

IMIX: A multivariate mixture model approach to integrative analysis of multiple types of omics data

Ziqiao Wang^{1,2} and Peng Wei^{1,*}

¹Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA

²The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences, Houston, Texas, USA

Abstract

Motivation: Integrative genomic analysis is a powerful tool to study the biological mechanisms underlying a complex disease or trait across multiplatform high-dimensional data, such as DNA methylation, copy number variation (CNV), and gene expression. It is common to perform large-scale genome-wide association analysis of an outcome for each data type separately and combine the results *ad hoc*, leading to loss of statistical power and uncontrolled overall false discovery rate (FDR).

Results: We propose a multivariate mixture model framework (IMIX) that integrates multiple types of genomic data and allows examining and relaxing the commonly adopted conditional independence assumption. We investigate across-data-type FDR control in IMIX, and show the gain in lower misclassification rates at controlled overall FDR compared with established individual data type analysis strategies, such as Benjamini-Hochberg FDR control, the q-value, and the local FDR control by extensive simulations. IMIX features statistically-principled model selection, FDR control, and computational efficiency. Applications to the Cancer Genome Atlas (TCGA) data provide novel multi-omic insights into the luminal/basal subtyping of bladder cancer and the prognosis of pancreatic cancer.

Availability and implementation: We have implemented our method in R package

*Correspondence should be addressed to: Peng Wei, PhD, 1400 Pressler St, Unit 1411, Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA; Email: pwei2@mdanderson.org

“IMIX” with instructions and examples available at <https://github.com/ziqiaow/IMIX>.

Keywords: cancer genomics; EM algorithm; false discovery rate; model selection; multivariate mixture model; omics data integration.

1 Introduction

With the development of high-throughput technology, integrative genomic analysis has become a powerful tool in biomedical research to study the biological mechanisms underlying a complex disease or trait. The mechanisms behind a certain disease outcome, such as the prognosis or the molecular subtypes of cancer, involve alterations in multiple pathways and biological processes including copy number variations (CNV), epigenetic changes, transcriptomic changes. It is a challenge to integrate all this information together to analyze the disease outcome. A common strategy is to assess the associations between genes and an outcome separately for each data type using Bonferroni correction or the Benjamini-Hochberg false discovery rate (BH-FDR) procedure (Benjamini and Hochberg, 1995) to adjust for multiple hypothesis testing. However, study on The Cancer Genome Atlas (TCGA) has found that omics data, such as gene expression, DNA methylation, and CNV, have several different triangular dependence structures (Sun et al., 2018). Furthermore, it remains unclear whether the correlation structures between data types vary according to the associations between different genes and the outcome of interest through those data types. Therefore, the heuristic separate analysis strategy loses statistical power by assuming that the data sources are independent of each other. The integration of multiple types of omics data by identifying the unknown dependence structures becomes essential to understand the intricacy of the genomic mechanisms behind complex diseases.

Previous developments in integrative genomics models mainly solve two types of problems: exploration (unsupervised) and prediction (supervised); see Richardson et al. (2016) for a comprehensive review. Unsupervised methods, such as clustering, focus on identifying similarity and uncovering heterogeneity among individuals, accounting for intersource asso-

ciations between multiple data types (Kirk et al., 2012; Mo et al., 2013; Shen et al., 2009). These methods are exploratory techniques rather than hypothesis-testing tools. Supervised methods, such as prediction or regression analysis, concentrate on outcome prediction and feature selections, such as penalized regression analysis (Tibshirani, 1996; Zou and Hastie, 2005) and nonparametric-based approach (Pineda et al., 2015). While these methods consider the association between genes and the outcome, to our best knowledge, most of them require individual-level data from the same set of samples and/or have no rigorous error-control procedure.

Here we propose a multivariate mixture model approach (IMIX) to integratively analyze omics data using summary statistics that incorporates the correlations and biological coordinations between multiple data sources to boost the statistical power for genomic discovery while controlling the across-data-type FDR. We use the EM algorithm to estimate the model parameters and propose an adaptive FDR control procedure. IMIX also features statistically-principled model selection and does not require the use of a common set of samples across data types, which relaxes the conditions of previously published methods for the integration of multiple omic data.

Through extensive simulation studies, we demonstrated that IMIX yielded better statistical power and overall FDR control compared with individual data type analysis strategies, such as BH-FDR, Bonferroni corrections, the q -value, and the local FDR control. We also observed that IMIX is computationally efficient. Lastly, we applied IMIX to study the molecular subtypes of bladder cancer through DNA methylation, CNV, and gene expression, as well as the prognosis of pancreatic cancer through gene expression, and CNV in the TCGA. Our applications of IMIX to the two TCGA data sets showed that different genomic data types can be correlated in both non-disease-associated and disease-associated genes, refuting the commonly adopted conditional independence assumption in integrative analysis of multiplatform genomic data.

2 Methods

In this section, we provide some background regarding IMIX (Section 2.1), and consider how these may be extended to allow us to perform integrative modeling of multiple datasets and model selection (Section 2.2). In section 2.3, we discuss model selection regarding the number of mixture components and the best model among proposed variants of the multivariate mixture model. We discuss local FDR (LFDR) and the adaptive procedure for across-data-type FDR control in Section 2.4.

2.1 Preliminaries and Notations for IMIX

Consider the association testing problem between gene $i, i = 1, 2, \dots, N$ and the outcome through data type $h = 1, 2, \dots, H$. For example, one wants to identify which gene is associated with a binary outcome, basal or luminal molecular subtype of bladder cancer, and assess the associations through $H = 3$ genomic data types: DNA methylation, gene expression, and CNV. It is a common practice to assume *a priori* that each genomic data type is independent and to perform statistical analysis for each dataset separately. Then, one can apply state-of-the-art methods such as the BH-FDR or the q-value for each data type to adjust for multiple hypothesis testing. Here, we group the genes into $K = 2^H$ latent states according to their associations with the h data types, and we introduce a vector of binary variables to denote each latent state k of gene i : $\mathbf{z}_{ik} = (z_{ik1}, z_{ik2}, \dots, z_{ikh})$. Without loss of generality, we assume $H = 3$. When $k = 1, 2, \dots, 8$, the potential latent states/classes of gene i are: $\mathbf{z}_{i1} = (0, 0, 0)$, $\mathbf{z}_{i2} = (1, 0, 0)$, $\mathbf{z}_{i3} = (0, 1, 0)$, $\mathbf{z}_{i4} = (0, 0, 1)$, $\mathbf{z}_{i5} = (1, 1, 0)$, $\mathbf{z}_{i6} = (1, 0, 1)$, $\mathbf{z}_{i7} = (0, 1, 1)$, and $\mathbf{z}_{i8} = (1, 1, 1)$. If $z_{ikh} = 1$, gene i is associated with the outcome through data type h in class k ; if $z_{ikh} = 0$, gene i is not associated with the outcome through data type h in class k . Dependeing on the latent state of gene i , i.e., whether it belongs to latent state k or not, we have $T_{ik} = 1$ or $T_{ik} = 0$, repsectively.

We assume that the data can be summarized as x_{ih} for each gene $i, i = 1, \dots, N$ and data

type $h = 1, 2, 3$. Here, x_{ih} is a Z-score that is transformed from the p-value of the association test for each data type (McLachlan et al., 2006; Wei and Pan, 2008). The transformation is given by $x_{ih} = \Phi^{-1}(1 - p_{ih})$, where Φ is the cumulative distribution function of the standard normal distribution $N(0, 1)$, and p_{ih} is the p-value for gene i and data type h .

We assume that $X_i = (x_{i1}, x_{i2}, x_{i3})^T$ comes from a mixture distribution with $K = 8$ mixture components:

$$f(X_i) = \sum_{k=1}^K \pi_k f_k(X_i).$$

Here, we assume the Z scores of genes in class k follow an H -dimensional multivariate distribution f_k , and the mixing proportions are $\pi_k, k = 1, \dots, 8$. To assess how likely gene i belongs to the latent state k , we estimate the posterior probability of the latent label T_{ik} :

$$Pr(T_{ik} = 1 | X_i) = \frac{\pi_k f_k(X_i)}{\sum_{j=1}^K \pi_j f_j(X_i)}.$$

2.2 Parametric Multivariate Mixture Model

Based on the principles of molecular biology and previous publications (Sun et al., 2018; Wei and Pan, 2012), when a gene is associated with an outcome through, for example, DNA methylation and gene expression, the two events can be correlated. In particular, Wei and Pan (2012) developed an integrative genomic method to improve the power of a two-component mixture model by considering the possible correlations between three data types. Here we extend this idea and formulate a multivariate mixture model of $K = 2^H$ components. We assume the k -th component distribution f_k to be multivariate normal. Normal mixture model can be used and is widely used to approximate many mixture distributions in real data applications (Sun and Cai, 2007). The marginal mixture density $f(\mathbf{X}_i)$ can then be written as

$$f(\mathbf{X}_i; \boldsymbol{\Psi}) = \sum_{k=1}^K \pi_k f_k(\mathbf{X}_i; \boldsymbol{\theta}_k),$$

where

$$f_k(\mathbf{X}_i; \boldsymbol{\theta}_k) = \Phi(\mathbf{X}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

Here, the vector $\boldsymbol{\Psi} = (\pi_1, \pi_2, \dots, \pi_{K-1}, \boldsymbol{\xi}^T)^T$ contains all the unknown parameters in the mixture model. $\boldsymbol{\xi}$ is the vector containing all the elements of the component means, $\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K$, and the elements of the covariance matrices, $\boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_K$, known *a priori* to be distinct. We use the expectation-maximization (EM) algorithm (Dempster et al., 1977) to estimate $\boldsymbol{\Psi}$. We call this generic multivariate Gaussian mixture model as **IMIX-Cor**. In the following sections we will describe three variants based on this model.

2.2.1 IMIX-Cor-Restrict: Correlated Mixture Model with Restrictions on Mean

To tackle the possible unidentifiability problem of $\boldsymbol{\Psi}$ due to the interchanging of component labels, we impose the following constraints on $\boldsymbol{\mu}_k$:

$$\begin{aligned} \boldsymbol{\mu}_1 &= (\mu_{10}, \mu_{20}, \mu_{30}); \boldsymbol{\mu}_2 = (\mu_{11}, \mu_{20}, \mu_{30}); \\ \boldsymbol{\mu}_3 &= (\mu_{10}, \mu_{21}, \mu_{30}); \boldsymbol{\mu}_4 = (\mu_{10}, \mu_{20}, \mu_{31}); \\ \boldsymbol{\mu}_5 &= (\mu_{11}, \mu_{21}, \mu_{30}); \boldsymbol{\mu}_6 = (\mu_{11}, \mu_{20}, \mu_{31}); \\ \boldsymbol{\mu}_7 &= (\mu_{10}, \mu_{21}, \mu_{31}); \boldsymbol{\mu}_8 = (\mu_{11}, \mu_{21}, \mu_{31}). \end{aligned} \tag{1}$$

These constraints correspond to the biological rational that for each data type, we assume that the means of the test statistics from the null and non-null groups are the same across the K classes. For example, μ_{10} , the mean of the null group in data type 1, is the same in $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_3$. The multivariate Gaussian mixture model with restrictions on the mean

is denoted as **IMIX-Cor-Restrict**. We estimate the parameters using the EM algorithm (Supplementary Materials, Section 1).

2.2.2 IMIX-Ind: Independent Mixture Model with Restrictions on Mean and Variance

If we assume there is no correlation between any two data types, then the covariance matrix in IMIX-Cor-Restrict, Σ_k , becomes diagonal. The model is reduced to

$$\begin{aligned} f(\mathbf{X}_i; \boldsymbol{\Psi}) &= \sum_{k=1}^K \pi_k f_k(\mathbf{X}_i; \boldsymbol{\theta}_k) \\ &= \sum_{k=1}^K \pi_k f_{k1}(\mathbf{x}_{i1}; \theta_{k1}) f_{k2}(\mathbf{x}_{i2}; \theta_{k2}) f_{k3}(\mathbf{x}_{i3}; \theta_{k3}), \end{aligned}$$

where $f_{kh}(\mathbf{x}_{ih}; \theta_{kh}) = \Phi(\mathbf{x}_{ih}; \mu_{kh}, \sigma_{kh})$ is a normal probability density function with mean μ_{kh} and variance σ_{kh}^2 , $k = 1, \dots, 8; h = 1, 2, 3$. Besides the constraints on the mean in (1), we impose the same constraints on the variance σ_{kh}^2 on the basis of the null and non-null genes for each data type. We call this model **IMIX-Ind**.

2.2.3 IMIX-Cor-Twostep: Correlated Mixture Model with Fixed Mean

To reduce the model complexity, i.e., to create a more parsimonious model, and to ease the computation time, we propose a third modification based on the previous models, the correlated mixture model with fixed mean. This model is similar to the previous model with constraints on the means; however, with the replacement of the means estimated from the independent model IMIX-Ind, we ease the complexity of estimating the mean and the covariance matrices at the same time in the EM algorithm. Based on our simulation study to be detailed later on, IMIX-Ind performed well in estimating the means; thus, to facilitate the correlation estimation between data types, we introduce the correlated mixture model with fixed mean, where we only estimate the covariance matrix for each component with a pre-specified mean vector. This provision would ease the time from estimating both the

mean and the covariance together; we will show later in the simulation study that this model achieves the best computational efficiency among IMIX models that consider the correlation structures. Here, the model estimation follows the same EM procedure as before. We call this model **IMIX-Cor-Twostep**.

2.3 Model Selection

In real data applications, one or a few classes out of K may be absent, for example, there may be no gene in class 8 in the TCGA bladder cancer example, where no gene was associated with the outcome through all three molecular mechanisms. Therefore, elimination of unnecessary components is crucial, which leads to more precise parameter estimation and well-calibrated FDR control. This is closely related to the question of how many components K to include in the mixture distribution to prevent overfitting. As pointed out by previous works (Leroux, 1992; McLachlan and Peel, 2004), the penalized log likelihood functions including Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are adequate for the number of components selection under a finite mixture distribution; in particular, under mild conditions, AIC and BIC do not underestimate the true number of components asymptotically. Specifically, AIC and BIC select the model that, respectively, minimizes

$$\begin{aligned} AIC &= -2 \times \text{loglik} + 2d, \\ BIC &= -2 \times \text{loglik} + d \log N, \end{aligned} \tag{2}$$

where d is the number of unknown parameters (i.e., degrees of freedom), N is the number of genes, and loglik is the maximized full log likelihood.

Along with model selection for the number of components K , we also select the best model for a fixed K among the different methods introduced in Section 2.2 regarding mean and covariance structures. We introduce the IMIX framework, where the data is fitted for all four IMIX methods (IMIX-Ind, IMIX-Cor, IMIX-Cor-Twostep, and IMIX-Cor-Restrict),

then AIC or BIC is utilized to select the best model among a number of candidate models, called IMIX-AIC or IMIX-BIC.

2.4 Local FDR Control and Adaptive Procedure for across-data-type FDR Control

For the trivariate mixture model of $H = 3$ data types, there are $K = 8$ latent classes, and we classify each gene i to component k_i^* , based on $k_i^* = \arg \max_k \Pr(T_{ik} = 1|X_i)$. The hypothesis to be tested is

$H_{0,i}$: Gene i does not belong to component k_i^* ;

$H_{1,i}$: Gene i belongs to component k_i^* .

The estimated posterior probability that gene i belongs to k_i^* is defined as $\hat{p}_i = \max_k \hat{Pr}(T_{ik} = 1|X_i), k = 1, \dots, 8$. The LFDR for gene i is defined as $\hat{q}_i = 1 - \hat{p}_i = \hat{Pr}(T_{ik^*} = 0|X_i)$. If \hat{q}_i is smaller than a threshold t , we reject $H_{0,i}$. This enables us to control LFDR according to the correspondence between the posterior probability and the threshold t (Efron et al., 2007).

In addition to controlling LFDR, Sun and Cai (2007) proposed an adaptive procedure to control the global FDR for mixture models. We propose to extend their method to the IMIX framework. For each component k , we construct the following hypothesis

$H_{0,i}^{(k)}$: Gene i does not belong to component k ;

$H_{1,i}^{(k)}$: Gene i belongs to component k .

Our method controls the across-data-type FDR for component k and it is defined as $\text{FDR}_k = E(N_{10,k}/R_k|R_k > 0)Pr(R_k > 0), k = 1, \dots, K$. Here $N_{10,k}$ is the number of false discoveries in component k and R_k is the total number of hypotheses claimed significant in component k . When no hypothesis is claimed significant, FDR_k is 0. The estimated posterior probability that gene i belongs to component k is defined as $\hat{p}_{i,k} = \hat{Pr}(T_{ik} = 1|X_i), k =$

1, \dots , 8. The local FDR (LFDR) for gene i is defined as $\hat{q}_{i,k} = 1 - \hat{p}_{i,k} = \hat{P}r(T_{ik} = 0|X_i)$.

The adaptive step-up procedure is described below:

$$\text{Let } m_k = \max\{i : \frac{1}{i} \sum_{j=1}^i \hat{q}_{(j),k} \leq \alpha\}; \quad (3)$$

then reject all $H_{0,(i)}$, $i = 1, \dots, m_k$,

where $\hat{q}_{(1),k}, \hat{q}_{(2),k}, \dots, \hat{q}_{(n),k}$ are the ranked values of the LFDR in component k . The adaptive procedure controls the marginal FDR (mFDR) for each component at level α asymptotically. Here mFDR $_k$ is defined as $E(N_{10,k})/E(R_k)$. The estimated mFDR becomes $\widehat{\text{mFDR}}_k = \sum_{j=1}^{m_k} \hat{q}_{(j),k}/m_k$. Genovese and Wasserman (2002) showed that under weak conditions, there exists an asymptotic relationship between mFDR and the across-data-type FDR of one component, in which $\text{mFDR}_k = \text{FDR}_k + O(N^{-1/2})$, where N is the number of hypotheses in component k and it is the same across all components. This adaptive procedure can be further used for a combination of components. For example, if we are interested in all the genes that are associated with the outcome through DNA methylation in a three-data type integration problem of DNA methylation (M), gene expression (E), and CNV, the procedure can be applied to the combination of component 2, 5, 6, and 8, i.e., (M+,E-,CNV-), (M+,E+,CNV-), (M+,E-,CNV+), and (M+,E+,CNV+).

3 Results

3.1 Simulation Studies

We performed two sets of simulation studies. Simulation study 1 assessed the performance of IMIX in terms of across-data-type FDR control, misclassification rate, and model calibration; simulation study 2 assessed the information criteria we proposed for model selection. We

consider the following multivariate normal mixture model for three data types:

$$\mathbf{X}_i \sim \sum_{k=1}^8 \pi_k \Phi(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \quad (4)$$

where $\boldsymbol{\mu}_k = (\mu_{k1}, \mu_{k2}, \mu_{k3})$ with μ_{kh} corresponds to the mean of data type h in component k , and $\boldsymbol{\Sigma}_k$ is a 3×3 matrix that contains the variance σ_{kh}^2 and the covariance $\sigma_{khh'}$ between data type h and h' in component k .

3.1.1 Across-data-type FDR Control and Misclassification Rate

To illustrate the gain in lower misclassification rate while controlling for the across-data-type FDR and LFDR of IMIX compared with the commonly used methods, we generated 1 000 simulated datasets of $N \times p = 20\,000 \times 3$ Z-scores x_{ih} following (4) in six scenarios. Scenario 1 assumed all three data types were independent with $\boldsymbol{\Sigma}_k = \text{diag}(1,1,1)$; the mean under the null was 0 and under the alternative was 3 for data type h ; the proportion of each component was balanced as $\pi_k = 0.125$. Scenarios 2-5 assumed the Z-scores were correlated under the alternative hypothesis by adding covariances (here they were also the correlations) $\sigma_{k12} = \sigma_{k13} = \sigma_{k23}$ in $\boldsymbol{\Sigma}_k$; here we only set the covariances to be non-zero for $k = 5, \dots, 8$, and each scenario corresponded to a covariance of 0.1, 0.3, 0.5, and 0.8, respectively. The rest of the parameters in Scenario 2-5 followed those of Scenario 1. Scenario 6 mimicked the real data in Section 3.2.1, where we analyzed the luminal/basal molecular subtypes through DNA methylation, gene expression, and CNV of the bladder cancer data in TCGA. We set the mean and covariance matrices in (4) equal to the empirical values estimated from Z-scores classified by separate analysis of each data type using BH-FDR method, and an unbalanced proportion equal to the estimated $\hat{\pi}$ using IMIX-Cor-Twostep (Supplementary Materials, Section 2.1). This simulation scenario thus did not favor either the separate analysis or the IMIX method.

We analyzed the simulated data by applying our proposed methods, including IMIX-Ind,

IMIX-Cor-Restrict, IMIX-Cor, IMIX-Cor-Twostep, IMIX-AIC, and IMIX-BIC; to compare the model performance with commonly used separate analysis methods, we also applied BH-FDR (Benjamini and Hochberg, 1995), Bonferroni correction, q-value (Storey, 2002; Storey et al., 2019), and the local FDR procedure (Efron et al., 2007). We set $\alpha = 0.2$ as the nominal error control level across all methods for comparisons. Note that we suggest an α value threshold between 0.05 and 0.2 for IMIX to discover interesting non-null genes while controlling the proportion of null genes (false positives) in the significant gene list.

The simulation results are presented in Figure 1 for the average of 1 000 simulations of the across-data-type FDR, which is the average of components 2-8, excluding the global null component 1; and the misclassification rate, which is the average of all components. Our proposed methods were able to robustly control the across-data-type FDR at the prespecified $\alpha = 0.2$ level. The separate analysis q-value failed to control the FDR. Bonferroni correction was designed to control the family-wise error rate (FWER), but we included it here to compare the misclassification rate with other methods as it is a popular error control procedure among researchers in biomedical sciences. The local FDR procedure deflated the FDR in Scenarios 3-5, which behaved similarly to the IMIX-Ind. Both methods were based on independent mixture distributions, and the reason why IMIX-Ind controlled the FDR slightly better than the local FDR procedure was that IMIX-Ind assumed a compound mixing proportion while the local FDR procedure only considered the mixing proportions for one data type at a time. For example, the mixing proportion π_1 of component 1 in the IMIX-Ind is only subject to the constraint $\sum_{k=1}^8 \pi_k = 1$, while π_1 in the local FDR procedure is subject to $\pi_1 = \pi_{10}\pi_{20}\pi_{30}$, where π_{h0} is the null mixing proportion for the separate analysis of data type $h = 1, 2, 3$. This was further illustrated in Scenario 6, where the underlying correlation structures and the generating model were more complicated: the local FDR procedure failed the across-data-type FDR control while IMIX-Ind controlled it robustly. BH-FDR returned slightly inflated FDR in Scenarios 1-5, and the realized FDR increased as the correlations increased among the three data types. In Scenario 6, it failed to control the across-data-type

FDR. As for the misclassification rate (Fig 1(b)), IMIX steadily achieved a lower number in all scenarios compared with commonly used methods. Our proposed methods can robustly control the across-data-type FDR and achieve a good operating characteristics under various scenarios.

In addition, we compared the computational time needed for the four IMIX models (Supplementary Material Section 2.4: table S4) using the simulation Scenario 3 assuming three data types with correlation 0.3 based on 1 000 simulations. IMIX-Ind converged the fastest within only 4.501 seconds and 67 iterations on average. Here, IMIX-Cor-Twostep achieved great computational advantages with an average of 217.379 seconds with only 42 iterations over IMIX-Cor and IMIX-Cor-Restrict, with 970.901 seconds and 417.531 seconds convergence time, 161 and 71 iterations respectively. This was processed on Intel(R) Xeon(R) CPU E5502 @ 1.87GHz with max CPU 1866 MHz and min CPU 1600 MHz.

3.1.2 Model Calibration and FDR Control

Newton et al. (2004) showed that the performance of the estimated FDR based on equation

$$FDR_{\text{estimated}}(t) = \frac{\sum_{i=1}^N q_i I(q_i \leq t)}{\sum_{i=1}^N I(q_i \leq t)}, \quad (5)$$

relies on how well the model fitting is. Thus, we need to assess the model calibration to ensure that IMIX framework is able to reliably control the realized FDR by the adaptive FDR procedures. We pursued this by comparing the realized and the estimated FDRs on the results fitted using IMIX from the six scenarios in simulation study 1. We compared the estimated and realized FDRs averaged across the 1 000 simulated datasets for each non-null component, and the sum up of non-nulls in comparison with the global null component 1. Fig S1-4 present the results of IMIX-Ind, IMIX-Cor-Restrict, IMIX-Cor, IMIX-Cor-Twostep in the six simulation scenarios. IMIX-Ind showed good model calibration in Scenarios 1 and 2, but as the correlation gradually increased from Scenario 3 to Scenario 5, the discrepancy

between estimated FDR and realized FDR increased. In Scenario 6 where we mimicked the real data, IMIX-Ind was slightly conservative as the realized FDR was slightly smaller than the estimated FDR. IMIX-Cor and IMIX-Cor-Restrict performed similarly where the estimated and realized FDRs were coincident in scenarios 1-5. In scenario 6, component 4 and component 6 showed slightly inflated realized FDR. This was because the proportion of these two components got as small as 6.9×10^{-3} and 8.9×10^{-3} . IMIX-Cor-Twostep also performed well in all scenarios except a slight shift in Scenario 5, where the correlations between data types were as high as 0.8. Since this model utilizes the estimated mean parameters from IMIX-Ind, it may have slightly affected the model calibration, but the gain in computation time was much better and can be shown in real-data based simulation Scenario 6.

In summary, the IMIX framework is rigorous and versatile to have good model calibrations under various data scenarios, which lead to a reliable and accurate FDR estimation, and thus a robust adaptive FDR control procedure.

3.1.3 Model Selection

We conduct simulation study 2 to evaluate how well AIC and BIC selected the number of components in the IMIX framework. We first generated 1 000 datasets following (4) for 16 scenarios that consisted of a combination of balanced and unbalanced mixing proportions of seven and eight components (Table S1). The unbalanced mixing proportions were based on the proportions of genes in the real-data example, we used the estimated $\hat{\pi}_k$ of the TCGA bladder cancer data set fitted by IMIX-Cor as the unbalanced proportions for the eight-component mixture model. The seven-component mixture model simulation simply eliminated the eighth component from the eight-component mixture model, i.e., genes that were associated with the outcome through all three data types. For each mixing proportion and number of components combination, we generated four scenarios with the mean and covariance parameters equal to the simulated parameters in simulation study 1 Scenarios 2-5. We fitted the IMIX framework without adding any constraint on the mean, assuming

models of one to eight components. One reason was that it was not necessary to impose the constraints on the mean for models with fewer components; another reason was that we were only interested in the final number of components of the selected model rather than the component each gene belongs to, for the purpose of model selection.

Fig 4 shows the number of components selected by AIC/BIC after averaging 1 000 simulation study for the unbalanced seven- and eight component simulated models. AIC selected the correct number of components for the simulation study in both balanced and unbalanced settings (Fig S5(a)-(d)). BIC performed similarly (Fig 4(a)(b)(d)), but it selected a more conservative number, i.e., a more parsimonious model, under the extreme unbalanced eight-component scenario in Fig 4(c). We consider this unbalanced eight-component setting very challenging for BIC or any model selection criterion since the smallest mixing proportion was only 4%. To further evaluate the ability of AIC and BIC to select the correct number of components when the mixing proportions were unbalanced, we conducted more simulation studies for eight-component multivariate Gaussian mixture model with varying levels of unbalanced settings as shown in Supplementary Material Section 2.3: both AIC and BIC performed well in identifying the correct number of components (Fig S6).

Together, AIC and BIC were both reliable model selection criteria under relatively balanced mixing proportions. AIC could select up to eight components in extremely unbalanced situations; however, previous studies (McLachlan and Peel, 2004; Steele and Raftery, 2009) have shown that AIC is prone to overestimating the number of components. We consider BIC to be more stable as it takes into account the number of genes in the penalty term, which can be as large as thousands under whole genome setting.

3.2 Real Data Applications

To demonstrate the proposed IMIX framework's versatility and efficiency in different disease outcomes, we applied our method to a binary outcome, the luminal and basal molecular subtypes of muscle-invasive bladder cancer, as well as a survival outcome for the prognosis

of pancreatic cancer in the TCGA dataset.

3.2.1 Molecular Subtypes of Bladder Cancer in the TCGA

Previous studies in bladder cancer identified molecular signatures associated with the pathological and clinical outcomes (Choi et al., 2014; Guo et al., 2019); in particular, those molecular subtypes have important implications for prognostication and treatment. Twenty-three gene expression markers have been reported to play a major role in these molecular subtypes. Here we applied IMIX to jointly analyze DNA methylation, gene expression, and CNV to investigate: (1) whether those gene expression markers also demonstrated difference at the DNA methylation and CNV levels, and (2) whether there were other genes associated with the molecular subtyping through any of the three data types. We analyzed the TCGA bladder cancer patient cohort that was profiled by three genomic platforms, DNA methylation, mRNA gene expression, and CNV. After quality control (Supplementary Materials, Section 3.1), we separately analyzed 373 DNA methylation samples, 391 RNA-Seq samples, and 387 CNV samples with $N = 15\,672$ genes with respect to the molecular subtypes adjusting for the clinical covariates, including age, sex, race, smoking status and pathologic stage. We applied IMIX, Bonferroni correction, and BH-FDR to the final summary statistics/Z-scores obtained from the association tests of individual-level data. The nominal error control level of Bonferroni correction and BH-FDR for separate analysis was set at $\alpha = 0.05$ and that of IMIX for integrative analysis was at $\alpha = 0.2$. We used IMIX-BIC to perform model selection, with the optimal model selected as IMIX-Cor-Twostep and the best number of components as eight based on BIC values. Table 1 shows the point estimates and 95% bootstrap-based confidence intervals ($B = 1\,000$) (McLachlan and Peel, 2004) for the parameters in the correlation matrices between DNA methylation, gene expression, and CNV. DNA methylation and gene expression were correlated for all components that involved non-null genes through at least one of these two data types: they were component 2 ($M+, E-, CNV-$), component 3 ($M-, E+, CNV-$), component 5 ($M+, E+, CNV-$), and component 8 ($M+, E+, CNV+$).

In particular, component 5 and component 8 showed moderate correlations between DNA methylation and gene expression where genes showed significant associations through both data types (M+ and E+). Another interesting finding was that the two data types DNA methylation and gene expression were correlated when genes were associated with the outcome through only one datatype as reflected in component 2 (M+ and E-) and component 3 (M- and E+). This also held for the correlations between gene expression and CNV in component 1 (E-,CNV-) and component 3 (E+,CNV-). These results further supported the IMIX model assumptions that the data types were correlated, both under the alternative and the null hypothesis, thus reinforcing that the IMIX method was effective by assuming multivariate distributions in all components instead of the commonly adopted conditional independence under the null hypothesis.

We compared the number of genes discovered in component 8 using BH-FDR, Bonferroni correction, and our method (Fig S7(a)). The genes that were detected by Bonferroni correction were identified by both our method and BH-FDR. The genes detected by IMIX had an overlap of 146 genes with the BH-FDR and included 116 new genes not discovered by either BH-FDR or Bonferroni correction. The estimated \widehat{mFDR}_8 of IMIX was 0.1995, close to the prespecified across-data-type FDR control level $\alpha=0.2$. Through simulation studies in Section 3.1.1, our method was more effective in controlling the across-data-type FDR compared with other methods. We also showed the levels of DNA methylation, gene expression, and CNV for the significant genes in component 8, i.e., genes that were associated with all the three data types in Fig 2(a); for the purpose of illustration, we only included the 61 significant genes after adaptive FDR control at $\alpha = 0.01$. We conducted Ingenuity Pathway Analysis (IPA, Ingenuity Systems (www.ingenuity.com)) on the 61 significant genes in component 8 at $\alpha = 0.01$. The results showed strong peroxisome proliferator activator receptor (PPAR) pathway activation (Fig S9) in luminal samples. This pathway was previously reported by Choi et al. (2014), who first proposed the molecular subtypes of muscle-invasive bladder cancer, that PPAR α and PPAR γ activation played essential roles in

regulating gene expression signature for the luminal subtype. Specifically, they exposed the PPAR γ -selective agonist rosiglitazone in two bladder cancer cell lines and further confirmed that rosiglitazone activated PPAR pathway and enriched gene signatures in primary luminal samples. Furthermore, we estimated the causal relationships between DNA methylation, gene expression, and CNV of the 61 genes in component 8 by applying Bayesian networks (Scutari, 2017) with the target nominal type I error rate at 0.01. Among which, 51 genes showed significant dependent structures between the three data types. The directed acyclic graphs (DAGs) based on conditional independence tests with a restriction of causal direction from CNV to E showed six different patterns of causal structures (Supplementary Materials, Section 3.2).

We present the levels of the luminal/basal markers for DNA methylation, gene expression, and CNV in Fig 2(b). Among the 23 markers, we found that six, fifteen, and one gene belonged to component 3 (M-,E+,CNV-), component 5 (M+,E+,CNV-), and component 8 (M+,E+,CNV+), respectively. In particular, PPARG belonged to component 8, i.e, associated with the subtypes via all three molecular mechanisms. Bayesian networks further confirmed that PPARG had a full model with dependence structures of CNV \rightarrow E, E \rightarrow M, M \rightarrow CNV (Fig S8(1)). This gene was reported to be one of the driver genes for the basal/luminal differentiation. As expected, PPARG showed higher gene expression level in the luminal samples than in the basal samples; furthermore, we discovered a concordant significant differential pattern in the methylation and CNV levels that have not been previously reported.

In summary, our analysis revealed that the luminal/basal markers demonstrated substantial differences in at least two data types (Fig 2(b)). By applying the IMIX framework, we successfully discovered novel genes that were associated with the molecular subtypes through all three data types (Fig 2(a)) and confirmed the PPAR/RXR activation canonical pathway that was previously reported to play a central role in luminal/basal differentiation (Choi et al., 2014).

3.2.2 Prognosis of Pancreatic Cancer in the TCGA

We further applied IMIX to a survival outcome to investigate the relationships between the prognosis of pancreatic cancer patients and two genomic datasets, gene expression and CNV in the TCGA. After quality control (Supplementary Materials, Section 3.1), we first applied the Cox proportional hazards model to each of the 15 472 genes respectively on 157 RNA-Seq samples and 161 CNV samples adjusting for age, gender, and smoking status. Next, we fitted IMIX, BH-FDR, and Bonferroni corrections on the summary statistics. After model selection based on BIC, IMIX-Cor-Twostep fitted the best. Table 2 shows the point estimates and 95% bootstrap-based confidence intervals ($B = 1000$) of the parameters in the correlation matrices between gene expression and CNV. In component 4 (E+,CNV+), where the detected genes were significantly associated with survival outcomes through both gene expression and CNV, the correlation between gene expression and CNV was $\hat{\rho} = 0.120$ with 95% confidence interval (0.071, 0.18). To assess the effect of the detected 104 genes in component 4 at $\alpha = 0.05$, we used iCluster (Shen et al., 2009) to group the patients based on gene expression and CNV data into two classes, here we only applied the 104 genes detected by IMIX with no feature selection in the clustering process. Fig 3(a) shows the Kaplan-Meier(KM) curve of the overall survival of the pancreatic cancer patients. The log-rank test resulted in $p = 0.016$, and the Cox model adjusting for patient pathologic stages resulted in $p = 0.04$. Furthermore, we also clustered the patients using the 991 genes discovered at adaptive FDR $\alpha = 0.2$, all patients but one were grouped into the same clusters as using the 104 genes at $\alpha = 0.05$; the KM curve returned the same results. This indicates that IMIX was able to capture the most important features and a controlled number of false discovered genes at $\alpha = 0.05$. Fig 3(b) shows the gene expression and CNV levels of the identified 104 genes associated with the pancreatic cancer prognosis. The gene expression and CNV were positively correlated as shown in the heatmap. We compared the results of IMIX, Bonferroni correction, and BH-FDR for component 4, i.e., genes that were associated with the survival outcomes through both gene expression and CNV. Bonferroni correction was not able to

discover any significant genes at the nominal level $\alpha = 0.05$. IMIX detected 104 genes at $\alpha = 0.05$ with an estimated $\widehat{\text{mFDR}}_4 = 0.0498$. BH-FDR detected 271 genes at $\alpha = 0.05$. IMIX identified fewer genes but it captured the important features as evidenced by the KM analysis/log-rank test with a controlled across-data-type FDR compared with BH-FDR. We showed in Section 3.1.1 that BH-FDR failed to control for FDR under the data integration settings. In addition, IMIX detected 991 genes at $\alpha = 0.2$ with an estimated $\widehat{\text{mFDR}}_4 = 0.2$; and the 271 genes detected by BH-FDR ($\alpha = 0.05$) were all included in the genes discovered by our method as shown in the Venn diagram (Fig S7(b)).

4 Discussion

We have proposed IMIX, a multivariate mixture model framework based on summary statistics for integrative genomic association analysis. Our model incorporates the correlation structures between different genomic datasets by assuming multivariate Gaussian mixture distribution of the Z scores (transformed from p-values) from regression analysis of individual-level data. The IMIX framework includes four models: IMIX-Cor, IMIX-Ind, IMIX-Cor-Restrict, and IMIX-Cor-Twostep, each of which best captures a specific type of data correlation structure arising from various data analysis problems. IMIX selects the optimal model based on AIC/BIC values among the four models. In addition, it features simultaneous model selection for the number of underlying latent states/components of the optimal mixture model with a specific correlation structure. We utilize the EM algorithm in parameter estimation, and the mixture model naturally produces the local FDR for each gene, which is easily derived from the posterior probability. Our model features an adaptive procedure to control the across-data-type FDR, where we take into account both the multiple testing of the gene, and the multiple data types under an integrative analysis setting. To our knowledge, we are the first to demonstrate such error-control property of integrative genomic models.

Our applications to the two TCGA data sets demonstrate that different genomic data types, such as DNA methylation, mRNA gene expression, and CNVs, can be correlated in both null and non-null genes, as shown in the bootstrap-based confidence intervals (Table 1 and 2). Therefore, it is necessary to consider the inter-source correlations of multiple datasets in integrative analysis. Based on simulation studies under various settings of correlation structures, including the one based on TCGA bladder cancer dataset, IMIX controlled the FDR precisely and yielded better statistical power compared with the independent separate analysis models, including BH-FDR, Bonferroni correction, q-value, and the local FDR procedure.

An advantage of our proposed method using summary statistics is that our method does not require the use of a common set of samples, which relaxes the conditions of previously published methods for the integration of multiple omics data (Richardson et al., 2016). In addition, IMIX is able to model summary statistics from independent or partially overlapped cohorts, as illustrated in the TCGA data examples in Section 3.2. The implementation of IMIX employs the EM algorithm, which in general converges fast, leading to great computational efficiency. One unique feature of IMIX is its constraint on the mixture component means, which is not only more biologically plausible than unconstrained means, but has also ensured model identifiability, a common challenge in mixture models. Another unique feature of IMIX is model selection: we let the data decide the number of mixture components and the correlation structure based on AIC or BIC.

We compare our work with a recently proposed method called Primo for quantitative trait loci (QTLs) mapping based on genome-wide association study (GWAS) summary statistics (Gleason et al., 2019). The two methods share a similar concept in utilizing the mixture model in data integration; however, there are several key differences. The main difference is how we approach the parameter estimations. Primo estimates the mixture model parameters by first assuming conditional independence between the data types and estimating the marginal null and alternative distributions for each data type with a fixed proportion of

non-null tests, and then approximating the correlation matrices under certain assumptions. In contrast, IMIX directly estimates the multivariate mixture model parameters, including means, covariance matrices and mixing proportions, using the EM algorithm, which allows examining and relaxing the conditional independence assumption. Furthermore, IMIX accommodates simultaneous model selection for both number of mixture components and the correlation structure, which is absent in Primo.

IMIX is a useful and versatile tool that can study various types of outcomes, including continuous, binary, and time-to-event outcomes in integrative genomic analysis. We have applied IMIX to two types of problems, the survival prognosis of pancreatic cancer and the luminal/basal molecular subtypes of bladder cancer, both providing novel biological insights. IMIX framework is not only applicable to cancer genomics, but also to other complex diseases and traits as afforded by ongoing large-scale multiple-omics projects, such as the NIH Trans-Omics for Precision Medicine (TOPMed) project (Brody et al., 2017) and the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium (Psaty et al., 2009), consisting of over 100,000 deeply phenotyped and sequenced individuals with multiple types of omics data, such as transcriptomic, epigenomic, metabolic, proteomic, and whole-genome sequencing data. This work, therefore, has a wide range of application potential to provide novel biological insights into disease mechanisms. We have implemented the integration model for two and three genomics data types in the simulation studies and data applications, which could be further generalized to four and more data types in the multivariate mixture model framework. We leave the details of this potential extension for future research. While we have relaxed the conditional independence assumptions for the data types in IMIX, we could further extend our method by assessing the correlations between genes within each data type, which is another important direction for future work.

We have implemented the proposed method in an R package “IMIX”, which is available at <https://github.com/ziqiaow/IMIX> and will be posted to R/CRAN.

Acknowledgements

This research was supported by the National Institutes of Health (NIH) grants R01HL116720 and R01CA169122; P.W. was supported by NIH grant P50CA091846. The results shown here are in whole or part based upon data generated by the TCGA Research Network: <https://www.cancer.gov/tcga>. We thank Ms. Sarah Bronson in Scientific Publications, Research Medical Library, The University of Texas MD Anderson Cancer Center, for editorial assistance. The authors declare no conflict of interest.

References

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* 57(1), 289–300.

Brody, J. A., A. C. Morrison, J. C. Bis, J. R. O’Connell, M. R. Brown, J. E. Huffman, D. C. Ames, A. Carroll, M. P. Conomos, S. Gabriel, et al. (2017). Analysis commons, a team approach to discovery in a big-data environment for genetic epidemiology. *Nature genetics* 49(11), 1560–1563.

Choi, W., S. Porten, S. Kim, D. Willis, E. R. Plimack, J. Hoffman-Censits, B. Roth, T. Cheng, M. Tran, I.-L. Lee, et al. (2014). Identification of distinct basal and luminal subtypes of muscle-invasive bladder cancer with different sensitivities to frontline chemotherapy. *Cancer cell* 25(2), 152–165.

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

Efron, B. et al. (2007). Size, power and false discovery rates. *The Annals of Statistics* 35(4), 1351–1377.

Genovese, C. and L. Wasserman (2002). Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64(3), 499–517.

Gleason, K. J., F. Yang, B. L. Pierce, X. He, and L. S. Chen (2019). Primo: integration of multiple gwas and omics qtl summary statistics for elucidation of molecular mechanisms of trait-associated snps and detection of pleiotropy in complex traits. *bioRxiv*, 579581.

Guo, C. C., T. Majewski, L. Zhang, H. Yao, J. Bondaruk, Y. Wang, S. Zhang, Z. Wang, J. G. Lee, S. Lee, et al. (2019). Dysregulation of emt drives the progression to clinically aggressive sarcomatoid bladder cancer. *Cell reports* 27(6), 1781–1793.

Kirk, P., J. E. Griffin, R. S. Savage, Z. Ghahramani, and D. L. Wild (2012). Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics* 28(24), 3290–3297.

Leroux, B. G. (1992). Consistent estimation of a mixing distribution. *The Annals of Statistics*, 1350–1360.

McLachlan, G. J., S.-K. Ng, and R. Bean (2006). Robust cluster analysis via mixture models. *Austrian Journal of Statistics* 35(2&3), 157–174.

McLachlan, G. J. and D. Peel (2004). *Finite mixture models*. John Wiley & Sons.

Mo, Q., S. Wang, V. E. Seshan, A. B. Olshen, N. Schultz, C. Sander, R. S. Powers, M. Ladanyi, and R. Shen (2013). Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proceedings of the National Academy of Sciences* 110(11), 4245–4250.

Newton, M. A., A. Noueiry, D. Sarkar, and P. Ahlquist (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* 5(2), 155–176.

Pineda, S., F. X. Real, M. Kogevinas, A. Carrato, S. J. Chanock, N. Malats, and K. Van Steen (2015). Integration analysis of three omics data using penalized regression methods: an application to bladder cancer. *PLoS genetics* 11(12).

Psaty, B. M., C. J. O'Donnell, V. Gudnason, K. L. Lunetta, A. R. Folsom, J. I. Rotter, A. G. Uitterlinden, T. B. Harris, J. C. Witteman, and E. Boerwinkle (2009). Cohorts for heart and aging research in genomic epidemiology (charge) consortium: Design of prospective meta-analyses of genome-wide association studies from 5 cohorts. *Circulation: Cardiovascular Genetics* 2(1), 73–80.

Richardson, S., G. C. Tseng, and W. Sun (2016). Statistical methods in integrative genomics. *Annual Review of Statistics and Its Application* 3(1), 181–209.

Scutari, M. (2017). Bayesian network constraint-based structure learning algorithms: Parallel and optimized implementations in the bnlearn r package. *Journal of Statistical Software, Articles* 77(2), 1–20.

Shen, R., A. B. Olshen, and M. Ladanyi (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* 25(22), 2906–2912.

Steele, R. J. and A. E. Raftery (2009). Performance of bayesian model selection criteria for gaussian mixture models. Technical Report No. 559, Department of Statistics, University of Washington.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64(3), 479–498.

Storey, J. D., A. J. Bass, A. Dabney, and D. Robinson (2019). *qvalue: Q-value estimation for false discovery rate control*. R package version 2.16.0.

Sun, W., P. Bunn, C. Jin, P. Little, V. Zhabotynsky, C. M. Perou, D. N. Hayes, M. Chen, and D.-Y. Lin (2018). The association between copy number aberration, dna methylation and gene expression in tumor samples. *Nucleic acids research* 46(6), 3009–3018.

Sun, W. and T. T. Cai (2007). Oracle and adaptive compound decision rules for false discovery rate control. *Journal of the American Statistical Association* 102(479), 901–912.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58(1), 267–288.

Wei, P. and W. Pan (2008). Incorporating gene networks into statistical tests for genomic data via a spatially correlated mixture model. *Bioinformatics* 24(3), 404–411.

Wei, P. and W. Pan (2012). Bayesian joint modeling of multiple gene networks and diverse genomic data to identify target genes of a transcription factor. *The annals of applied statistics* 6(1), 334.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)* 67(2), 301–320.

Figures and Tables

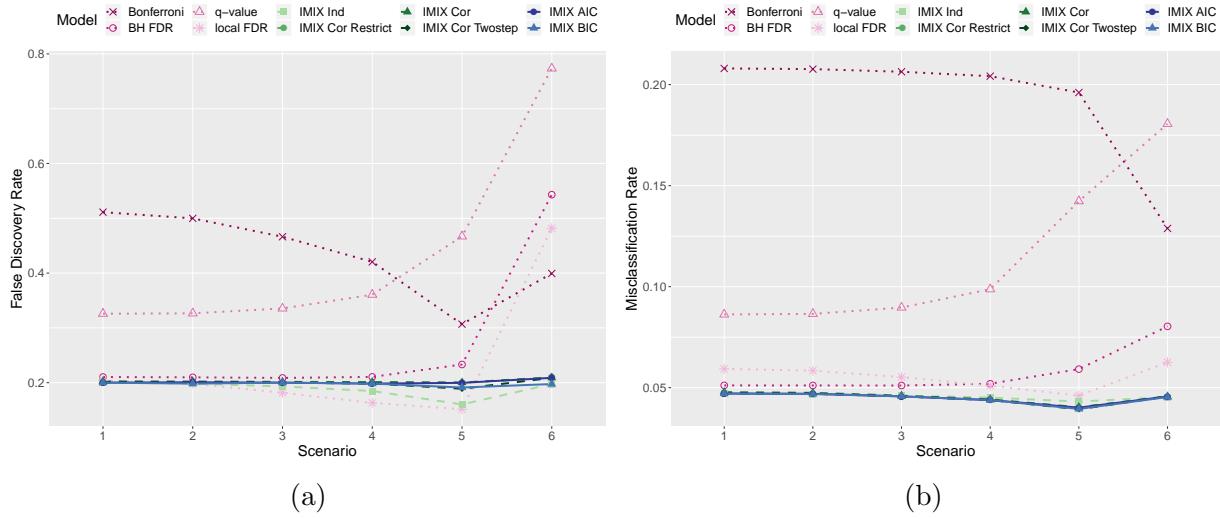


Figure 1: Comparison of IMIX-Ind, IMIX-Cor, IMIX-Cor-Restrict, IMIX-Cor-Twostep, IMIX-AIC, IMIX-BIC, BH-FDR, Bonferroni correction, q-value, and the local FDR procedure at $\alpha = 0.2$: (a) across-data-type FDR control, the results are the average of 1000 simulations of the average of components 2-8, excluding the global null component 1. (b) misclassification rate, the results are the average of 1000 simulations of the average of all components.

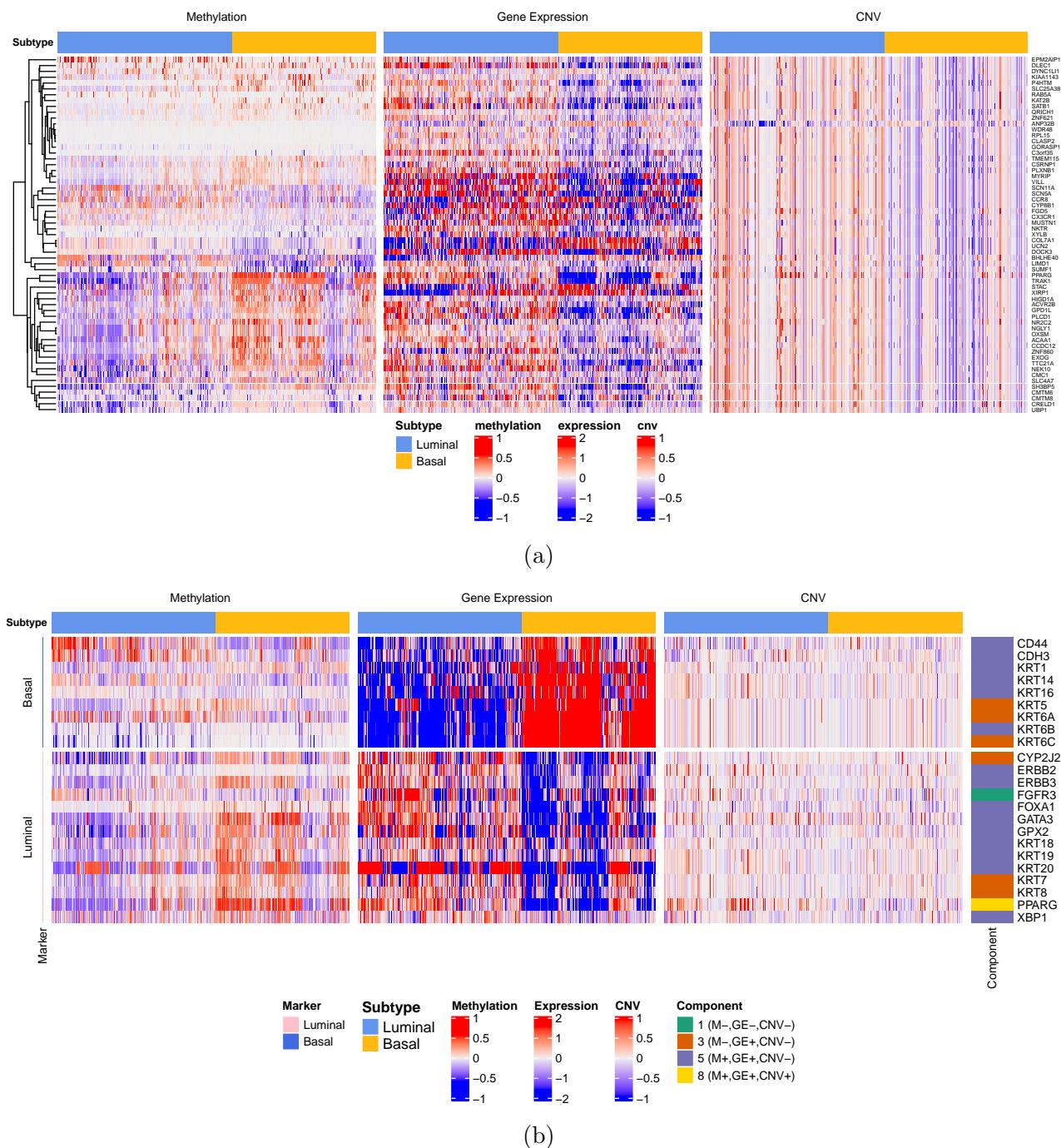
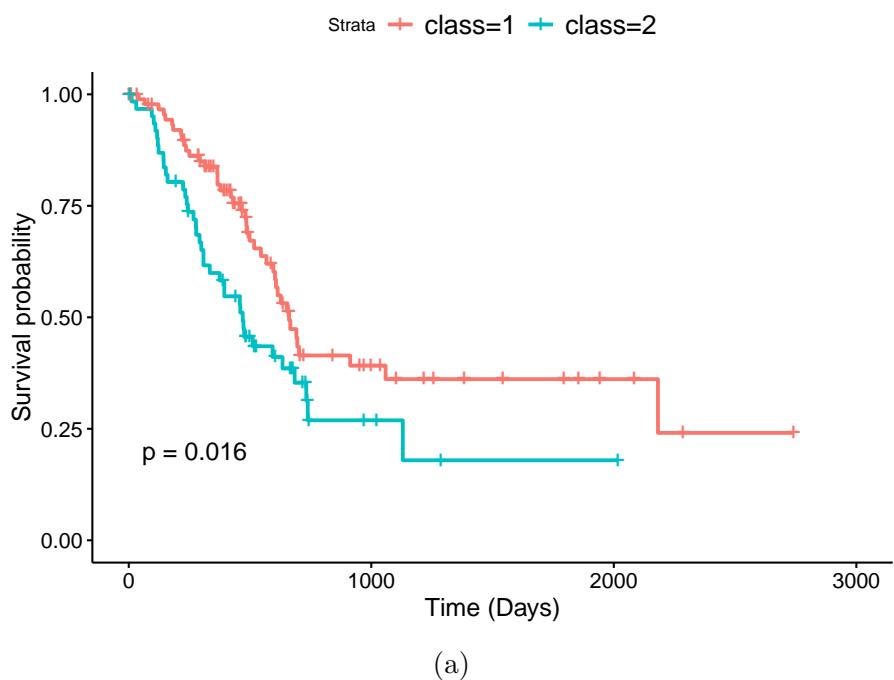
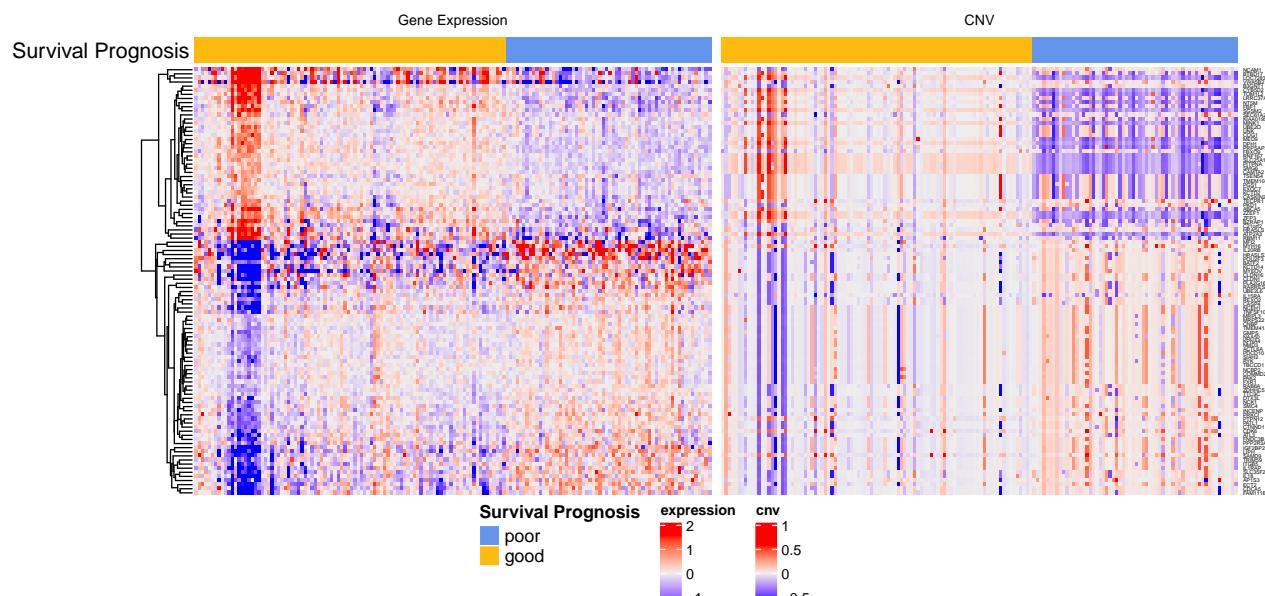


Figure 2: Heatmaps of genes in IMIX analysis for bladder cancer molecular subtypes in the TCGA. (a) Methylation, gene expression and copy number variation (CNV) patterns of top significant genes associated with the three data types (M+, GE+, CNV+) identified by IMIX in molecular subtypes of The Cancer Genome Atlas (TCGA) muscle-invasive bladder cancer patients, with adaptive false discovery rate (FDR) control at $\alpha = 0.01$, estimated marginal FDR ($\widehat{\text{mFDR}}_8 = 0.0098$). (b) Expression patterns of luminal/basal markers of TCGA bladder cancer cohort.



(a)



(b)

Figure 3: Results of IMIX analysis for pancreatic cancer prognosis in the TCGA. (a) Kaplan-Meier curves for pancreatic cancer patient survival in The Cancer Genome Atlas (TCGA). Samples were clustered based on the 104 genes identified by IMIX, with adaptive false discovery rate (FDR) control at $\alpha = 0.05$, estimated marginal FDR ($\widehat{\text{mFDR}}_4 = 0.0498$). (b) Expression patterns of the 104 genes selected by IMIX that were associated with the prognosis of the TCGA pancreatic cancer patients.

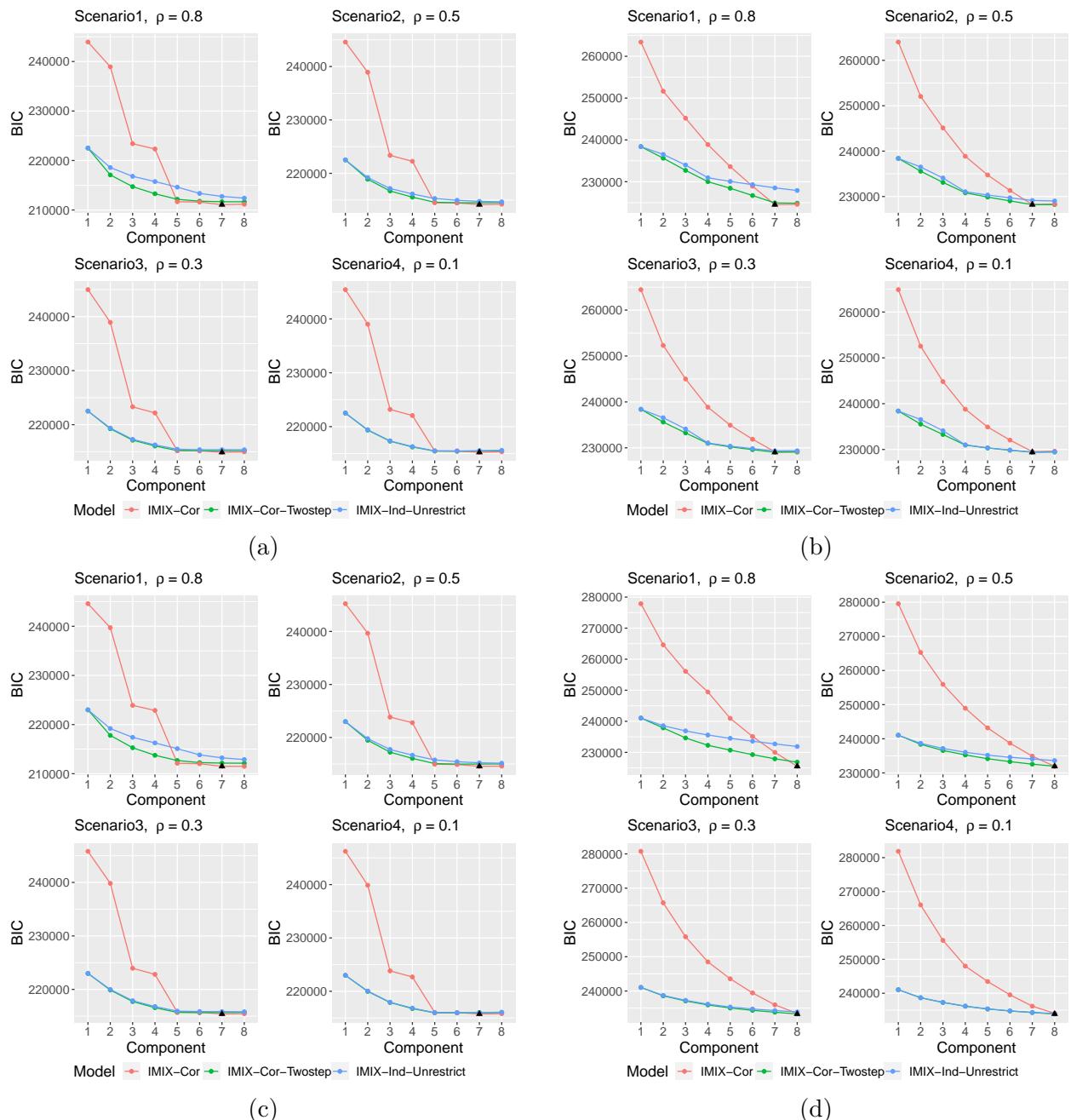


Figure 4: Simulation study on selecting the number of mixture components using BIC. (a) Unbalanced setting, 7-component mixture model; (b) Balanced setting, 7-component mixture model; (c) Unbalanced setting, 8-component mixture model; (d) Balanced setting, 8-component mixture model. Black triangle represents the model BIC selects. ρ is the correlation between data types.

Component	M & E	M & CNV	E & CNV
1 (M-,E-,CNV-)	0.015 (-0.039,0.072)	0.091 (0.037,0.14)	-0.016 (-0.070,0.043)
2 (M+,E-,CNV-)	0.070 (0.014,0.12)	-0.012 (-0.060,0.031)	0.0049 (-0.064,0.075)
3 (M-,E+,CNV-)	0.14 (0.051,0.23)	0.090 (0.016,0.17)	-0.10 (-0.19,-0.022)
4 (M-,E-,CNV+)	-0.099 (-0.59,0.40)	0.35 (-0.085,0.73)	-0.059 (-0.53,0.36)
5 (M+,E+,CNV-)	0.25 (0.20,0.31)	-0.037 (-0.082,0.013)	0.038 (-0.023,0.097)
6 (M+,E-,CNV+)	0.038 (-0.22,0.25)	-0.037 (-0.27,0.22)	-0.088 (-0.40,0.26)
7 (M-,E+,CNV+)	0.21 (-0.16,0.56)	0.13 (-0.27,0.49)	0.12 (-0.29,0.50)
8 (M+,E+,CNV+)	0.19 (0.021,0.38)	0.10 (-0.049,0.26)	0.080 (-0.13,0.28)

Table 1: Estimated correlations between the transformed z scores (from p value) of DNA methylation (M) and gene expression (E), DNA methylation and CNV, gene expression and CNV with 95% bootstrap-based confidence intervals (B = 1 000) for TCGA bladder cancer data integration analysis by IMIX-BIC.

Component	E & CNV
1 (E-,CNV-)	0.040 (-0.047,0.13)
2 (E+,CNV-)	0.11 (-0.012,0.21)
3 (E-,CNV+)	0.016 (-0.019,0.049)
4 (E+,CNV+)	0.12 (0.071,0.18)

Table 2: Estimated correlations between the transformed z scores (from p value) of gene expression (E) and CNV with 95% bootstrap-based confidence intervals (B = 1 000) for TCGA pancreatic cancer data integration analysis by IMIX-BIC.