METHODS
published: 23 April 2021
doi: 10.3389/fgene.2021.665416

CBP-JMF: An Improved Joint Matrix Tri-Factorization Method for Characterizing Complex Biological Processes of Diseases

Bingbo Wang[1]*, **Xiujuan Ma**[1], **Minghui Xie**[1], **Yue Wu**[1], **Yajun Wang**[2], **Ran Duan**[1], **Chenxing Zhang**[1], **Liang Yu**[1], **Xingli Guo**[1] **and Lin Gao**[1]*

[1] *School of Computer Science and Technology, Xidian University, Xi'an, China,* [2] *School of Humanities and Foreign Languages, Xi'an University of Technology, Xi'an, China*

Multi-omics molecules regulate complex biological processes (CBPs), which reflect the activities of various molecules in living organisms. Meanwhile, the applications to represent disease subtypes and cell types have created an urgent need for sample grouping and associated CBP-inferring tools. In this paper, we present CBP-JMF, a practical tool primarily for discovering CBPs, which underlie sample groups as disease subtypes in applications. Differently from existing methods, CBP-JMF is based on a joint non-negative matrix tri-factorization framework and is implemented in Python. As a pragmatic application, we apply CBP-JMF to identify CBPs for four subtypes of breast cancer. The result shows significant overlapping between genes extracted from CBPs and known subtype pathways. We verify the effectiveness of our tool in detecting CBPs that interpret subtypes of disease.

Keywords: non-negative matrix factorization, complex biological processes, multi-dimensional genomic data, disease, subtype

OPEN ACCESS

Edited by:
Jianing Xi,
Northwestern Polytechnical University, China

Reviewed by:
Hao Wu,
Shandong University, China
Feng Li,
Qufu Normal University, China
Kai Shi,
Guilin University of Technology, China

Correspondence:
Bingbo Wang
bingbowang@xidian.edu.cn
Lin Gao
lgao@mail.xidian.edu.cn

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *08 February 2021*
Accepted: *01 March 2021*
Published: *23 April 2021*

Citation:
Wang B, Ma X, Xie M, Wu Y, Wang Y, Duan R, Zhang C, Yu L, Guo X and Gao L (2021) CBP-JMF: An Improved Joint Matrix Tri-Factorization Method for Characterizing Complex Biological Processes of Diseases. Front. Genet. 12:665416. doi: 10.3389/fgene.2021.665416

INTRODUCTION

Complex biological processes (CBPs) are the coordinated effect of multiple molecules, which result in some functional pathways and the vital processes occurring in living organisms. In addition, the vast amounts of multi-omics data, such as genomics, epigenomics, transcriptomics, proteomics, and metabolomics, can be integrated to understand systems biology accurately (Suravajhala et al., 2016). Hasin et al. (2017) pointed out that a deeper and better understanding of important biological processes and modules can be obtained through multi-omics studies. However, practical tools are still missing to integrate diverse multi-omics data at different biological levels and reveal the CBPs and other problems like the causes of diseases.

Non-negative matrix factorization (NMF) (Lee and Seung, 1999) is a powerful tool for dimension reduction and feature extraction. It has been increasingly applied to diverse fields, including bioinformatics (*e.g.*, high-dimensional genomic data analysis). For example, Brunet et al. (2004) applied NMF and consensus clustering to the gene expression data of leukemia to discover metagenes and molecular patterns. Xi et al. (2018) detected driver genes from pan-cancer data based on another matrix decomposition framework called matrix tri-factorization. Up to now, several variants of NMF have been proposed, including tri-factorization NMF (Ding et al., 2006), graph-regularized NMF (Cai et al., 2011), joint NMF (Zhang et al., 2012), iNMF (Yang and Michailidis, 2016), *etc*. (more details are in **Supplementary Note 1** of the **Supplementary Materials**). In 2012, jNMF (Zhang et al., 2012) was proposed to identify multi-omics modules by integrating cancer's DNA methylation data, gene expression data, and miRNA expression data. Chen and Zhang (2018) applied joint matrix tri-factorization to discover two-level modular organization from matched genes and miRNA expression data, gene expression data, and drug response data.

Omics data across the same samples contain signal values from expression counts, methylation levels, and protein concentrations, which control biological systems, resulting in so-called multi-dimensional genomic (MG) data. The natural representation of these diverse MG data is a series of matrices with measured values in rows and individual samples in columns. Recently, there are integrative analysis tools based on NMF technique that reveal low-dimensional structure patterns. The low-dimensional structure patterns reflect CBPs and sample groups while preserving as much information as possible from high-dimensional MG data (Stein-O'Brien et al., 2018).

In general, most particular matrix factorization techniques are being developed to enhance their applicability to specific biological problems. Meanwhile, the applications to represent disease subtypes (Biton et al., 2014) and cell types (Fan et al., 2016) have created an urgent need for sample grouping and associated CBP-inferring tools. Moreover, cancer and other complex diseases are heterogeneous, *i.e.*, there are various subgroups for a cancer or a complex disease. The study of the heterogeneity of cancer and complex diseases will help us understand the disease further and provide better opportunities to disease treatment (Xi et al., 2020). To address this issue, we extend traditional jNMF and develop CBP-JMF, an improved joint matrix tri-factorization framework for characterizing CBPs that represent sample groups, and implement a Python package. This package takes labeled samples as the prior information and integrates MG data (*e.g.*, copy number variation, gene expression, microRNA expression, and/or molecule interaction network) to identify the underlying CBPs which characterize the specific functional properties of each group. CBP-JMF can be used to mark unlabeled samples with groups of known labels. For ease of use, CBP-JMF can recommend reasonable parameter settings for users. CBPs found by CBP-JMF are connected network markers, and they are distinguished between sample groups. These markers usually have specific biological functions and play important roles in phenotypes. As an example, CBPs for subtypes of breast cancer are obtained by CBP-JMF, but they may not have been collected in any reference database yet.

The rest of this paper is organized as follows. Section "Framework of CBP-JMF" deals with the problem formulation of CBP-JMF and the implementation of it. Then, Section "Results" exemplifies our approach by applying CBP-JMF to identify CBPs for different subtypes of breast cancers and compares the results of classifying unlabeled samples with CBP-JMF and its several variants. Finally, Section "Discussion" discusses our results and lists our expectations of our method and the limitations of it. Section "Conclusions" highlights our method.

FRAMEWORK OF CBP-JMF

Problem Definition

Given a non-negative matrix $\mathbf{X} \in \mathbf{R}^{m\times n}$, it can be factorized into three non-negative matrix factors based on matrix tri-factorization: $\mathbf{X} \approx \mathbf{USV}$, where $\mathbf{U} \in \mathbf{R}^{m\times k}$, $\mathbf{S} \in \mathbf{R}^{k\times k}$, and $\mathbf{V} \in \mathbf{R}^{k\times n}$. Factored matrix $\mathbf{S}$ cannot only absorb scale difference between $\mathbf{U}$ and $\mathbf{V}$ but also indicates relationships between the identified k modules.

In CBP-JMF, given a MG dataset composed of P omics, it can be presented by multiple matrices $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, ..., \mathbf{X}^{(P)}$, as illustrated in **Figure 1**. For each matrix, the rows indicate molecules like genes, and the columns indicate samples; the values in it are related to the meaning of omics. If $\mathbf{X}^{(p)}$ ($p \in [1, P]$) is a matrix of gene expression data, $\mathbf{X}_{ij}^{(p)}$ represents the expression value of the gene in the i-th row on the j-th sample. Basically, each non-negative matrix $\mathbf{X}^{(p)} \in \mathbf{R}^{m\times n}$, $p = 1, 2, ..., P$ is factorized into three non-negative matrix factors based on matrix tri-factorization: $\mathbf{X}^{(p)} \approx \mathbf{U}^{(p)}\mathbf{S}^{(p)}\mathbf{V}$, where molecular coefficient matrix (MCM) $\mathbf{U}^{(p)} \in \mathbf{R}^{m\times k}$ and sample basis matrix (SBM) $\mathbf{V} \in \mathbf{R}^{k\times n}$ are the pattern indicator matrices of k CBPs and k sample groups, respectively. Scale absorbing matrix (SAM) $\mathbf{S}^{(p)} \in \mathbf{R}^{k\times k}$ explores the relationships between them. Furthermore, MCM describes the structure pattern between molecules (*e.g.*, genes), SBM indicates the structure pattern between samples, and SAM absorbs the difference of scales between MCM and SBM (**Figure 1**). Each column of the MCM infers a latent feature associated with a CBP, and the continuous values in it represent the relative contribution of each molecule in the CBP. Meanwhile, each row of the SBM describes the relative contributions of the samples to a latent feature. The sample groups can be detected by comparing the relative weights in each row of the SBM.

Overall, $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, ..., \mathbf{X}^{(P)}$ can be jointly factorized into specific $\mathbf{U}^{(1)}, \mathbf{U}^{(2)}, ..., \mathbf{U}^{(P)}$, $\mathbf{S}^{(1)}, \mathbf{S}^{(2)}, ..., \mathbf{S}^{(P)}$, and a common matrix $\mathbf{V}$. $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, ..., \mathbf{X}^{(P)}$ are across the same samples, and $\mathbf{V}$ reveals consistent sample groups of multi-omics data. In CBP-JMF, $\mathbf{V}$ can be divided into $\mathbf{V}^{L}$ and $\mathbf{V}^{UL}$ according to input data, where L and UL mean "labeled" samples and "unlabeled" samples, respectively.

Objective Function of CBP-JMF

Considering that different datasets may play different roles in data integration, we adopted a method that can learn the weights of different input data through a weighted joint tri-NMF:

$$\begin{aligned} &\min \sum_{p=1}^{P} \pi^{(p)} \left\| \mathbf{X}^{(p)} - \mathbf{U}^{(p)}\mathbf{S}^{(p)}\mathbf{V} \right\|_F^2 + \omega \| \Pi \|^2 \\ &s.t.\ \pi^{(p)} > 0, \sum_{p=1}^{P} \pi^{(p)} = 1 \end{aligned} \tag{1}$$

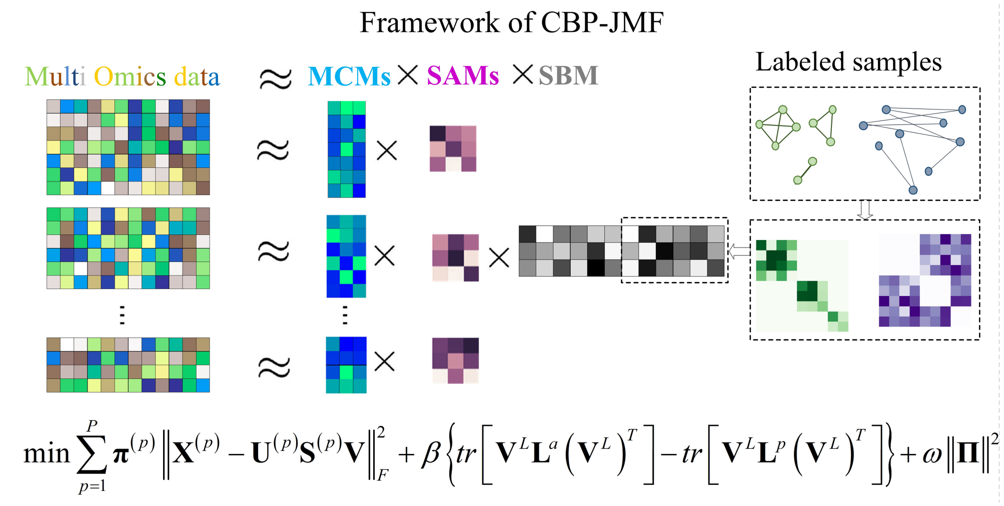
Framework of CBP-JMF
Multi Omics data
≈ MCMs × SAMs × SBM
Labeled samples
$\min\sum_{p=1}^{P}\pi^{(p)}\left\|\mathbf{X}^{(p)}-\mathbf{U}^{(p)}\mathbf{S}^{(p)}\mathbf{V}\right\|_F^2+\beta\left\{tr\left[\mathbf{V}^L\mathbf{L}^a\left(\mathbf{V}^L\right)^T\right]-tr\left[\mathbf{V}^L\mathbf{L}^p\left(\mathbf{V}^L\right)^T\right]\right\}+\omega\left\|\mathbf{\Pi}\right\|^2$

FIGURE 1 | Illustration of the framework and optimization objective function of complex biological processes–joint matrix tri-factorization.

where $\Pi = \left(\pi^{(1)}, \pi^{(2)}, ..., \pi^{(P)}\right)$. CBP-JMF differentiates the importance of datasets by the weight constraint $\|\Pi\|^2$, and $\pi^{(p)}$ will get a weight to represent the contribution of data $\mathbf{X}^{(p)}$ to objective function after optimization. If $\mathbf{X}^{(p)}$ contributes to the optimization of cost function, then it will be given a higher weight $\pi^{(p)}$, or if $\mathbf{X}^{(p)}$ contains lots of noises which hinder the optimization of objective function, it will be given a lower weight $\pi^{(p)}$.

In addition, $\mathbf{V}$ can be divided into labeled $\mathbf{V}^L$ and unlabeled $\mathbf{V}^{UL}$ parts according to the labeled samples and unlabeled samples. In order to learn the correlation between labeled samples, we use a graph Laplacian to represent the distance of labeled sample in latent space (Guan et al., 2015). We use Equations (2) and (3) to denote the distance between labeled samples from the same class and different class in the learned latent space, respectively,

$$\sum_{i=1}^{N^L}\sum_{j=1}^{N^L}\mathbf{W}^a_{ij}\left\|\mathbf{v}^L_i-\mathbf{v}^L_j\right\|_2^2 = tr\left[\mathbf{V}^L\mathbf{L}^a(\mathbf{V}^L)^T\right] \quad (2)$$

$$\sum_{i=1}^{N^L}\sum_{j=1}^{N^L}\mathbf{W}^p_{ij}\left\|\mathbf{v}^L_i-\mathbf{v}^L_j\right\|_2^2 = tr\left[\mathbf{V}^L\mathbf{L}^p(\mathbf{V}^L)^T\right] \quad (3)$$

where N^L is the number of labeled samples in $\mathbf{V}$, and $\mathbf{W}^a$ ($\mathbf{W}^{affinity}$) and $\mathbf{W}^p$ ($\mathbf{W}^{penalty}$) are the weighted adjacency matrices (see **Supplementary Note 2** in SM) corresponding to intra-group and inter-group samples respectively. $\mathbf{L}^a$ ($\mathbf{L}^{affinity}$) and $\mathbf{L}^p$ ($\mathbf{L}^{penalty}$) are the Laplacian matrix of $\mathbf{W}^a$ and $\mathbf{W}^p$, respectively, where $\mathbf{L}^a$=$\mathbf{D}^a - \mathbf{W}^a$, $\mathbf{L}^p$=$\mathbf{D}^p - \mathbf{W}^p$, $\mathbf{D}^a$=$\sum_{j=1}^{N^L}\mathbf{W}^a_{ij}$. In machine learning, people try to make samples from the same class near each other in the learned latent space and samples from different class far from each other. This principle can be written as

$$\min\left(tr\left[\mathbf{V}^L\mathbf{L}^a(\mathbf{V}^L)^T\right] - tr\left[\mathbf{V}^L\mathbf{L}^p(\mathbf{V}^L)^T\right]\right) \quad (4)$$

Combining weighted joint tri-NMF and the constraints of correlation between labeled samples mentioned above, we give the formulation of the optimization objective function of CBP-JMF as follows (**Figure 1**):

$$\min_{\{\mathbf{U}^{(p)}\}_{p=1}^{P},\{\mathbf{S}^{(p)}\}_{p=1}^{P},\mathbf{V}}\sum_{p=1}^{P}\pi^{(p)}\left\|\mathbf{X}^{(p)}-\mathbf{U}^{(p)}\mathbf{S}^{(p)}\mathbf{V}\right\|_F^2$$
$$+\beta\left\{tr\left[\mathbf{V}^L\mathbf{L}^a(\mathbf{V}^L)^T\right]-tr\left[\mathbf{V}^L\mathbf{L}^p(\mathbf{V}^L)^T\right]\right\}+\omega\|\Pi\|^2$$
$$s.t.\ \forall p, \mathbf{U}^{(p)}_{ij}\geq 0, \mathbf{V}_{ij}\geq 0, \pi^{(p)}\geq 0, \sum_{p=1}^{P}\pi^{(p)}=1 \quad (5)$$

Parameters β and ω represent the importance of the graph Laplacian regularization and weight constraint $\|\Pi\|^2$. In total, each $\mathbf{X}^{(p)}$ is factorized into individual molecular matrix $\mathbf{U}^{(p)}$ and scale matrix $\mathbf{S}^{(p)}$ and a common sample matrix $\mathbf{V}$. We allowed all matrices to share the same sample matrix $\mathbf{V}$ for finding common factors in MG data. There is only a part of samples labeled (subtype or subpopulation or subgroup is known as prior information); we incorporate this prior information with graph Laplacian. We can also learn the weights of different input data to conclude the roles that different data matrices play in CBP-JMF.

Optimization and Update Rules of CBP-JMF

To solve the problem of factorization $\mathbf{X} \approx \mathbf{USV}$, we firstly randomly initialize the solution of $\mathbf{U}$, $\mathbf{S}$, and $\mathbf{V}$ and then apply iterative multiplicative updates as the optimization

Algorithm 1 | The CBP-JMF algorithm.

```
Input:
$P$ data matrices $X^{(1)}, X^{(2)}, ..., X^{(P)}$, parameters $\beta \quad \omega$
Output:
$P$ basis matrices $\mathbf{U}^{(1)}, \mathbf{U}^{(2)}, ..., \mathbf{U}^{(P)}$, $P$ relation matrices $\mathbf{S}^{(1)}, \mathbf{S}^{(2)}, ..., \mathbf{S}^{(P)}$, factor matrices $\mathbf{V}$, weight vector $\Pi = \left(\pi^{(1)}, \pi^{(2)}, ..., \pi^{(P)}\right)$
Begin
    Initialize$\mathbf{U}^{(1)}, \mathbf{U}^{(2)}, ..., \mathbf{U}^{(P)}, \mathbf{S}^{(1)}, \mathbf{S}^{(2)}, ..., \mathbf{S}^{(P)}, \mathbf{V}$
    Initialize $\left(\pi^{(1)}, \pi^{(2)}, ..., \pi^{(P)}\right) = \left(\frac{1}{P}, \frac{1}{P}, ..., \frac{1}{P}\right)$
    loop
        for $p$=1 to $P$ do
            Fix $\mathbf{V}$, update $\mathbf{U}^{(p)}$, $\mathbf{S}^{(p)}$
        end for
        Fix $\mathbf{U}^{(1)}, \mathbf{U}^{(2)}, ..., \mathbf{U}^{(P)}$, update $\mathbf{V}^{L}$
        Fix $\mathbf{U}^{(1)}, \mathbf{U}^{(2)}, ..., \mathbf{U}^{(P)}$, update $\mathbf{V}^{UL}$
        for $p$=1 to $P$ do
            Fix $\mathbf{U}, \mathbf{S}, \mathbf{V}$, compute $c^{(p)} = \left\|\mathbf{X}^{(p)} - \mathbf{U}^{(p)}\mathbf{S}^{(p)}\mathbf{V}\right\|_F^2$
        end for
        Update $\Pi$
        break loop if convergence
End
```

approach similar to EM algorithms (Dempster et al., 1977). The optimization procedure of CBP-JMF is as follows.

To clarify the update rules of the objective function of CBP-JMF, we define $O(\mathbf{U}, \mathbf{V}, \mathbf{S}, \Pi) = \sum_{p=1}^{P} \pi^{(p)} \left\|\mathbf{X}^{(p)} - \mathbf{U}^{(p)}\mathbf{S}^{(p)}\mathbf{V}\right\|_F^2 + \beta \left\{tr\left[\mathbf{V}^L\mathbf{L}^a\left(\mathbf{V}^L\right)^T\right] - tr\left[\mathbf{V}^L\mathbf{L}^p\left(\mathbf{V}^L\right)^T\right]\right\} + \omega\|\Pi\|^2$. Firstly, we fix $\mathbf{V}$ and $\mathbf{S}$ and update $\mathbf{U}$; then, we can get the Lagrange function and let Ψ be the Lagrange multiplier for the constraints $\mathbf{U}_{ij}^{(p)} > 0$.

$$L\left(\mathbf{U}^{(P)}\right) = \mathrm{O}\left(\mathbf{U}^{(P)}\right) + tr\left(\Psi^T\mathbf{U}^{(P)}\right) \tag{6}$$

The partial derivatives of $L\left(\mathbf{U}^{(P)}\right)$ with $\mathbf{U}$ is:

$$\frac{\partial L\left(\mathbf{U}^{(P)}\right)}{\partial \mathbf{U}^{(P)}} = -2\mathbf{X}^{(p)}\mathbf{V}^T\left(\mathbf{S}^{(p)}\right)^T + 2\mathbf{U}^{(P)}\mathbf{S}^{(p)}\mathbf{V}\mathbf{V}^T\left(\mathbf{S}^{(p)}\right)^T + \Psi \tag{7}$$

Based on the KKT conditions $\Psi_{ij}\mathbf{U}_{ij} = 0$, we can get the following update rules:

$$\mathbf{U}^{(P)} \leftarrow \mathbf{U}^{(P)} \circ \frac{\mathbf{X}^{(P)}\mathbf{V}^T\left(\mathbf{S}^{(p)}\right)^T}{\mathbf{U}^{(P)}\mathbf{S}^{(p)}\mathbf{V}\mathbf{V}^T\left(\mathbf{S}^{(p)}\right)^T} \tag{8}$$

Similarly, we can get the update rules for $\mathbf{W}$, $\mathbf{V}^L$, and $\mathbf{V}^{UL}$:

$$\mathbf{S}^{(P)} \leftarrow \mathbf{S}^{(P)} \circ \frac{\left(\mathbf{U}^{(p)}\right)^T\mathbf{X}^{(p)}\mathbf{V}^T}{\left(\mathbf{U}^{(p)}\right)^T\mathbf{U}^{(P)}\mathbf{S}^{(p)}\mathbf{V}\mathbf{V}^T} \tag{9}$$

$$\mathbf{V}^L \leftarrow \mathbf{V}^L \circ \frac{\sum_{p=1}^{P} \pi^{(p)}\left(\left(\mathbf{S}^{(p)}\right)^T\left(\mathbf{U}^{(p)}\right)^T\mathbf{X}^{L(p)}\right) + \beta\mathbf{V}^L\left(\mathbf{D}^p + \mathbf{S}^a\right)}{\sum_{p=1}^{P} \pi^{(p)}\left(\mathbf{S}^{(p)}\right)^T\left(\mathbf{U}^{(p)}\right)^T\mathbf{U}^{(p)}\mathbf{S}^{(p)}\mathbf{V}^L + \beta\mathbf{V}^L\left(\mathbf{D}^a + \mathbf{S}^p\right)} \tag{10}$$

$$\mathbf{V}^{UL} \leftarrow \mathbf{V}^{UL} \circ \frac{\sum_{p=1}^{P} \pi^{(p)}\left(\left(\mathbf{S}^{(p)}\right)^T\left(\mathbf{U}^{(p)}\right)^T\mathbf{X}^{UL(p)}\right)}{\sum_{p=1}^{P} \pi^{(p)}\left(\mathbf{S}^{(p)}\right)^T\left(\mathbf{U}^{(p)}\right)^T\mathbf{U}^{(p)}\mathbf{S}^{(p)}\mathbf{V}^{UL}} \tag{11}$$

As for updating of π, when $\mathbf{U}$,$\mathbf{V}$, and $\mathbf{S}$ are fixed, minimization of $O(\pi)$ is a convex optimization, and we use convex optimization toolbox to update π.

CBPs Obtained From CBP-JMF

Values in each column of $\mathbf{U}^{(p)}$ represent the relative contribution of each molecule in each module, and values in each row of $\mathbf{V}$ represent the degree of each sample involved in each module. According to the rules of matrix multiplication, the i-th column of basis matrix $\mathbf{U}^{(p)}, p = 1, 2, ..., P$ corresponds to the i-th row of coefficient matrix $\mathbf{V}$, so there is a one-to-one correspondence between subtype and multi-omics module discovered from the columns of $\mathbf{U}^{(p)}$ matrix. Firstly, we need to know the relationship between k modules and subtypes by counting each subtype's value in each module from $\mathbf{V}^{(p)}$ matrix (see **Supplementary Note 3** in **Supplementary Material**).

To select features associated with each module, CBP-JMF calculates the z-scores of each molecule for each column vector of $\mathbf{U}^{(p)}$ as $z = (x - \bar{x})/S_x$, where $\bar{x} = \frac{1}{n}\sum_i x_i$, $S_x^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2$. Let $\mathbf{u}_j^{(p)}$ be the j-th column of $\mathbf{U}^{(p)}$ and infer a latent feature associated with j-th CBP. The continuous value $\mathbf{u}_{ij}^{(p)}$ represents the relative contribution of molecule i in the j-th CBP. $\mathbf{u}_{ij}^{(p)}$ can be regarded as x_i, and the length of $\mathbf{u}_j^{(p)}$ can be regarded as n in Equation (12). CBP-JMF calculates a z-score for each value in $\mathbf{u}_j^{(p)}$ and obtains CBP's members through a given cutoff (z-score >2 in our tests). Then, they are mapped to a built-in molecule interaction network (see "Section 'Results'") to extract their connected components as the final CBP.

RESULTS

We applied CBP-JMF to BRCA with multi-omics data. The reason we chose BRCA as example is that breast cancer is a heterogeneous complex disease, and it is the most commonly occurring cancer. BRCA is also a type of cancer that can be divided into smaller groups based on certain characteristics of the cancer cells. Distinct complex biological processes represent different subtypes. Characterizing the processes can provide us comprehensive insights into the mechanisms of how multiple

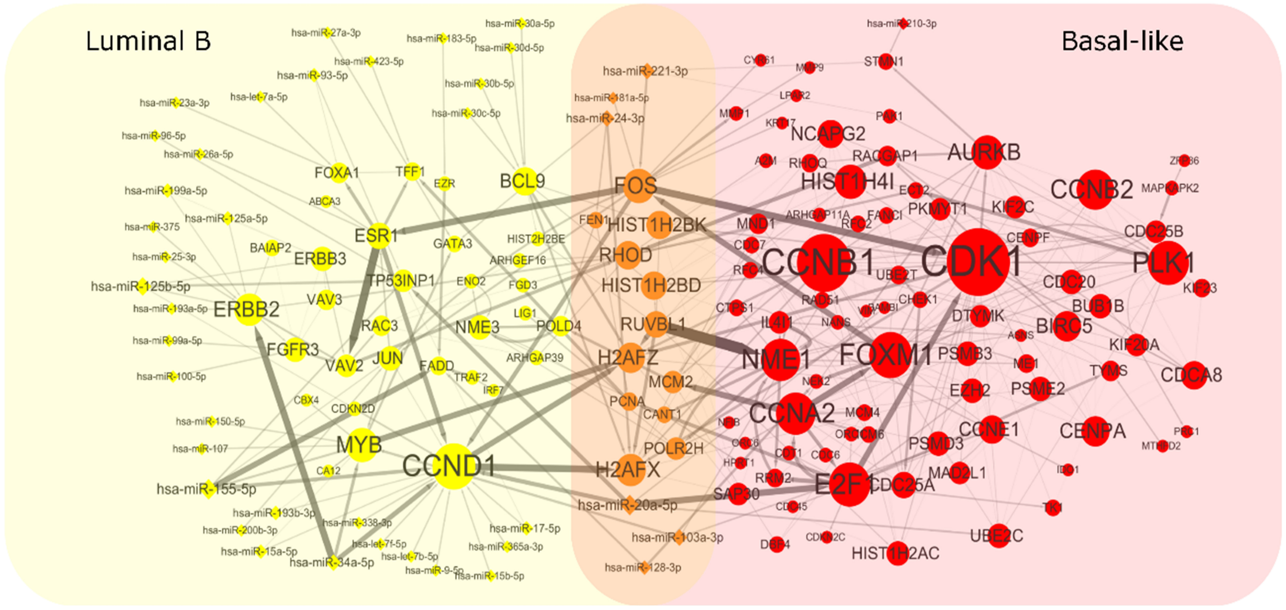

FIGURE 2 | Complex biological processes of luminal B and basal-like subtype. We mapped the genes and miRNAs obtained from luminal B's module and basal-like's module to an integrated gene regulation network. The network was obtained through integrating three databases including Reactom, Kyoto Encyclopedia of Genes and Genomes, and Nci-PID Pathway Interaction Database. The interactions between genes and miRNAs were obtained from miRTarBase. The size of the node is proportional to the size of the degree. The thickness of the edges indicates the strength of the regulatory relationship expressed by the Pearson correlation coefficient between microRNA and gene.

TABLE 1 | Enrichment analysis of the extracted module gene across six datasets.

Dataset	Online mendelian inheritance in man	CGC	Virhostome	Kinome	Drug target	BRCA pathway
Total	51	43	947	516	61	102
Overlapped nodes	2	5	13	6	3	6
P-value	0.049	0.0003	0.007	0.008	0.010	0.012

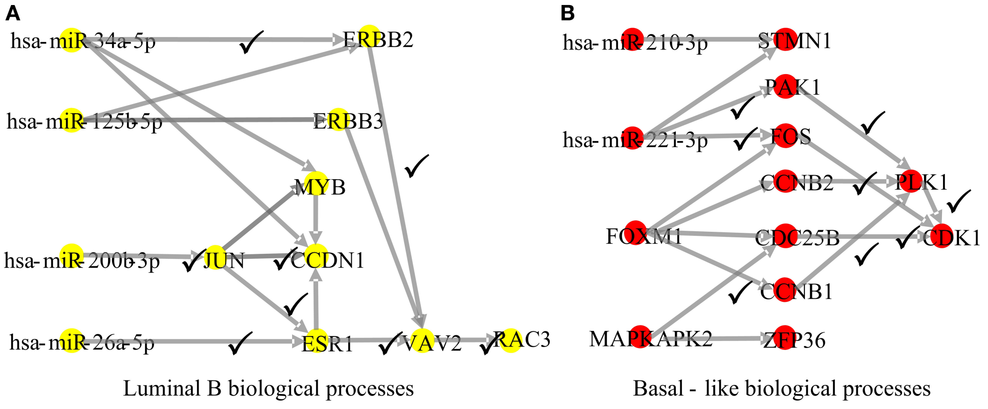

FIGURE 3 | Part of complex biological processes luminal B and basal-like. The edges with checkmarks are the interactions that have been documented. **(A)** Luminal B's biological processes: luminal subtypes are driven by the estrogen/ER pathway. Among all nodes, ERBB2, ERBB3, and ESR1 are involved in the estrogen/ER pathway. **(B)** Basal-like's biological processes: basal-like subtype is driven by the deregulation of various signaling pathways (Notch, MAPK, FoxO signaling pathway, and Wnt/beta-catenin). Among all nodes, MAPKAPK2, CDC25B, CCNB1, CCNB2, PAK1, and STMN1 are known to exist in multiple signaling pathways.

TABLE 2 | Evidences of luminal B's complex biological processes.

Interactions	Literatures	Descriptions
miR-34a->ERBB2	Wang et al., 2017	MiR-34a modulates ErbB2 in breast cancer
ERBB2->VAV2	Wang et al., 2006	ErbB2 colocalizes with Vav2 *via* activation of PI3K
VAV2->RAC3	Rosenberg et al., 2017	Vav2 promotes Rac3 activation at invadopodia
miR-200b->JUN	Jin et al., 2017	MiR-200b upregulates JUN in breast cancer
JUN->CCND1	Cicatiello et al., 2004	CCND1 promoter activation by estrogens in human breast cancer cells is mediated by the recruitment of a c-Jun/c-Fos/estrogen receptor
JUN->ESR1	Stossi et al., 2012	The activation of ESR1 gene locus in a process that was dependent upon activation and recruitment of the c-Jun transcription factor
miR-26a->ESR1	Howard and Yang, 2018	MiR-26a modulates ESR1 in breast cancer
ESR1->VAV2	Grassilli et al., 2014	ESR1 upregulates VAV2 in breast cancer cell lines

TABLE 3 | Evidences of basal-like's complex biological processes.

Interactions	Literatures	Descriptions
CCNB1(CCNB2)->PLK1->CDK1	Li et al., 2019	CCNB1 (CCNB2), PLK1, and CDK1 have interactions in chicken breast muscle
miR221->FOS	Yao et al., 2016	miR221 modulates FOS
miR221->PAK1	Ergun et al., 2015	miR221 modulates PAK1 in breast cancer cell lines
PAK1->PLK1	Maroto et al., 2008	PAK1 regulates PLK1
MAPKAPK2->CDC25B	MAPK signaling pathway	MAPKAPK2 and CDC25B are involved in MAPK signaling pathway
CDC25B->CDK1	Timofeev et al., 2010	Timely assembly of CDK1 required CDC25B

levels of molecules interact with each other and the heterogeneity of breast cancers.

Data

Firstly, we downloaded the Gene Expression (GE) data, miRNA expression (ME) data, and copy number variation (CNV) data across the same set of 738 breast cancer samples from UCSC Xena (Goldman et al., 2018). Secondly, we obtained the sample label information which is classified by PAM50 from The Cancer Genome Atlas Network (Koboldt et al., 2012). Among 738 samples, there are 522 breast cancer samples with labels, including 231 luminal A, 127 luminal B, 98 triple negative/basal-like, 58 HER2-enriched, and eight normal-like. Thirdly, we filtered out some samples, in which more than 90% of the genes have an expression value of zero. For genes and miRNAs, we filtered the genes and miRNAs with an expression value of zero in more than 20% of the samples. Fourthly, we did differential expression analysis for genes using edgeR package (Robinson et al., 2009) in R with P-value < 0.01 and $|\log(\text{fold change})| > 0.5$ to filter out genes which are not associated with breast cancer. Fifthly, we imputed missing miRNA data using knnimpute package in MATLAB. About the CNV data, the GISTIC2 (Mermel et al., 2011) thresholded the estimated values of CNV to -2, -1, 0, 1, and 2, which represent homozygous deletion, single copy deletion, diploid normal copy, low-level copy number amplification, or high-level number amplification. Finally, we obtained the GE data $\mathbf{X}^{(1)} \in \mathbf{R}^{2913\times725}$ and ME data $\mathbf{X}^{(2)} \in \mathbf{R}^{516\times725}$. Among 725 samples, 179 samples are marked with subtype labels (80 luminal A, 38 luminal B, 39 basal-like, 22 HER2-enriched) and shared between GE, ME, and CNV datasets. Furthermore, we calculated the Pearson correlation of 179 labeled samples using CNV data to construct $\mathbf{W}^{a} \in \mathbf{R}^{179\times179}, \mathbf{W}^{p} \in \mathbf{R}^{179\times179}$, and their Laplacian matrices to form the graph Laplacian regularization $tr\left[\mathbf{V}^{L}\mathbf{L}^{a}\left(\mathbf{V}^{L}\right)^{T}\right] - tr\left[\mathbf{V}^{L}\mathbf{L}^{p}\left(\mathbf{V}^{L}\right)^{T}\right]$.

Complex Biological Processes for Breast Cancer Subtypes

In our example, we set parameters $k = 4$, β=10, and ω=100, 000. Other parameters and more details can be found in **Supplementary Note 2** of **Supplementary Material**. As a result, we obtained unique matrices $\mathbf{U}^{(1)} \in \mathbf{R}^{2913\times4}$, $\mathbf{U}^{(2)} \in \mathbf{R}^{516\times4}$, $\mathbf{S}^{(1)} \in \mathbf{R}^{4\times4}$, and $\mathbf{S}^{(2)} \in \mathbf{R}^{4\times4}$ and a common matrix $\mathbf{V} \in \mathbf{R}^{4\times725}$.

To get heterogeneous CBPs (**Supplementary Table 1**), directed regulatory pathways containing miRNAs and genes, which correspond to each cancer subtype we put subtype-specific multi-omics modules obtained from matrix $\mathbf{U}^{(p)}, p = 1, 2$ onto an integrated gene regulation network from Reactome (Croft et al., 2014), Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000), and Nci-PID pathway (Schaefer et al., 2009). Then, we add directed regulatory edges from miRNA to the gene supported by miRTarBase (Chou et al., 2018). Finally, we extracted the maximum connected component of the regulation network and showed the discovered characteristic CBPs underlying luminal B and basal-like subtypes in **Figure 2**.

To explore whether the genes in the CBPs of luminal B and basal-like subtype have significant biological importance or not, we performed an enrichment analysis with all 124 genes from **Figure 2** across six datasets. The datasets are from OMIM (Hamosh et al., 2005), CGC (Futreal et al., 2004), virhostome, kinome (Manning et al., 2002), drug target (Wishart et al., 2008), KEGG pathway of BRCA (Kanehisa and Goto, 2000). Genes associated with breast cancer or breast tissue in the six datasets are selected as the set of enrichment analysis. Genes extracted through CBP-JMF have significant overlapping with known datasets (**Table 1**). Furthermore, for each subtype's CBP, functional enrichment analysis (**Supplementary Figure 4**) shows that four CBPs are mainly enriched in known biological processes and pathways associated with breast cancer, such as cell cycle and various signaling pathways (including p53 signaling pathway and estrogen pathway). However, each CBP also has its specific biological processes and path. This may explain differences between subtypes. As a demonstration, we take the CBPs of luminal B and basal-like as example. Based on the study of the subtypes of BRCA, luminal B is mainly driven by the estrogen/ER pathway (Zhang et al., 2014). In our discovered CBPs, we found several CBPs containing genes like ERBB2, ERBB3, and ESR1 that are related to the estrogen/ER pathway. Besides that, through literature review, miRNAs in luminal B's CBP can regulate the estrogen/ER pathway, such as miR-34a, miR-125b, miR-200b, and so on (**Figure 3**, **Table 2**). In addition, basal-like subtype is mainly driven by the deregulation of various signaling pathways including Notch, MAPK, and wnt/β-catenin signaling pathway (King et al., 2012). In our discovered CBPs, we found genes involved in the above-mentioned pathways, such as MAPKAPK2, CDC25B, PLK1, and so on. Besides that, we also found that miRNAs in CBPs of basal-like, such as miR-221 and miR-210, may regulate the genes above in basal-like subtype (**Figure 3**, **Table 3**). In summary, subtype-specific biological processes can be identified by CBP-JMF, and CBP-JMF can help users discover potential biological targets.

Meanwhile, to classify unlabeled samples into subtypes, CBP-JMF returned predicted labels for unlabeled samples (**Supplementary Note 4** in **Supplementary Material**). **Figure 4** shows the Kaplan–Meier (KM) survival analysis using survival package (Therneau, 2015) on unlabeled samples based on their clinical data in TCGA. We compared our results with other NMF methods (**Supplementary Note 4** of **Supplementary Material**) and found that CBP-JMF achieves more accurate subtype classification results. Unlabeled samples are classified by using GE data and ME data. **Figure 4** indicates that the survival analysis for unlabeled samples has the most significant Cox (Lin and Zelterman, 2002) p-value 0.031 and similar survival curves like the labeled samples. This proves that the CBP-JMF framework is useful for cancer subtyping, as the framework incorporates integration of multi-omics data and samples' prior information.

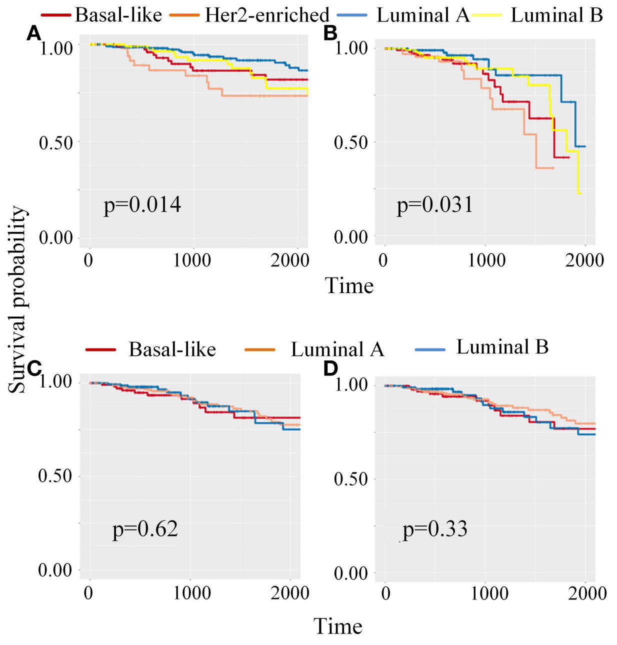

FIGURE 4 | Kaplan–Meier (K–M) survival analysis for patients which are classified using different methods. **(A)** KM survival curve for labeled samples, whose subtypes are known in advance. **(B)** KM survival analysis for unlabeled samples, which are classified using complex biological processes–joint matrix tri-factorization (CBP-JMF) on mRNA expression and miRNA expression data. **(C)** KM survival analysis for unlabeled patients, which are classified using CBP-JMF only on mRNA expression. **(D)** KM survival analysis for unlabeled patients, which are classified on mRNA expression and miRNA expression data without graph embedding regularization.

DISCUSSION

Understanding CBPs is vital to help us further understand the development of disease and intervene in the disease. NMF is an effective tool for dimension reduction and data mining in high-throughput genomic data. In this paper, we proposed CBP-JMF, an improved method of multi-view data analysis. It is designed for heterogeneous biological data based on NMF. Moreover, we created an easy-to-use package in Python. CBP-JMF analyzes multi-dimensional genomic data across the same samples integrally. Our method can discover CBPs that underlie sample groups and classify unlabeled samples through learning the relationship between labeled samples.

We tested this framework on the gene expression data and miRNA expression data of BRCA. CBP-JMF discovered subtype-specific biological processes and classified unlabeled samples into four subtypes. We did survival analysis and function analysis, and the results showed that CBP-JMF has great performance. Furthermore, CBP-JMF is a weighted joint tri-NMF framework in essence. We expect that it can be applied to vast fields including disease subtypes, cell types, and population stratification. Meanwhile, we expect that CBP-JMF can be used to identify hub genes or predict the association between genes or non-coding mRNA and diseases by integrating a variety of data. Though CBP-JMF is efficient to uncover CBPs by integrating multi-omics data, CBP-JMF must integrate different multi-omics data that have the same samples. This weakness limits the use of more types of data and integrates more information to obtain more significant results.

CONCLUSIONS

In this article, we develop CBP-JMF, a matrix tri-factorization and weighted joint integration tool, for detecting CBPs, which characterize prior disease subtypes and cell groups in Python. We improve its usability by estimating the parameters, such as determining the number of features through consensus clustering. CBP-JMF always gives reference values of all parameters. In applications, CBP-JMF characterizes the CBPs of four subtypes of BRCA based on gene and miRNA expression data from TCGA, and we find the significantly different functional pathways that characterized luminal B and basal-like subtypes.

DATA AVAILABILITY STATEMENT

The datasets presented in this study are publicly available and the addresses for finding them are listed within the article. Prediction results and a reference implementation of CBP-JMF in Python are available at: https://github.com/wangbingbo2019/CBP-JMF.

AUTHOR CONTRIBUTIONS

BW, YWu, and XM conceived and designed the experiments. YWu and MX performed the experiments. XM, RD, CZ, LY, XG, and LG analyzed the data. BW, YWu, XM, and YWa proofread the paper. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Natural Science Foundation of China under Grant Nos. 61772395 and 61873198, the Fundamental Research Funds for the Central Universities (Nos. JB190306 and ZD2009), and the Shanghai Municipal Science and Technology Major Project (No. 2018SHZDZX01). We thank LCNBI and ZJLab for the financial support.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.665416/full#supplementary-material

REFERENCES

Biton, A., Bernard-Pierrot, I., Lou, Y., Krucker, C., Chapeaublanc, E., Rubio-Pérez, C., et al. (2014). Independent component analysis uncovers the landscape of the bladder tumor transcriptome and reveals insights into luminal and basal subtypes. *Cell Rep.* 9, 1235–1245. doi: 10.1016/j.celrep.2014.10.035

Brunet, J. P., Tamayo, P., Golub, T. R., and Mesirov, J. P. (2004). Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl. Acad. Sci. U.S.A.* 101, 4164–4169. doi: 10.1073/pnas.0308531101

Cai, D., He, X., Han, J., and Huang, T. S. (2011). graph regularized nonnegative matrix factorization for data representation. *IEEE Trans. Pattern Anal. Mach. Intell.* 22, 1548–1560. doi: 10.1109/TPAMI.2010.231

Chen, J., and Zhang, S. (2018). Discovery of two-level modular organization from matched genomic data via joint matrix tri-factorization. *Nucleic Acids Res.* 46, 5967–5976. doi: 10.1093/nar/gky440

Chou, C. H., Shrestha, S., Yang, C. D., Chang, N. W., Lin, Y. L., Liao, K. W., et al. (2018). MiRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. *Nucleic Acids Res.* 46, D296–D302. doi: 10.1093/nar/gkx1067

Cicatiello, L., Addeo, R., Sasso, A., Altucci, L., Petrizzi, V. B., Borgo, R., et al. (2004). Estrogens and Progesterone promote persistent CCND1 gene activation during G1 by inducing transcriptional derepression via c-Jun/c-Fos/estrogen receptor (progesterone receptor) complex assembly to a distal regulatory element and recruitment of Cyclin D1 t. *Mol. Cell. Biol.* 24, 7260–7274. doi: 10.1128/MCB.24.16.7260-7274.2004

Croft, D., Mundo, A. F., Haw, R., Milacic, M., Weiser, J., Wu, G., et al. (2014). The Reactome pathway knowledgebase. *Nucleic Acids Res.* 42, 472–477. doi: 10.1093/nar/gkt1102

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B* 39, 1–22. doi: 10.1111/j.2517-6161.1977.tb01600.x

Ding, C., Li, T., Peng, W., and Park, H. (2006). "Orthogonal nonnegative matrix tri-factorizations for clustering," in *KDD '06: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Philadelphia, PA), 126–135. doi: 10.1145/1150402.1150420

Ergun, S., Tayeb, T. S., Arslan, A., Temiz, E., Arman, K., Safdar, M., et al. (2015). The investigation of miR-221-3p and PAK1 gene expressions in breast cancer cell lines. *Gene* 555, 377–381. doi: 10.1016/j.gene.2014.11.036

Fan, J., Salathia, N., Liu, R., Kaeser, G. E., Yung, Y. C., Herman, J. L., et al. (2016). Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nat. Methods* 13, 241–244. doi: 10.1038/nmeth.3734

Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., et al. (2004). A census of human cancer genes. *Nat. Rev. Cancer* 4, 177–183. doi: 10.1038/nrc1299

Goldman, M., Craft, B., Hastie, M., Repečka, K., McDade, F., Kamath, A., et al. (2018). The UCSC Xena platform for public and private cancer genomics data visualization and interpretation. *bioRxiv*, 1–16. doi: 10.1101/326470

Grassilli, S., Brugnoli, F., Lattanzio, R., Rossi, C., Perracchio, L., Mottolese, M., et al. (2014). High nuclear level of Vav1 is a positive prognostic factor in early invasive breast tumors: a role in modulating genes related to the efficiency of metastatic process. *Oncotarget* 5, 4320–4336. doi: 10.18632/oncotarget.2011

Guan, Z., Zhang, L., Peng, J., and Fan, J. (2015). Multi-view concept learning for data representation. *IEEE Trans. Knowl. Data Eng.* 27, 3016–3028. doi: 10.1109/TKDE.2015.2448542

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., and McKusick, V. A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* 33, 514–517. doi: 10.1093/nar/gki033

Hasin, Y., Seldin, M., and Lusis, A. (2017). Multi-omics approaches to disease. *Genome Biol.* 18, 1–15. doi: 10.1186/s13059-017-1215-1

Howard, E. W., and Yang, X. (2018). MicroRNA regulation in estrogen receptor-positive breast cancer and endocrine therapy. *Biol. Proced. Online* 20, 1–19. doi: 10.1186/s12575-018-0082-9

Jin, T., Kim, H. S., Choi, S. K., Hwang, E. H., Woo, J., Ryu, H. S., et al. (2017). microRNA-200c/141 upregulates SerpinB2 to promote breast cancer cell metastasis and reduce patient survival. *Oncotarget* 8, 32769–32782. doi: 10.18632/oncotarget.15680

Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto Encyclopedia of genes and genomes. *Oxford Univ. Press Nucleic Acids Res.* 28, 27–30. doi: 10.1093/nar/28.1.27

King, T. D., Suto, M. J., and Li, Y. (2012). The wnt/β-catenin signaling pathway: a potential therapeutic target in the treatment of triple negative breast cancer. *J. Cell. Biochem.* 113, 13–18. doi: 10.1002/jcb.23350

Koboldt, D. C., Fulton, R. S., McLellan, M. D., Schmidt, H., Kalicki-Veizer, J., McMichael, J. F., et al. (2012). Comprehensive molecular portraits of human breast tumours. *Nature* 490, 61–70. doi: 10.1038/nature11412

Lee, D. D., and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature* 401, 788–791. doi: 10.1038/44565

Li, Y., Chen, Y., Jin, W., Fu, S., Li, D., Zhang, Y., et al. (2019). Analyses of microRNA and mRNA expression profiles reveal the crucial interaction networks and pathways for regulation of chicken breast muscle development. *Front. Genet.* 10, 1–15. doi: 10.3389/fgene.2019.00197

Lin, H., and Zelterman, D. (2002). Modeling survival data: extending the cox model. *Technometrics* 44, 85–86. doi: 10.1198/tech.2002.s656

Manning, G., Whyte, D. B., Martinez, R., Hunter, T., and Sudarsanam, S. (2002). The protein kinase complement of the human genome. Science 298, 1912–1934. doi: 10.1126/science.1075762

Maroto, B., Ye, M. B., Von Lohneysen, K., Schnelzer, A., and Knaus, U. G. (2008). P21-activated kinase is required for mitotic progression and regulates Plk1. *Oncogene* 27, 4900–4908. doi: 10.1038/onc.2008.131

Mermel, C. H., Schumacher, S. E., Hill, B., Meyerson, M. L., Beroukhim, R., and Getz, G. (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* 12, 1–14. doi: 10.1186/gb-2011-12-4-r41

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2009). edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi: 10.1093/bioinformatics/btp616

Rosenberg, B. J., Gil-Henn, H., Mader, C. C., Halo, T., Yin, T., Condeelis, J., et al. (2017). Phosphorylated cortactin recruits Vav2 guanine nucleotide exchange factor to activate Rac3 and promote invadopodial function in invasive breast cancer cells. *Mol. Biol. Cell* 28, 1347–1360. doi: 10.1091/mbc.e16-12-0885

Schaefer, C. F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., et al. (2009). PID: the pathway interaction database. *Nucleic Acids Res.* 37, 674–679. doi: 10.1093/nar/gkn653

Stein-O'Brien, G. L., Arora, R., Culhane, A. C., Favorov, A. V., Garmire, L. X., Greene, C. S., et al. (2018). Enter the matrix: factorization uncovers knowledge from omics. *Trends Genet.* 34, 790–805. doi: 10.1016/j.tig.2018.07.003

Stossi, F., Madak-Erdogan, Z., and Katzenellenbogen, B. S. (2012). Macrophage-elicited loss of estrogen receptor-α in breast cancer cells via involvement of MAPK and c-Jun at the ESR1 genomic locus. *Oncogene* 31, 1825–1834. doi: 10.1038/onc.2011.370

Suravajhala, P., Kogelman, L. J. A., and Kadarmideen, H. N. (2016). Multi-omic data integration and analysis using systems genomics approaches: methods and applications in animal production, health and welfare. *Genet. Sel. Evol.* 48, 1–14. doi: 10.1186/s12711-016-0217-x

Therneau, T. M. (2015). *A Package for Survival Analysis in S. Version 2.38*. Available online at: https://cran.r-project.org/package=survival.

Timofeev, O., Cizmecioglu, O., Settele, F., Kempf, T., and Hoffmann, I. (2010). Cdc25 phosphatases are required for timely assembly of CDK1-cyclin B at the G2/M transition. *J. Biol. Chem.* 285, 16978–16990. doi: 10.1074/jbc.M109.096552

Wang, S. E., Shin, I., Wu, F. Y., Friedman, D. B., and Arteaga, C. L. (2006). HER2/Neu (ErbB2) signaling to Rac1-Pak1 is temporally and spatially modulated by transforming growth factor β. *Cancer Res.* 66, 9591–9600. doi: 10.1158/0008-5472.CAN-06-2071

Wang, Y., Zhang, X., Chao, Z., Kung, H. F., Lin, M. C., Dress, A., et al. (2017). MiR-34a modulates ErbB2 in breast cancer. *Cell Biol. Int.* 41, 93–101. doi: 10.1002/cbin.10700

Wishart, D. S., Knox, C., Guo, A. C., Cheng, D., Shrivastava, S., Tzur, D., et al. (2008). DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* 36, 901–906. doi: 10.1093/nar/gkm958

Xi, J., Li, A., and Wang, M. (2018). A novel unsupervised learning model for detecting driver genes from pan-cancer data through matrix tri-factorization framework with pairwise similarities constraints. *Neurocomputing* 296, 64–73. doi: 10.1016/j.neucom.2018.03.026

Xi, J., Yuan, X., Wang, M., Li, A., Li, X., and Huang, Q. (2020). Inferring subgroup-specific driver genes from heterogeneous cancer samples via subspace learning with subgroup indication. *Bioinformatics* 36, 1855–1863. doi: 10.1093/bioinformatics/btz793

Yang, Z., and Michailidis, G. (2016). A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. *Bioinformatics* 32, 1–8. doi: 10.1093/bioinformatics/btv544

Yao, M., Gao, W., Yang, J., Liang, X., Luo, J., and Huang, T. (2016). The regulation roles of miR-125b, miR-221 and miR-27b in porcine Salmonella infection signalling pathway. *Biosci. Rep.* 36, 1–11. doi: 10.1042/B.S.R.20160243

Zhang, M. H., Man, H. T., Zhao, X. D, Dong, N., and Ma, S. L. (2014). Estrogen receptor-positive breast cancer molecular signatures and therapeutic potentials (review). *Biomed. Rep.* 2, 41–52. doi: 10.3892/br.2013.187

Zhang, S., Liu, C. C., Li, W., Shen, H., Laird, P. W., and Zhou, X. J. (2012). Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Res.* 40, 9379–9391. doi: 10.1093/nar/gks725

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.