been removed from each group. Let $r$ be the sum of the ranks of the observations in group 1 among all 2n observations. Let r0 be the sum of the ranks of the non-zero scores of group 1. Then

$$r = \begin{cases} r_0 + m\dfrac{(2n - m + 1 + 2n)}{2}, & \text{if } n_1 \leq n_2 \\ r_0, & \text{if } n_1 \geq n_2 \end{cases} \tag{5.7}$$

and under the null $f_1 = f_2$, the Wilcoxon rank-sum statistic, $s = r - N(2N +$

$1)/2$, satisfies

$$E(s = r - n(2n+1)/2|n1, n2) = \begin{cases} mn/2, & \text{if } n_1 \leq n_2 \\ -mn/2 & \text{if } n_1 \geq n_2 \end{cases} \tag{5.8}$$

Then

$$\begin{aligned} Var(s|n_1, n_2) &= Var(r|n_1, n_2) \\ &= Var(r_o|n_1, n_2) \\ &= n(nm)(2nm+1)/12 \\ &= n^3/6 + nm^2/12 - mn^2/4 + n^2/12 - nm/12 \end{aligned} \tag{5.9}$$

Let $\mu_{i,j} = E((n/N)^i(m/N)^j)$. Then

$$E(Var(s|n_1, n_2)) = N^3(\mu_{3,0}/6 + \mu_{1,2}/12 - \mu_{2,1}/4) + N^2(\mu_{2,0} - \mu_{1,1})/12 \tag{5.10}$$

Under the null hypothesis, it is equally likely that $n_1$ is less than or greater than $n_2$, so $E(s) = 0$ and $Var(E(s|n_1, n_2) = E((mn/2)2) = N^4\mu_{2,2}/4$. Since $Var(s) = Var(E(s|n_1, n_2) + E(Var(s|n_1, n_2))$, it follows

$$Var(s) = N^4\mu_{2,2}/4 + N3(\mu_{3,0}/6 + \mu_{1,2}/12 - \mu_{2,1}/4) + N^2(\mu_{2,0} - \mu_{1,1})/12$$

It is defined that the zero-inflated Wilcoxon rank sum test by $W = s/\sqrt{Var(s)}$.

Next, key species were identified by implementing zero-inflated Wilcoxon rank sum test between treatment and control group for each species and adjusting for the multiple testing, e.g. using the Benjamini-Hochberg procedure.

## 5.3 Simulation

For the simulation section, based on the different nature of immunology data and microbiome data, we apply distinct strategies for the simulation study. First, for the simulation of low-dimensional immunology data, we plan to implement small scale simulation study focuses on the stepwise pairwise log-ratio approach. More specifically, we aim to identify the signal variables from noise, and recover the underlying correlation structure among the selected variables.

The cellular fractions data induced from Colorectal Adenocarcinoma TCGA Pan-Cancer study that we used in the real data analysis contain 254 patients, including 58 African American and 196 European American. To mimic this, We examined six sample sizes, namely $n = 200$ with 2 groups, and we generate equal number of samples in the two groups. In addition, the immunology data we use in the real data analysis contains 9 celltypes. To mimic this, we will consider three different composition lengths, $D = (3, 6, 9)$, with varying numbers of signal features.

For the given number of components and sample sizes, we first generated data from a $Dirichlet(\alpha)$ distribution, where $Dirichlet(\alpha)$ is a family of continuous multivariate probability distributions parameterized by a matrix $\alpha$. We generated the $\alpha$ matrix from a truncated multivariate normal distribution, $MVN(\mu, \Sigma)$. The mean of the $\alpha$ matrix $\mu$ is set for the underlying clus-

tering structure, more specifically, we choose different mean structure for the two groups. Next, in the multivariate normal distribution, the covariance matrix $\Sigma$ is a diagonal matrix, by setting different level for diagonal elements of the covariance matrix, we can set different level of signals. The larger the variance is, the stronger the signal is. In addition, we can control the correlation of the variables by setting the elements of the covariance matrix.

We evaluate the performance of the stepwise pairwise log-ratio approach by checking its detection power, which is defined as ability to separate out signal features from noise features. In addition, we will check the performance of stepwise pairwise log-ratio approach on whether it can detect features that are highly correlated to each other, by looking at the numbers on the edges shown in the stepwise procedure, where large number of links indicate more variation for that feature. Finally, we will evaluate that if the features in the stepwise pairwise log-ratio approach can preserve the underlying clustering structure. In particular, we will apply PCA on the features selected from the simulation, and apply K-means algorithm for the sample clustering. Sensitivity and specificity will be used to evaluate the performance.

For example, if we set sample size $n = 200$, and generate $\alpha$ matrix using truncated multivariate normal where $\mu = (50, 50, 50, 50, 50, 50, 50, 50)$ and the covariance matrix

$$\boldsymbol{\Sigma} = \begin{array}{c} \begin{matrix} \text{V1} & \text{V2} & \text{V3} & \text{V4} & \text{V5} & \text{V6} & \text{V7} & \text{V8} \end{matrix} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 100 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 100 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 100 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 100 \end{bmatrix} \end{array}$$

We have the first 10 rows of $\alpha$ matrix shown in Table 5.1. Next, we generated simulation data from a $Dirichlet(\alpha)$ distribution, and the first 10 samples of the compositional simulation data is shown in Table 5.2:

| v1 | v2 | v3 | v4 | v5 | v6 | v7 | v8 |
|---|---|---|---|---|---|---|---|
| 49.038 | 49.707 | 52.588 | 38.479 | 50.196 | 50.030 | 50.854 | 61.166 |
| 48.781 | 51.267 | 42.552 | 38.688 | 49.284 | 50.253 | 51.520 | 46.923 |
| 49.047 | 49.352 | 62.243 | 51.998 | 49.422 | 49.058 | 47.963 | 33.335 |
| 49.516 | 49.259 | 61.606 | 60.121 | 49.928 | 48.863 | 59.006 | 58.518 |
| 50.728 | 50.737 | 46.479 | 57.055 | 51.300 | 50.038 | 40.207 | 57.938 |
| 50.787 | 49.690 | 66.989 | 42.054 | 50.348 | 47.735 | 48.378 | 61.309 |
| 49.544 | 49.101 | 57.268 | 41.906 | 50.267 | 48.263 | 35.886 | 45.464 |
| 48.965 | 51.362 | 59.175 | 42.149 | 50.574 | 50.918 | 52.563 | 53.520 |
| 51.174 | 49.519 | 45.812 | 59.551 | 48.711 | 50.186 | 49.687 | 54.671 |
| 51.024 | 50.267 | 52.318 | 57.476 | 51.217 | 50.383 | 40.119 | 48.431 |

Table 5.1: First 10 rows of $\alpha$ matrix

| sample | v1 | v2 | v3 | v4 | v5 | v6 | v7 | v8 |
|---|---|---|---|---|---|---|---|---|
| 1 | 0.108 | 0.129 | 0.116 | 0.090 | 0.118 | 0.141 | 0.138 | 0.159 |
| 2 | 0.107 | 0.128 | 0.103 | 0.141 | 0.144 | 0.143 | 0.117 | 0.117 |
| 3 | 0.104 | 0.135 | 0.160 | 0.136 | 0.106 | 0.136 | 0.155 | 0.067 |
| 4 | 0.134 | 0.111 | 0.123 | 0.147 | 0.096 | 0.118 | 0.127 | 0.144 |
| 5 | 0.107 | 0.154 | 0.142 | 0.117 | 0.131 | 0.115 | 0.094 | 0.142 |
| 6 | 0.110 | 0.126 | 0.155 | 0.086 | 0.125 | 0.108 | 0.152 | 0.139 |
| 7 | 0.154 | 0.140 | 0.153 | 0.100 | 0.130 | 0.126 | 0.089 | 0.109 |
| 8 | 0.122 | 0.121 | 0.137 | 0.097 | 0.117 | 0.141 | 0.158 | 0.108 |
| 9 | 0.139 | 0.119 | 0.114 | 0.146 | 0.128 | 0.131 | 0.095 | 0.128 |
| 10 | 0.135 | 0.114 | 0.118 | 0.143 | 0.155 | 0.105 | 0.108 | 0.122 |

Table 5.2: First 10 samples of the compositional simulation data

| | cummulative_explained_variance | median | 2.50% | 97.50% |
|---|---|---|---|---|
| v3/v4 | 0.248 | -0.039 | -0.827 | 0.684 |
| v3/v8 | 0.475 | 0.028 | -0.743 | 0.814 |
| v4/v7 | 0.682 | -0.005 | -0.727 | 0.672 |
| v6/v7 | 0.792 | 0.014 | -0.550 | 0.635 |
| v1/v8 | 0.872 | 0.016 | -0.518 | 0.735 |
| v4/v5 | 0.939 | -0.004 | -0.717 | 0.478 |
| v2/v5 | 1.000 | 0.007 | -0.358 | 0.390 |

Table 5.3: Sequence of logratios of markers entering in a stepwise search, explaining the logratio variance of the whole compositional data set.

Table 5.3 presents the sequence of logratios of markers entering in a stepwise search, explaining the logratio variance of the simulation compositional data set. Table 5.3 also reports the medians of these ratios, as well as their reference ranges based on the estimated 0.025 and 0.975 quantiles. As we can observe from Table 5.3, the true signal v3, v4, v7, v8 was detected by the first 3 logratios. Notice that the top 3 logratios in Table 5.3 are v3/v4, v3/v8 and v4/v7, and they represent all the pairwise combination of v3, v4, v7, v8. ($v3/v7 = v3/v4 \times v4/v7$) And the top 3 logratio explained 68.2% of the total variance.

In order to evaluate the proposed approach for the analysis of microbiome data, we plan to implement comprehensive simulation studies that take into account key aspects of microbiome data, especially the large proportion of zeros (sparsity) and high-dimensionality. The pooled samples from the metagenomic data we use in the real data analysis section contain 387 CRC cases and 384 healthy controls. To simulate this, we will consider varying numbers of samples (100, 200, 300, ...,1000) with 2 groups, where cases and controls are generated with 1:1 ratio. The metagenomic data we use in this study contains 719 features. To mimic this, we will consider 700 features, with varying numbers of signal features (10, 50, 100). The matrix of the metagenomic data we use in this paper contains 85% zeros. By considering this, we will consider varying proportions of zeros when generating the simulation data. In addition, we will consider the case without zeros (0%) to define the baseline with perfect information and to quantify relative information loss.

To mimic the microbiome data where excessive numbers of zero values are presented, we will simulate each feature from a zero-inflated Beta (ZIB) distribution. More specifically, to model structural zeros, for $j$-th species, we assume $Y_j$ follows ZIB distribution with parameters $(\pi_j, a_j, b_j)$, where $\pi_j$ represents the probability of non-zeros. We can generate data from ZIB distribution by using the two-step approach. Specifically, we first generate the latent non-zero indicator $Z_j$ from Bernoulli distribution of parameter $\pi_j$. If $Z_j = 0$, then we set $Y_j = 0$. If $Z_j = 1$, then we generate $Y_j$ from Beta distribution

with parameters $(a_j, b_j)$.

We will set the parameters $(\pi_j, a_j, b_j)$ so that it can closely match the levels of zero-inflation in the real metagenome data. Specifically, we will obtain the empirical distribution of proportions of zeros from the metagenome data and generate $\pi_j$ from this empirical distribution. Likewise, we will fit Beta distribution to each feature in the metagenome data and obtain MLEs or MOMs (or their robust versions) of $a_j$, $b_j$. We will use these estimated $a_j, b_j$ to generate non-zero part of data.

We evaluate the performance of the zero-inflated Wilcoxon rank sum test by considering the simulation univariate, and apply the test on each variables. After controling for FDR, features with adjusted p-value less than 0.5 will be selected as significant feature. Sensitivity and specificity will be calucated to evaluate the accuracy of detection.

| CRC | | | Control | | |
|---|---|---|---|---|---|
| shape1 | shape2 | % nonzero | shape1 | shape2 | % nonzero |
| 0.05 | 27.77 | 0.81 | 0.09 | 120.28 | 0.54 |
| 0.23 | 102.69 | 0.68 | 0.07 | 56.77 | 0.4 |
| 0.13 | 92.35 | 0.4 | 0.23 | 784.83 | 0.05 |
| 0.17 | 107.77 | 0.42 | 0.06 | 212.54 | 0.07 |
| 0.1 | 97.62 | 0.39 | 0.1 | 586.44 | 0.08 |
| 0.05 | 11.97 | 0.57 | 0.03 | 63.87 | 0.2 |
| 0.13 | 34.87 | 0.45 | 0.03 | 34.9 | 0.09 |
| 0.06 | 6.06 | 0.32 | 0.07 | 99.49 | 0.06 |
| 0.12 | 14.18 | 0.12 | 24.88 | 519449.5 | 0.01 |
| 0.03 | 6.36 | 0.17 | 1.61 | 20701.7 | 0.01 |
| 0.1 | 104.35 | 0.67 | 0.05 | 395.16 | 0.42 |
| 0.09 | 66.7 | 0.1 | 11.76 | 76473.82 | 0.01 |
| 0.14 | 289.1 | 0.12 | 0.67 | 22929.41 | 0.02 |
| 0.21 | 72.36 | 0.12 | 0.54 | 9032 | 0.02 |
| 0.33 | 1304.11 | 0.08 | 0.94 | 139234.4 | 0.01 |

Table 5.4: Parameter estimation of signal species using MOMs approach, shape 1 and shape 2 are the parameter estimation of nonzero elements using beta distribution, % nonzero is the percent of nonzero element of that signal species

| CRC | | | Control | | |
|---|---|---|---|---|---|
| shape1 | shape2 | % nonzero | shape1 | shape2 | % nonzero |
| 0.28 | 636.35 | 0.02 | 0.53 | 2069.32 | 0.03 |
| 0.2 | 412.11 | 0.03 | 0.16 | 307.49 | 0.04 |
| 0.36 | 89.33 | 0.13 | 0.27 | 115.72 | 0.12 |
| 0.39 | 33.8 | 0.86 | 0.51 | 44.25 | 0.86 |
| 67.81 | 17713.99 | 0.01 | 2.99 | 1541.99 | 0.01 |
| 0.18 | 59.06 | 0.18 | 0.07 | 20.28 | 0.17 |
| 0.27 | 627.46 | 0.04 | 0.39 | 1181.77 | 0.04 |
| 0.29 | 377.83 | 0.05 | 0.67 | 1188.22 | 0.04 |
| 0.36 | 9442.15 | 0.19 | 1.05 | 35626.28 | 0.2 |
| 4.87 | 97085.22 | 0.04 | 1.89 | 35215.35 | 0.05 |
| 1.02 | 16898.4 | 0.17 | 0.31 | 4390.55 | 0.18 |
| 0.36 | 14992.73 | 0.01 | 0.53 | 2820.13 | 0.01 |
| 0.54 | 48.74 | 0.01 | 0.18 | 0.66 | 0.01 |
| 0.21 | 53.24 | 0.33 | 0.15 | 21.92 | 0.32 |
| 0.33 | 1304.11 | 0.08 | 0.94 | 139234.4 | 0.01 |

Table 5.5: Parameter estimation of noise species using MOMs approach

| | Rank Sum test | Zero-inflated Wilcoxon test |
|---|---|---|
| Sensitivity | 0.970 | 0.990 |
| Specificity | 0.990 | 0.983 |

Table 5.6: Sensitivity and Specificity for Rank sum test versus Zero-inflated Wilcoxon test

For example, we pick the signal species using pvalue cutoff at 0.001 (A species is singal is $pvalue < 0.001$) and we pick the noise with pvalue cutoff at 0.7 (A species is singal is $pvalue > 0.7$). Table 5.4 and Table 5.5 shows the parameter estimation of signal and noise species using MOMs approach, respectively. Shape 1 and shape 2 are the parameter estimation of nonzero elements under beta distribution assumption, "% nonzero" in Table 5.4 indicates the percent of nonzero element of that signal species. Simulation samples are generated by randomly select a set of $(\pi, a, b)$ for CRC and control group. In this simulation study, we choose sample size n as 600, with 300 samples in both CRC and control. We set the first 100 species as signal, and the next 600 species as noise.

Table 5.6 shows the comparison of the performance between Rank sum test and Zero-inflated Wilcoxon test. In this example, Zero-inflated Wilcoxon test shows better performance in detecting true signal, while Rank Sum test is slightly better for specificity.

## 5.4 Real data analysis

As an application of the proposed approach, we used the immune cellular fractions data for Colorectal Adenocarcinoma TCGA PanCancer study characterized by CIBERSORT [9]. TCGA data for colorectal cancer previously analyzed by Thorsson and colleagues [48] can be accessed through the National Cancer Institute Genomic Data Commons. We downloaded the data from the cBioportal database (http://www.cbioportal.org/) and kept all colorectal cancer patients with immunophenotype data for whom race, sex and survival data were available. In this data of 254 patients, 58 (23%) were African American and 196 (77%) were European American. 126 were female (50%) and 128 were male (50%). Previous studies have shown remarkable discrepancy exists in outcomes between different sex [59] [60] and race [61]. In this compositional data analysis setting, we sought to identify key immune subtypes and potentially expose sex and race as predictors of response to cancer immunotherapy.

| | Weight | Procruste Correlation |
|---|---|---|
| Macrophage | 0.453 | 0.602 |
| T.cells.CD8 | 0.142 | 0.562 |
| Mast.cells | 0.088 | 0.528 |
| T.cells.CD4 | 0.165 | 0.524 |
| B.cells | 0.072 | 0.523 |
| NK.cells | 0.051 | 0.487 |
| Neutrophils | 0.008 | 0.461 |
| Dendritic.cells | 0.017 | 0.415 |
| Eosinophils | 0.003 | 0.398 |

Table 5.7: Results for ALRs using each part in turn as the reference one in the denominator: weight is the average proportion of the reference part, used in the weighted analysis; Procrustes correlation measures similarity between the multidimensional geometry of the samples in the ALR space and that of the samples using all logratios.

The categorization of the leukocyte composition associated with each colon cancer sample within TCGA was discussed in the Immune Landscape of Cancer [50]. There were 3 types of aggregation described in the Supplementary Materials of Thorsson et al. We used “Aggregate 2” that was implemented in the study, where immune cell subsets are aggregated into nine classes with respect to the cytokine network, including CD8 T cells, CD4 T cells (naïve, memory, resting and activated), B cells (naïve and memory), NK cells (resting and activated), macrophage (M0, M1, M2), dendritic cells (resting, activated), mast cells (resting and activated), neutrophils and eosinophils; After aggregation of the immune cell subsets, we re-normalized the immune cellular fractions so that they can sum to 100% as compositional data.

| | cummulative explained variance | median | 2.50% | 97.50% |
|---|---|---|---|---|
| T.cells.CD4/Macrophage | 0.320 | -1.078 | -8.131 | 0.591 |
| Macrophage/Dendritic.cells | 0.563 | 3.891 | 1.250 | 12.124 |
| T.cells.CD8/Mast.cells | 0.699 | 0.435 | -2.531 | 5.194 |
| T.cells.CD8/Macrophage | 0.805 | -1.222 | -4.014 | 0.666 |
| B.cells/Macrophage | 0.892 | -2.175 | -5.662 | 0.445 |
| NK.cells/Macrophage | 0.955 | -2.375 | -5.981 | -0.449 |
| Macrophage/Neutrophils | 0.988 | 7.724 | 1.779 | 8.612 |
| Macrophage/Eosinophils | 1.000 | 8.974 | 2.392 | 9.534 |

Table 5.8: Sequence of logratios of markers entering in a stepwise search, explaining the logratio variance of the whole compositional data set.

The stepwise procedure starts by selecting, from the 36 logratios in this example, the one that explains the most logratio variance, using redundancy analysis (RDA). The sequence of ratios and their accumulated explained variances are given in Table 5.8. In addition, Table 5.8 reports the medians of these ratios, as well as their reference ranges based on the estimated 0.025 and 0.975 quantiles (i.e., 2.5 and 97.5% percentiles, respectively). The logratio of T.cells.CD4/Macrophage turned out to be the best, explaining 32.0% of the variance. The second best is Macrophage/Dendritic.cells, explaining an additional 24.3%; so the variance explained is now 56.3%. Then, T.cells.CD8/Mast.cells and T.cells.CD8/Macrophage brings the variance explained up to 80.5%, and so on. The number of links to Macrophags being 8 indicates large variation.

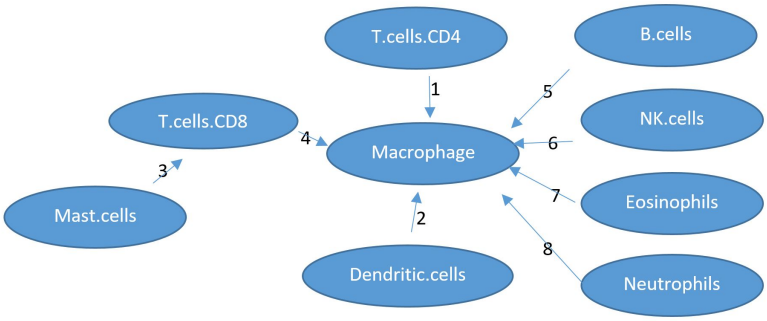



Figure 5.1: Graph of the eight ratios chosen in a stepwise procedure to explain maximum logratio variance, with numbers indicating their rank in the variable selection.

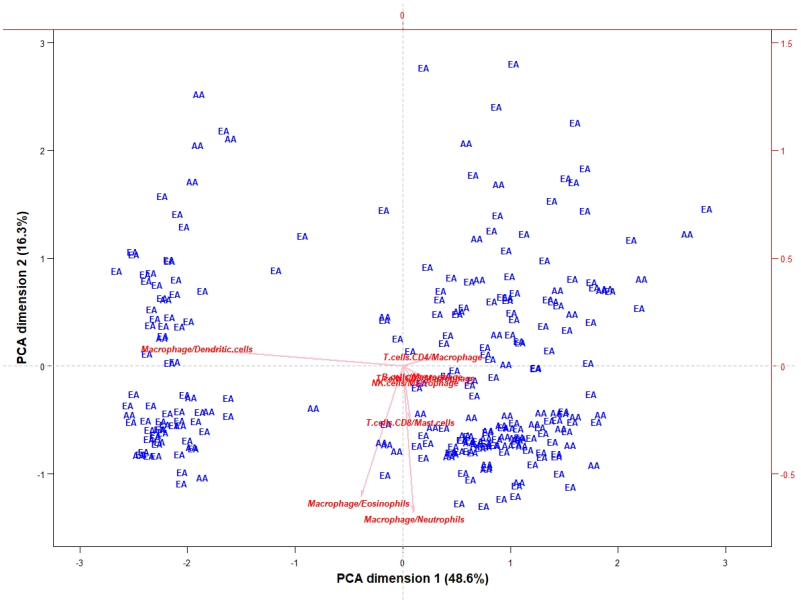



Figure 5.2: PCA contribution biplot of the eight ratios chosen in a stepwise procedure to explain maximum logratio variance.

Figure 5.1 represents the set of ratios in its acyclic graph connecting all the parts, where the numbers on the edges show their order of entry in the stepwise procedure, large number of links indicate more variation for that immune cell type. Figure 5.1 shows that the number of links to Macrophags is 8, it indicates Macrophags presents the largest variation revealed in the tumor immune microenvironment. Based on the ranking shown in the Figure 5.1, Macrophages, CD4 T cells, dendritic cells, CD8 T cells, and mast cells reveals most variation in the tumor immune microenvironment, which is consistent to the cell types identified in recent study ( [62]).

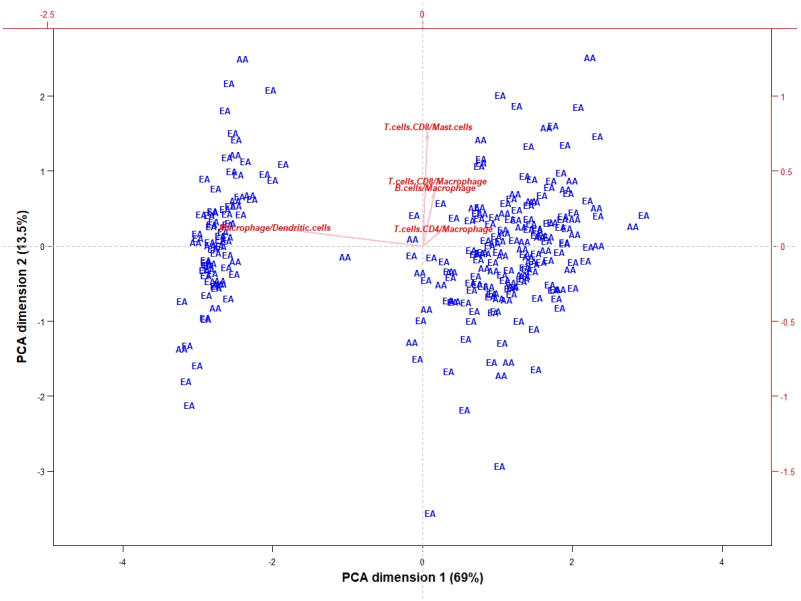



Figure 5.3: PCA contribution biplot of the top five ratios chosen in a stepwise procedure to explain maximum logratio variance.

The logratio biplot of using the 10 identified logratios in Figure 5.2 provides more insight on the choice of the ratios. These are the 10 identified logratios in Table 5.8, which explain 100% of the logratio variance. PCA is applied on these 10 identified logratios, and Figure 5.2 presents the first two dimension of the dimension reduction result. The first two dimension explains combined 64.6% of the total variance. The Macrophage versus Dendritic.cells opposition appears as the most important along the first axis, which clarifies the choice of the second ratio as Macrophage/Dendritic.cells. In addition, the samples are labeled as AA and EA, represent African American and European American respectively. We can find certain clustering pattern where AA are mostly clustered in the top-left and bottom-right in the biplot.

Figure 5.3 presents the biplot of the top five ratios chosen in a stepwise procedure to explain maximum logratio variance. Here the first two dimension explains combined 72.5% of the total variance. Moreover, most African American samples are clustered in the bottom-left of the biplot, which indicates that the top five ratios chosen in a stepwise procedure reveals certain level of racial difference.

We also include the biplot of the top five ratios with the samples labeled with both sex and racial information, which can be found in the Supplemental Figure A.5 and A.6. It is challenging to identify sex and racial subgroups from these two plots.

For the real data analysis of microbiome data, we downloaded the preprocessed metagenomic data of six cross-sectional studies of colonrectal cancer from the curatedMetagenomicData R package [63], which is supplemental to the meta-analysis by Thomas, A. M. [64], et al. The data includes information of metadata, taxonomic and functional composition, and relative abundance of each species and metabolic pathway. Since the preprocessed metagenomic data for the other two validation studies used in the meta-analysis were not available, we downloaded their raw fastq files from the DNA Data Bank of Japan database (project No. DRA00668432) and European Nucleotide Archive (project No. PRJEB2792814), respectively. Data preprocessing was performed strictly as described in the meta-analysis to make it consistent with the other six cohorts. Overall, we use the name of the country where the cohorts were recruited to denote the eight studies: ThomasAM_2018a represents cohort in Italy (ITA 1), ThomasAM_2018b represents cohort in Italy (ITA 2), FengQ_2015 represents cohort in Austria (AUS), VogtmannE_2016 represents cohort in the United States (USA), YuJ_2015 represents cohort in China (CHI), and ZellerG_2014 represents cohort in France (FRA), project No. DRA006684 represents cohort in Japan (JAP), and project No. PRJEB27928 represent cohorts in Germany (GEM). In total, the pooled samples contain 387 CRC cases and 384 healthy controls.

| | # of control | # of CRC | T-test | Rank sum test | Zero-inflated Wilcoxon |
|---|---|---|---|---|---|
| FengQ_2015 | 61 | 46 | 0.000 | 0.057 | 0.081 |
| ThomasAM_2018a | 24 | 29 | 0.000 | 0.000 | 0.000 |
| ThomasAM_2018b | 28 | 32 | 0.000 | 0.010 | 0.010 |
| VogtmannE_2016 | 52 | 52 | 0.000 | 0.000 | 0.011 |
| YuJ_2015 | 53 | 75 | 0.000 | 0.045 | 0.078 |
| ZellerG_2014 | 61 | 53 | 0.000 | 0.021 | 0.036 |
| PRJDB4176 | 40 | 40 | 0.000 | 0.009 | 0.023 |
| PRJEB27928 | 64 | 60 | 0.004 | 0.174 | 0.190 |
| Pooled | 383 | 387 | 0.037 | 0.181 | 0.200 |

Table 5.9: Proportion of species selected using T-test, Wilcoxon rank sum test and Zero-inflated Wilcoxon test.

Table 5.9 presents the proportion of species selected using T-test, Wilcoxon rank sum test and Zero-inflated Wilcoxon test. We can observe from the table that t-test has the least power in identifying the signal species. Rank sum test and zero-inflated Wilcoxon test show similar performance. Zero-inflated Wilcoxon test is slightly more powerful in finding key species, where the proportion of species selected are higher across all cohorts and meta data analysis. The large proportion of zeros played an important role in the microbiome data analysis. For the rank sum test, the excessive number of zeros results in many ties when calculate the rank sum, which lead to loss of power.

| | # of selected species | Top 3 species | | |
|---|---|---|---|---|
| FengQ_2015 | 28 | Prevotella_copri | Fusobacterium_nucleatum | Porphyromonas_asaccharolytica |
| ThomasAM_2018a | 0 | NA | NA | NA |
| ThomasAM_2018b | 4 | Gemella_morbillorum | Parvimonas_micra | Parvimonas_unclassified |
| VogtmannE_2016 | 0 | NA | NA | NA |
| YuJ_2015 | 22 | Peptostreptococcus_stomatis | Parvimonas_unclassified | Gemella_morbillorum |
| ZellerG_2014 | 11 | Fusobacterium_nucleatum | Peptostreptococcus_stomatis | Porphyromonas_asaccharolytica |
| PRJDB4176 | 4 | Parvimonas_unclassified | Gemella_morbillorum | Peptostreptococcus_stomatis |
| PRJEB27928 | 78 | Anaerotruncus_unclassified | Parvimonas_unclassified | Solobacterium_moorei |
| pooled | 126 | Parvimonas_unclassified | Peptostreptococcus_stomatis | Fusobacterium_nucleatum |

Table 5.10: Number of significant species and the top 3 species identified using Wilcoxon rank sum test

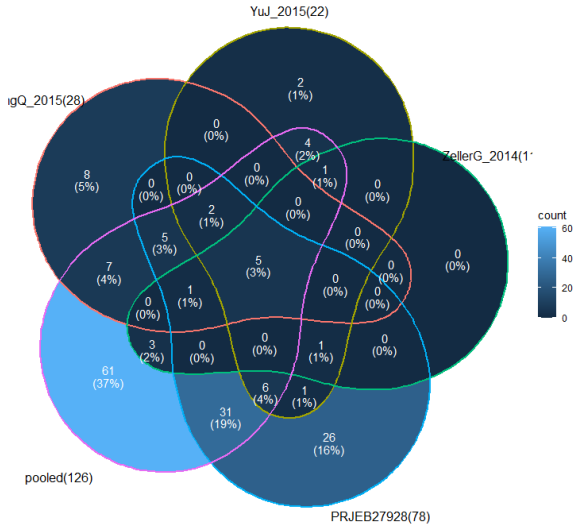



Figure 5.4: Venn Diagram of the cohorts and metadata for significant species identified using Wilcoxon rank sum test

| | # of selected species | Top 3 species | | |
|---|---|---|---|---|
| FengQ_2015 | 40 | Fusobacterium_nucleatum | Prevotella_copri | Porphyromonas_asaccharolytica |
| ThomasAM_2018a | 0 | NA | NA | NA |
| ThomasAM_2018b | 4 | Gemella_morbillorum | Parvimonas_micra | Parvimonas_unclassified |
| VogtmannE_2016 | 5 | Fusobacterium_nucleatum | Porphyromonas_uenonis | Gemella_morbillorum |
| YuJ_2015 | 38 | Peptostreptococcus_stomatis | Parvimonas_unclassified | Parvimonas_micra |
| ZellerG_2014 | 19 | Fusobacterium_nucleatum | Peptostreptococcus_stomatis | Porphyromonas_asaccharolytica |
| PRJDB4176 | 10 | Gemella_morbillorum | Parvimonas_unclassified | Peptostreptococcus_stomatis |
| PRJEB27928 | 85 | Parvimonas_unclassified | Anaerotruncus_unclassified | Parvimonas_micra |
| pooled | 139 | Clostridium_hathewayi | Clostridium_symbiosum | Fusobacterium_nucleatum |

Table 5.11: Number of significant species and the top 3 species identified using Zero-inflated Wilcoxon test

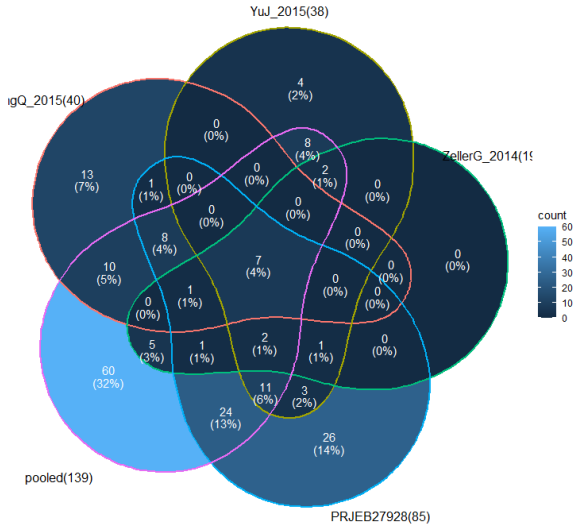



Figure 5.5: Venn Diagram of the cohorts and metadata for significant species identified using Zero-inflated Wilcoxon test

Figure 5.4 present the Venn Diagram of the significant species identified using Wilcoxon rank sum test comparing the cohorts FengQ_2015, YuJ_2015, ZellerG_2014, PRJEB27928, and the metadata including all 8 cohorts studies. The lighter the color in the plot, the large count it is. Also, the number of total number of significant species for each cohort is listed beside the study name in the plot. Figure 5.4 shows that there are 5 species that are shared by the 5 categories. And those 5 species are "Clostridium_hathewayi", "Fusobacterium_nucleatum", "Parvimonas_micra", "Peptostreptococcus_stomatis", and "Porphyromonas_asaccharolytica". Clostridium_hathewayi is a newly described gram-negative, endospore-forming, rod-shaped bacterium, and Fusobacterium nucleatum is an oral bacterium, commensal to the human oral cavity, that plays a role in periodontal disease. Parvimonas micra is a rare pathogen for psoas abscess and a Gram-positive anaerobic coccus. Peptostreptococcus stomat is a species of anaerobic, gram-positive coccoid bacteria belonging to the genus. Porphyromonas asaccharolytica is a rare causative agent for Lemierre's Syndrome.

On the other hand, the Venn Diagram in Figure 5.5 shows the significant species identified using Zero-inflated Wilcoxon test comparing the cohorts FengQ_2015, YuJ_2015, ZellerG_2014, PRJEB27928, and the metadata including all 8 cohorts studies. Figure 5.5 shows that there are 7 species that are shared by the 5 categories. And those 5 species are "Clostridium_hathewayi", "Fusobacterium_nucleatum", "Parvimonas_micra", "Peptostreptococcus_stomatis",

"Porphyromonas_asaccharolytica", "Parvimonas_unclassified" and "Gemella_morbillorum". The first 5 species are the same as the 5 species identified using Wilcoxon rank sum test. The two species that are uniquely identified using Zero-inflated Wilcoxon test are "Parvimonas_unclassified" and "Gemella_morbillorum". There are 18 items that were listed as unclassified Parvimonas in the taxonomy browswer in NCBI. Limited information is available for "Parvimonas_unclassified". As for "Gemella_morbillorum", it is a species of bacteria within the genus Gemella. It is a facultative anaerobic Gram positive coccus usually preferring capnophilic or microaerophilic environments. Zero-inflated Wilcoxon shows higher power in detecting signal species.

Table 5.10 presents the number of significant species and 3 species with smallest p-values using Wilcoxon rank sum test. If we consider only the cohorts with number of selected species larger than 0. There is only 1 species that are shared by the rest of the categories, which is shown in the Supplemental Figure A.1, and that one species is "Peptostreptococcus_stomatis". In the meantime, Supplemental Figure A.2 presents Venn Diagram of significant species identified using Zero-inflated Wilcoxon test for the same 6 cohorts and metadata. Using Zero-inflated Wilcoxon test, 4 species are shared among all categories. More detailed Venn Diagram comparing the two approach for 4 cohorts only are included in the Supplemental Figure A.3 and A.4. Overall, we observe higher number of selected species identified using Zero-inflated Wilcoxon test for almost all the cases.

## 5.5 Conclusions

In this aim, we investigated variable selection in compositional data analysis with application to immunology data and microbiome data. For low-dimensional microbiome data, we applied stepwise pairwise log-ratio procedure for variable selection and identified key immune subtypes using cellular fractions data induced from Colorectal Adenocarcinoma TCGA PanCancer study. Macrophags presents the largest variation in the tumor immune microenvironment. It is also shown that Macrophages, CD4 T cells, dendritic cells, CD8 T cells, and mast cells are the top ranking immnue subtypes with most variation. By applying the stepwise pairwise log-ratio procedure, we identified the key immune subtypes and revealed the relationship among the top ranking subtypes. As for the microbiome data, we took into account key aspects of the data including large proportion of zeros and high dimensionality. Key species are identified in the metagenomic data of six cross-sectional studies of CRC by applying zero-inflated Wilcoxon test for variable selection. Overall, various variable selection approaches investigated in compositional data analysis for immunology data and microbiome data in this study will provide researchers meaningful insight for identifying key features in cancer study.

## REFERENCES


[1] Cancer facts & figures 2018 | american cancer society. https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2018.html. (Accessed on 07/22/2020).

[2] The Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609–615, June 2011.

[3] The Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407):330–337, July 2012.

[4] The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature*, 513(7517):202–209, September 2014.

[5] Cécile Chauvel, Alexei Novoloaca, Pierre Veyre, Frédéric Reynier, and Jérémie Becker. Evaluation of integrative clustering methods for the

analysis of multi-omics data. *Briefings in Bioinformatics*, 21(2):541–552, 02 2019.

[6] Wei Wei, Zequn Sun, Willian A da Silveira, Zhenning Yu, Andrew Lawson, Gary Hardiman, Linda E Kelemen, and Dongjun Chung. Semi-supervised identification of cancer subgroups using survival outcomes and overlapping grouping information. *Statistical Methods in Medical Research*, 28(7):2137–2149, July 2019.

[7] J. Aitchison. Logratio analysis of compositions. *The Statistical Analysis of Compositional Data*, page 141–183, 1986.

[8] Michael Greenacre. Variable Selection in Compositional Data Analysis Using Pairwise Logratios. *Mathematical Geosciences*, 51(5):649–682, July 2019.

[9] Aaron M Newman, Chih Long Liu, Michael R Green, Andrew J Gentles, Weiguo Feng, Yue Xu, Chuong D Hoang, Maximilian Diehn, and Ash A Alizadeh. Robust enumeration of cell subsets from tissue expression profiles. *Nature Methods*, 12(5):453–457, May 2015.

[10] Alfred P. Hallstrom. A modified Wilcoxon test for non-negative distributions with a clump of zeros. *Statistics in Medicine*, 29(3):391–400, February 2010.

[11] Ronglai Shen, Adam B. Olshen, and Marc Ladanyi. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, 25(22):2906–2912, November 2009.

[12] Ronglai Shen, Sijian Wang, and Qianxing Mo. Sparse integrative clustering of multiple omics data sets. *The Annals of Applied Statistics*, 7(1):269–294, March 2013.

[13] Ronglai Shen, Sijian Wang, and Qianxing Mo. Sparse integrative clustering of multiple omics data sets. *Ann. Appl. Stat.*, 7(1):269–294, 03 2013.

[14] Qianxing Mo, Ronglai Shen, Cui Guo, Marina Vannucci, Keith S Chan, and Susan G Hilsenbeck. A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data. *Biostatistics*, 19(1):71–86, 05 2017.

[15] Eric F. Lock, Katherine A. Hoadley, J. S. Marron, and Andrew B. Nobel. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *The Annals of Applied Statistics*, 7(1):523–542, March 2013.

[16] Zi Yang and George Michailidis. A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. *Bioinformatics*, page btv544, September 2015.

[17] Johan Trygg and Svante Wold. O2-PLS, a two-block (X-Y) latent variable regression (LVR) method with an integral OSC filter. *Journal of Chemometrics*, 17(1):53–64, January 2003.

[18] Eric F. Lock and David B. Dunson. Bayesian consensus clustering. *Bioinformatics*, 29(20):2610–2616, October 2013.

[19] Paul Kirk, Jim E. Griffin, Richard S. Savage, Zoubin Ghahramani, and David L. Wild. Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics*, 28(24):3290–3297, December 2012.

[20] Haisu Ma and Hongyu Zhao. iFad: an integrative factor analysis model for drug-pathway association inference†. *Bioinformatics*, 28(14):1911–1918, July 2012.

[21] Haisu Ma and Hongyu Zhao. FacPad: Bayesian sparse factor modeling for the inference of pathways responsive to drug treatment. *Bioinformatics*, 28(20):2662–2670, October 2012.

[22] Binbin Chen, Michael Khodadoust, Chih Liu, Aaron Newman, and Ash Alizadeh. *Profiling tumor infiltrating immune cells with CIBERSORT*, volume 1711, pages 243–259. 01 2018.

[23] Gregor Sturm, Francesca Finotello, Florent Petitprez, Jitao David Zhang, Jan Baumbach, Wolf H Fridman, Markus List, and Tatsiana Aneichyk.

Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. *Bioinformatics*, 35(14):i436–i445, 07 2019.

[24] Karel Hron, Peter Filzmoser, Sandra Donevska, and Eva Fišerová. Covariance-Based Variable Selection for Compositional Data. *Mathematical Geosciences*, 45(4):487–498, May 2013.

[25] Javier Palarea-Albaladejo and Josep Antoni Martín-Fernández. zCompositions — R package for multivariate imputation of left-censored data under a compositional approach. *Chemometrics and Intelligent Laboratory Systems*, 143:85–96, April 2015.

[26] J. Palarea-Albaladejo and J.A. Martín-Fernández. A modified EM alr-algorithm for replacing rounded zeros in compositional data sets. *Computers & Geosciences*, 34(8):902–917, August 2008.

[27] J.A. Martín-Fernández, K. Hron, M. Templ, P. Filzmoser, and J. Palarea-Albaladejo. Model-based replacement of rounded zeros in compositional data: Classical and robust approaches. *Computational Statistics & Data Analysis*, 56(9):2688–2704, September 2012.

[28] A. Goodspeed, L. M. Heiser, J. W. Gray, and J. C. Costello. Tumor-Derived Cell Lines as Molecular Models of Cancer Pharmacogenomics. *Molecular Cancer Research*, 14(1):3–13, January 2016.

[29] James T. Webber, Swati Kaushik, and Sourav Bandyopadhyay. Integration of Tumor Genomic Data with Cell Lines Using Multi-dimensional Network Modules Improves Cancer Pharmacogenomics. *Cell Systems*, 7(5):526–536.e6, November 2018.

[30] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[31] Robert Tibshirani. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, January 1996.

[32] Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, October 1999.

[33] Elias Chaibub Neto, J. Christopher Bare, and Adam A. Margolin. Simulation Studies as Designed Experiments: The Comparison of Penalized Regression Models in the "Large p, Small n" Setting. *PLoS ONE*, 9(10):e107957, October 2014.

[34] Dongjun Chung and Sunduz Keles. Sparse Partial Least Squares Classification for High Dimensional Data. *Statistical Applications in Genetics and Molecular Biology*, 9(1), January 2010.

[35] Yujing Jiang, Mei-Ling Ting Lee, Xin He, Bernard Rosner, and Jun Yan. Wilcoxon Rank-Based Tests for Clustered Data with *R* Package **clusrank**. *Journal of Statistical Software*, 96(6), 2020.

[36] Karl Pearson F.R.S. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[37] Jonathan W. Pillow and James Scott. Fully bayesian inference for neural models with negative-binomial spiking. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1898–1906. Curran Associates, Inc., 2012.

[38] Nicholas G. Polson, James G. Scott, and Jesse Windle. Bayesian Inference for Logistic Models Using Pólya–Gamma Latent Variables. *Journal of the American Statistical Association*, 108(504):1339–1349, December 2013.

[39] Edward I. George and Robert E. McCulloch. Approaches for bayesian variable selection. *Statistica Sinica*, 7(2):339–373, 1997.

[40] Mingyuan Zhou and Lawrence Carin. Negative Binomial Process Count and Mixture Modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):307–320, February 2015.

[41] Siamak Zamani Dadaneh, Mingyuan Zhou, and Xiaoning Qian. Covariate-dependent negative binomial factor analysis of RNA sequencing data. *Bioinformatics*, 34(13):i61–i69, July 2018.

[42] Hedibert Lopes and Mike West. Bayesian model assessment in factor analysis. *Statistica Sinica*, 14:41–67, 01 2004.

[43] S. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

[44] S. Jones, X. Zhang, D. W. Parsons, J. C.-H. Lin, R. J. Leary, P. Angenendt, P. Mankoo, H. Carter, H. Kamiyama, A. Jimeno, S.-M. Hong, B. Fu, M.-T. Lin, E. S. Calhoun, M. Kamiyama, K. Walter, T. Nikolskaya, Y. Nikolsky, J. Hartigan, D. R. Smith, M. Hidalgo, S. D. Leach, A. P. Klein, E. M. Jaffee, M. Goggins, A. Maitra, C. Iacobuzio-Donahue, J. R. Eshleman, S. E. Kern, R. H. Hruban, R. Karchin, N. Papadopoulos, G. Parmigiani, B. Vogelstein, V. E. Velculescu, and K. W. Kinzler. Core Signaling Pathways in Human Pancreatic Cancers Revealed by Global Genomic Analyses. *Science*, 321(5897):1801–1806, September 2008.

[45] Boris J. Winterhoff, Makayla Maile, Amit Kumar Mitra, Attila Sebe, Martina Bazzaro, Melissa A. Geller, Juan E. Abrahante, Molly Klein, Raffaele Hellweg, Sally A. Mullany, Kenneth Beckman, Jerry Daniel, and Timothy K. Starr. Single cell sequencing reveals heterogeneity within ovarian

cancer epithelium and cancer associated stromal cells. *Gynecologic Oncology*, 144(3):598–606, March 2017.

[46] Peter Scarbrough, Rachel Weber, Edwin Iversen, Yonathan Brhane, Christopher Amos, Peter Kraft, Rayjean Hung, Thomas Sellers, John Witte, Paul Pharaoh, Brian Henderson, David Hunter, Judy Garber, Amit Joshi, Kevin McDonnell, Douglas Easton, Ros Eeles, Zsofia Kote-Jarai, and Joellen Schildkraut. A cross-cancer genetic association analysis of the dna repair and dna damage signaling pathways for lung, ovary, prostate, breast, and colorectal cancer. *Cancer Epidemiology Biomarkers Prevention*, 25, 12 2015.

[47] A. P. Reynolds, G. Richards, B. de la Iglesia, and V. J. Rayward-Smith. Clustering Rules: A Comparison of Partitioning and Hierarchical Clustering Algorithms. *Journal of Mathematical Modelling and Algorithms*, 5(4):475–504, December 2006.

[48] Jonathan Thorsen, Asker Brejnrod, Martin Mortensen, Morten A. Rasmussen, Jakob Stokholm, Waleed Abu Al-Soud, Søren Sørensen, Hans Bisgaard, and Johannes Waage. Large-scale benchmarking reveals false discoveries and count transformation sensitivity in 16S rRNA gene amplicon data analysis methods used in microbiome studies. *Microbiome*, 4(1):62, December 2016.

[49] Jui-Ling Yu and Sophia R.-J. Jang. A mathematical model of tumor-immune interactions with an immune checkpoint inhibitor. *Applied Mathematics and Computation*, 362:124523, December 2019.

[50] Thorsson. The Immune Landscape of Cancer. *Immunity*, 51(2):411–412, August 2019.

[51] Eduardo R. De Arantes E Oliveira. Theoretical foundations of the finite element method. *International Journal of Solids and Structures*, 4(10):929–952, October 1968.

[52] D. Venet, F. Pecasse, C. Maenhaut, and H. Bersini. Separation of samples into their constituents using gene expression data. *Bioinformatics*, 17(Suppl 1):S279–S287, June 2001.

[53] Timo Erkkilä, Saara Lehmusvaara, Pekka Ruusuvuori, Tapio Visakorpi, Ilya Shmulevich, and Harri Lähdesmäki. Probabilistic analysis of gene expression measurements from heterogeneous tissues. *Bioinformatics*, 26(20):2571–2577, October 2010.

[54] Shai S Shen-Orr, Robert Tibshirani, Purvesh Khatri, Dale L Bodian, Frank Staedtler, Nicholas M Perry, Trevor Hastie, Minnie M Sarwal, Mark M Davis, and Atul J Butte. Cell type–specific gene expression differences in complex tissues. *Nature Methods*, 7(4):287–289, April 2010.

[55] Jason E Shoemaker, Tiago JS Lopes, Samik Ghosh, Yukiko Matsuoka, Yoshihiro Kawaoka, and Hiroaki Kitano. CTen: a web-based platform for identifying enriched cell types from heterogeneous microarray data. *BMC Genomics*, 13(1):460, 2012.

[56] Kosuke Yoshihara, Maria Shahmoradgoli, Emmanuel Martínez, Rahulsimham Vegesna, Hoon Kim, Wandaliz Torres-Garcia, Victor Treviño, Hui Shen, Peter W. Laird, Douglas A. Levine, Scott L. Carter, Gad Getz, Katherine Stemke-Hale, Gordon B. Mills, and Roel G.W. Verhaak. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nature Communications*, 4(1):2612, December 2013.

[57] Arnold L. van den Wollenberg. Redundancy analysis an alternative for canonical correlation analysis. *Psychometrika*, 42(2):207–219, June 1977.

[58] J. C. Gower and Garmt B. Dijksterhuis. *Procrustes problems*. Number 30 in Oxford statistical science series. Oxford University Press, Oxford ; New York, 2004. OCLC: ocm53156636.

[59] Ching Ying Lin, Hyunwoo Kwon, Guillermo O. Rangel Rivera, Xue Li, Dongjun Chung, and Zihai Li. Sex differences in using systemic inflammatory markers to prognosticate patients with head and neck squa-

mous cell carcinoma. *Cancer Epidemiology and Prevention Biomarkers*, 27(10):1176–1185, 2018.

[60] Andreia Ribeiro, Thomas Ritter, Matthew Griffin, and Rhodri Ceredig. Corrigendum to: "Development of a flow cytometry-based potency assay for measuring the in vitro immunomodulatory properties of mesenchymal stromal cells" [J. Immunol. Lett. 177 (2016) 38–46]. *Immunology Letters*, 180:81, December 2016.

[61] Jeronay King Thomas, Hina Mir, Neeraj Kapur, and Shailesh Singh. Racial Differences in Immunological Landscape Modifiers Contributing to Disparity in Prostate Cancer. *Cancers*, 11(12):1857, November 2019.

[62] Director's Challenge Consortium for the Molecular Classification of Lung Adenocarcinoma. Gene expression–based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nature Medicine*, 14(8):822–827, August 2008.

[63] Edoardo Pasolli, Lucas Schiffer, Paolo Manghi, Audrey Renson, Valerie Obenchain, Duy Tin Truong, Francesco Beghini, Faizan Malik, Marcel Ramos, Jennifer B Dowd, Curtis Huttenhower, Martin Morgan, Nicola Segata, and Levi Waldron. Accessible, curated metagenomic data through ExperimentHub. *Nature Methods*, 14(11):1023–1024, November 2017.

[64] Andrew Maltez Thomas, Paolo Manghi, Francesco Asnicar, Edoardo Pasolli, Federica Armanini, Moreno Zolfo, Francesco Beghini, Serena Manara, Nicolai Karcher, Chiara Pozzi, Sara Gandini, Davide Serrano, Sonia Tarallo, Antonio Francavilla, Gaetano Gallo, Mario Trompetto, Giulio Ferrero, Sayaka Mizutani, Hirotsugu Shiroma, Satoshi Shiba, Tatsuhiro Shibata, Shinichi Yachida, Takuji Yamada, Jakob Wirbel, Petra Schrotz-King, Cornelia M. Ulrich, Hermann Brenner, Manimozhiyan Arumugam, Peer Bork, Georg Zeller, Francesca Cordero, Emmanuel Dias-Neto, João Carlos Setubal, Adrian Tett, Barbara Pardini, Maria Rescigno, Levi Waldron, Alessio Naccarati, and Nicola Segata. Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nature Medicine*, 25(4):667–678, April 2019.

# A. APPENDIX: Supplementary Figures and Tables

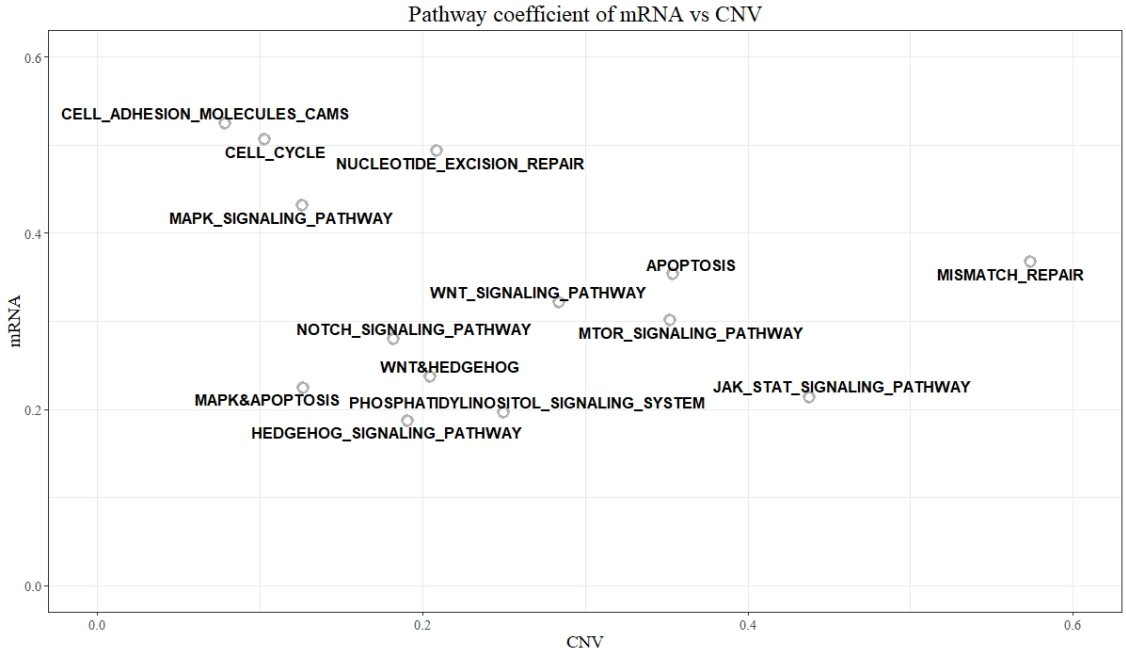



Fig. S1. Pathway Coefficient of mRNA vs CNV.

| | |
|---|---|
| KEGG_HEDGEHOG_SIGNALING_PATHWAY | 32 |
| KEGG_MTOR_SIGNALING_PATHWAY | 2 |
| KEGG_NOTCH_SIGNALING_PATHWAY | 10 |
| KEGG_NUCLEOTIDE_EXCISION_REPAIR | 21 |
| KEGG_CELL_CYCLE | 24 |
| KEGG_CELL_ADHESION_MOLECULES_CAMS | 0 |
| KEGG_JAK_STAT_SIGNALING_PATHWAY | 6 |
| KEGG_PHOSPHATIDYLINOSITOL_SIGNALING_SYSTEM | 4 |
| KEGG_MAPK_SIGNALING_PATHWAY | 0 |
| KEGG_MISMATCH_REPAIR | 14 |
| KEGG_APOPTOSIS | 0 |
| KEGG_WNT_SIGNALING_PATHWAY | 46 |
| KEGG_BASE_EXCISION_REPAIR | 10 |
| KEGG_NON_HOMOLOGOUS_END_JOINING | 1 |
| KEGG_TGF_BETA_SIGNALING_PATHWAY | 24 |

Table. S1. Gene membership of WNT & MTOR gene set

| | |
|---|---|
| KEGG_HEDGEHOG_SIGNALING_PATHWAY | 3 |
| KEGG_MTOR_SIGNALING_PATHWAY | 16 |
| KEGG_NOTCH_SIGNALING_PATHWAY | 0 |
| KEGG_NUCLEOTIDE_EXCISION_REPAIR | 0 |
| KEGG_CELL_CYCLE | 10 |
| KEGG_CELL_ADHESION_MOLECULES_CAMS | 0 |
| KEGG_JAK_STAT_SIGNALING_PATHWAY | 18 |
| KEGG_PHOSPHATIDYLINOSITOL_SIGNALING_SYSTEM | 11 |
| KEGG_MAPK_SIGNALING_PATHWAY | 62 |
| KEGG_MISMATCH_REPAIR | 0 |
| KEGG_APOPTOSIS | 42 |
| KEGG_WNT_SIGNALING_PATHWAY | 26 |
| KEGG_BASE_EXCISION_REPAIR | 0 |
| KEGG_NON_HOMOLOGOUS_END_JOINING | 0 |
| KEGG_TGF_BETA_SIGNALING_PATHWAY | 9 |

Table. S2. Gene membership of MARK & APOPTOSIS gene set

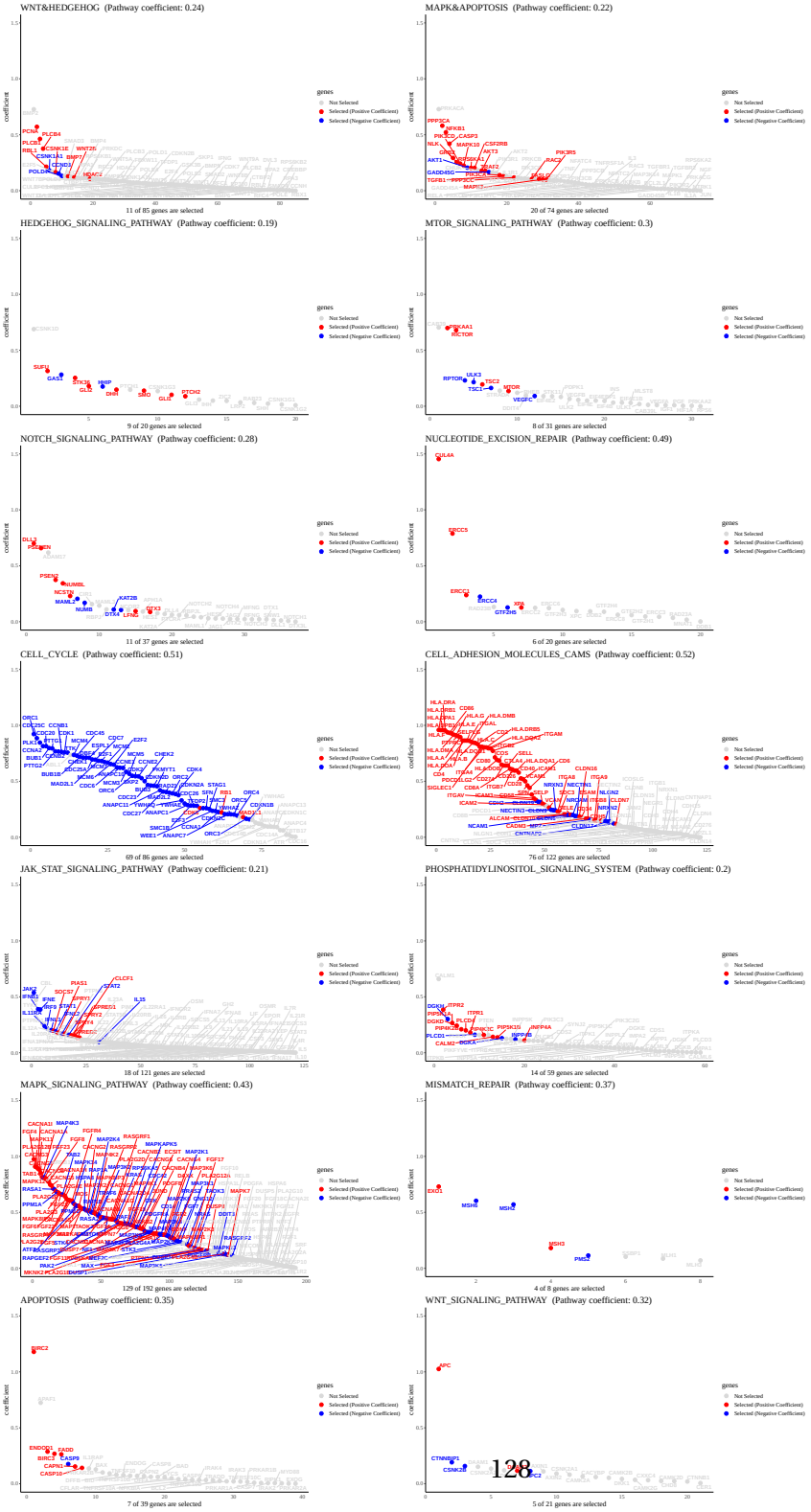



Fig. S2. Coefficients of genes for each pathway for mRNA data.

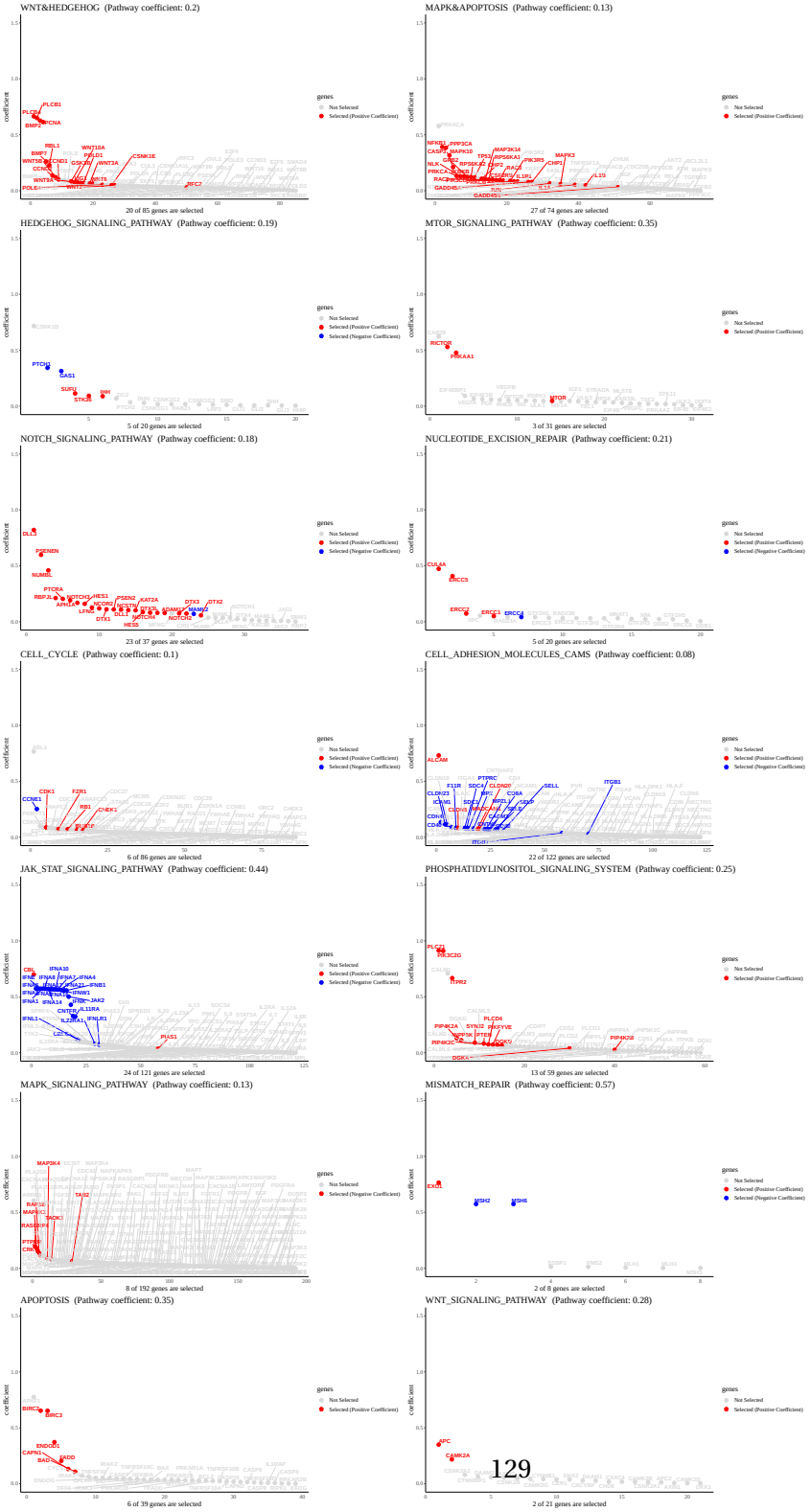



Fig. S3. Coefficients of genes for each pathway for CNV data.

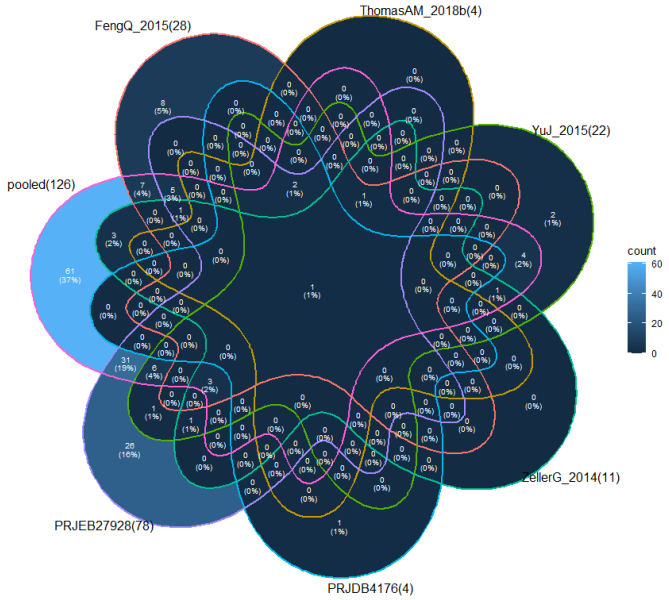



Figure A.1: Venn Diagram of the 8 cohorts and metadata for significant species identified using Wilcoxon rank sum test

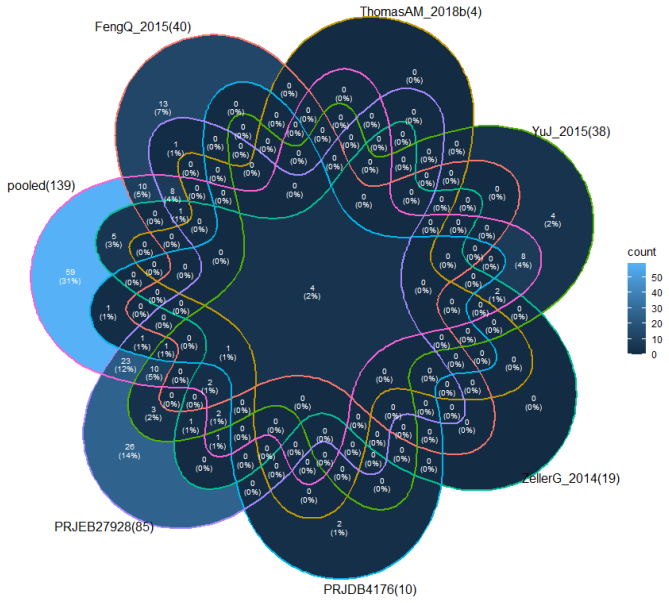



Figure A.2: Venn Diagram of the 8 cohorts and metadata for significant species identified using Zero-inflated Wilcoxon test

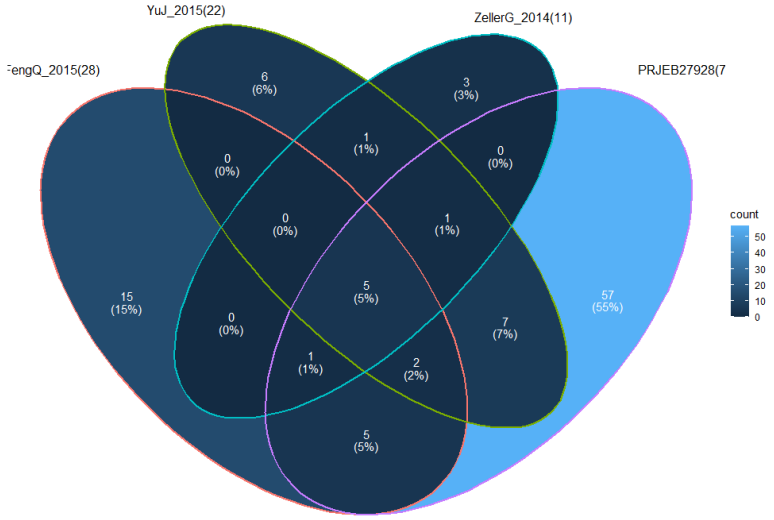



Figure A.3: Venn Diagram of the 4 cohorts for significant species identified using Wilcoxon rank sum test

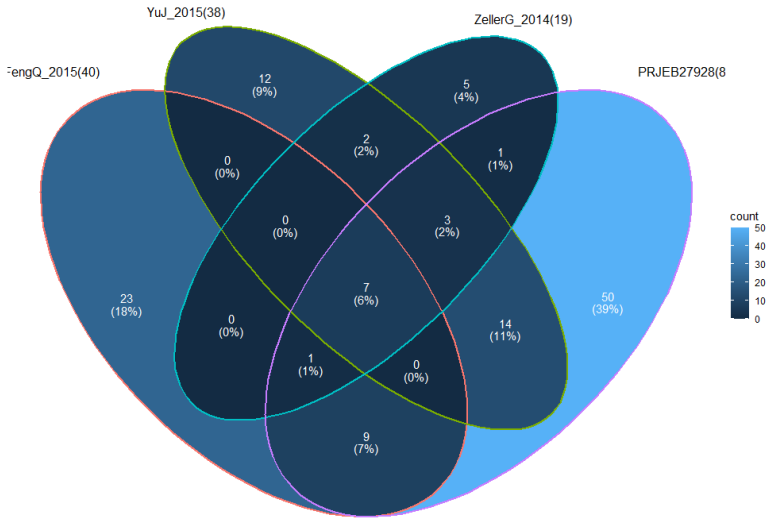



Figure A.4: Venn Diagram of the 4 cohorts for significant species identified using Zero-inflated Wilcoxon test

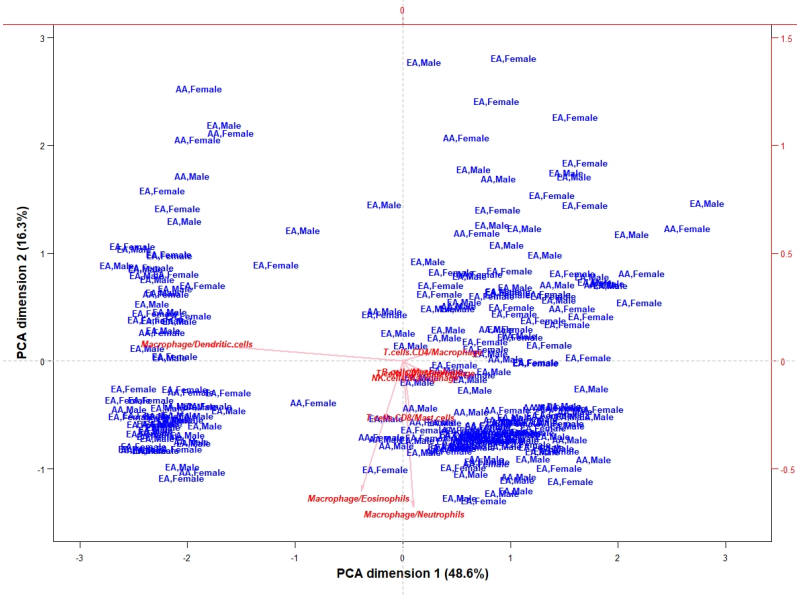



Figure A.5: PCA contribution biplot of the eight ratios chosen in a stepwise procedure to explain maximum logratio variance (each sample labeled with sex and race information).

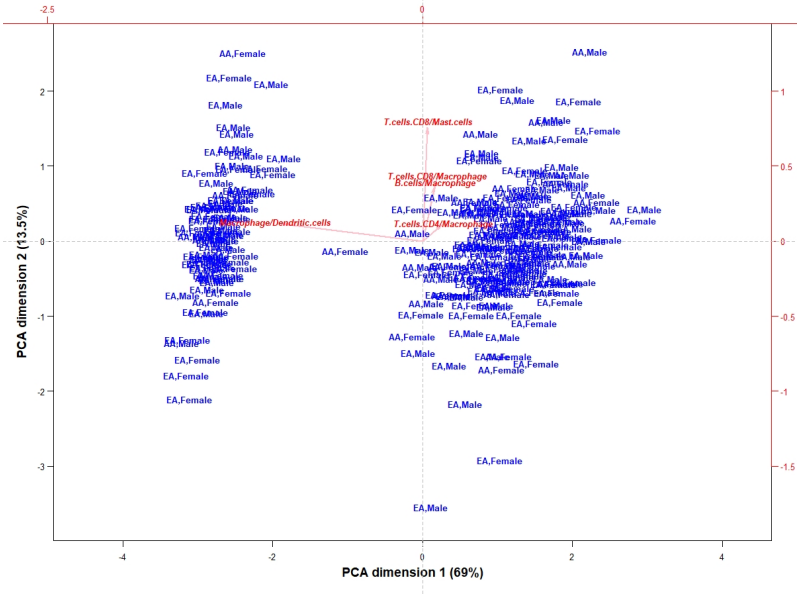



Figure A.6: PCA contribution biplot of the top 5 ratios chosen in a stepwise procedure to explain maximum logratio variance (each sample labeled with sex and race information).