Hindawi
Complexity
Volume 2020, Article ID 3917812, 10 pages
https://doi.org/10.1155/2020/3917812

Research Article

Joint Nonnegative Matrix Factorization Based on Sparse and Graph Laplacian Regularization for Clustering and Co-Differential Expression Genes Analysis

Ling-Yun Dai, Rong Zhu, and Juan Wang

School of Computer Science, Qufu Normal University, Rizhao 276826, China

Correspondence should be addressed to Ling-Yun Dai; dailingyun_1@163.com

Received 12 June 2020; Revised 8 October 2020; Accepted 12 October 2020; Published 16 November 2020

Academic Editor: Jia Wu

The explosion of multiomics data poses new challenges to existing data mining methods. Joint analysis of multiomics data can make the best of the complementary information that is provided by different types of data. Therefore, they can more accurately explore the biological mechanism of diseases. In this article, two forms of joint nonnegative matrix factorization based on the sparse and graph Laplacian regularization (SG-jNMF) method are proposed. In the method, the graph regularization constraint can preserve the local geometric structure of data. $L_{2,1}$-norm regularization can enhance the sparsity among the rows and remove redundant features in the data. First, SG-jNMF1 projects multiomics data into a common subspace and applies the multiomics fusion characteristic matrix to mine the important information closely related to diseases. Second, multiomics data of the same disease are mapped into the common sample space by SG-jNMF2, and the cluster structures are detected clearly. Experimental results show that SG-jNMF can achieve significant improvement in sample clustering compared with existing joint analysis frameworks. SG-jNMF also effectively integrates multiomics data to identify co-differentially expressed genes (Co-DEGs). SG-jNMF provides an efficient integrative analysis method for mining the biological information hidden in heterogeneous multiomics data.

1. Introduction

With the development of state-of-the-art sequencing technology, a large quantity of effective experimental data has been collected. These data may imply some unknown molecular mechanisms. Bioinformatics is faced with the task of analyzing massive omics data. The Cancer Gene Atlas (TCGA, https://tcgadata.nci.nih.gov/tcga/) includes gene expression profile data (GE), DNA methylation data (DM), copy number variation data (CNV), protein expression data, and drug sensitivity data. These data are from approximately 15,000 clinical samples of more than 30 kinds of cancers [1]. These massive data enable researchers to study the mechanisms of cancer production, diagnosis, and treatment at different biological levels.

The joint analysis of multiomics data can make up for lost or unreliable information in single omics data. In recent years, scientists have performed considerable research on the cancer mechanisms based on the joint analysis of cancer multiomics data. For example, Christina et al. integrated the gene expression data and copy number variations of breast cancer, identified possible pathogenic genes, and discovered new subtypes of breast cancer [2]. Wang and Wang used similarity network fusion to jointly analyze mRNA, DM, and microRNA (miRNA) data and identify cancer subtypes further [3]. In the existing joint analysis methods, those based on matrix decomposition are remarkable. Liu et al. integrated mRNA, somatic cell mutation, DNA methylation, and copy number variation data. They established a block constraint-based RPCA model to identify differentially expressed genes (DEGs) [4]. Integration and analysis of these heterogeneous multiomics data provide an in-depth understanding of the pathogenesis of cancer and promote the development of precision medicine. Recently, unsupervised integrative methods based on matrix decomposition have attracted considerable

attention among the existing methods for integrating and analyzing multiomics data. Zhang et al. constructed a joint matrix factorization framework (jNMF) to discover multidimensional modules of genomic data [5]. Yang and Michailidis introduced a new method named integrative NMF (iNMF) for heterogeneous multiomics data [6]. Strazar et al. incorporated orthogonality regularization into iNMF (iONMF) to integrate and analyze multiple data sources [7]. Joint nonnegative matrix decomposition meta-analysis (jNMFMA) [8], multiomics factor analysis (MOFA) [9], and Bayesian joint analysis [10] have been successfully applied to the integration and analysis of cancer omics data. To avoid the influence of redundant information, many sparse modeling methods have been proposed. Typical applications are as follows: The weighted sparse representation classifier (WSRC) model combined with global coding (GE) [11] was used to predict interactions between proteins based on protein sequence information. The network regularization sparse logic regression model (NSLR) [12] was used to predict survival risk and discover biomarkers. Sparse coregularization matrix decomposition was used to find mutant driver genes and so on [13].

In recent years, graph/network-based analysis as a powerful data representation tool has been applied to the modeling and analysis of complex systems [14–17]. In general, entities can be regarded as nodes, and the interaction between entities can be regarded as edges in the graph. Graph-based approaches can explore the local subspace structure and obtain the low-dimensional representation of high-dimensional data. Zhang and Ma proposed a subspace clustering algorithm based on a graph to detect the common modules highly correlated with cancer by jointly analyzing the gene expression and protein interaction networks [18]. Mixed-norm Laplacian regularized low-rank representation (MLLRR) was used to cluster samples [19]. Cui proposed an improved graph-based method to predict drug-target interactions [20]. Liu et al. introduced the contributions of deep neural networks, deep graph embedding, and graph neural networks along with the opportunities and challenges they faced [21]. Wu et al. proposed a multigraph learning algorithm called gMGFL that search and choose a group of decision subgraphs as features to move bags and bag labels to the instance [22].

Recently, sparse regularization has played a very important role in data analysis. The L_0-norm, L_1-norm, $L_{2,1}$-norm, etc. are all typical sparse regularization methods. Among these many sparse constraints, $L_{2,1}$-norm regularization stands out in terms of computational time and performance. The $L_{2,1}$-norm can obtain a sparse projection matrix in rows to learn discriminative features in the subspace. Zhang used the $L_{2,1}$-norm constraint on the coefficients to ensure that they are sparse in rows [23]. The $L_{2,1}$-norm was applied to the predictor to ensure that it is robust to noise and outliers [24].

Considering the role of graph regularizations and $L_{2,1}$-norm constraints in matrix factorization, we propose joint nonnegative matrix factorization based on sparse and graph Laplacian regularization (SG-jNMF). SG-jNMF can make the best of the potential associations and complementary information among multiomics data. The main highlights of this approach are as follows.

(1) Graph regularization is incorporated into the joint nonnegative matrix factorization model, and undirected graphs are constructed for input data in this method. Local graph regularization can preserve the local geometrical structure of the data space. Therefore, SG-jNMF can use the low-dimensional characteristics of the observed data to find intrinsic laws and improve the performance of the integrated analysis method.

(2) $L_{2,1}$-norm regularization can deal with each row of the matrix as a whole and can enhance the sparsity among the rows. Therefore, involving the $L_{2,1}$-norm can remove redundant features and noise in the data and further explore the clear cluster structure.

(3) Two forms of SG-jNMF are proposed. SG-jNMF1 projects multiomics data into a fusion feature space. The fusion matrix contains complementary and differential information provided by multiomics data, so that more accurate results can be obtained when identifying Co-DEGs. SG-jNMF2 projects multiomics data into a common sample space, which results in more accurate clustering results.

The rest of this paper is arranged as follows: In Section 2, we start with a brief review of jNMF. Next, we introduce the SG-jNMF method, optimization process, and computational complexity analysis. Section 3 gives out the experimental results of clustering and feature selection. Finally, we summarize the whole paper and give some suggestions for future work in Section 4.

2. Materials and Methods

2.1. Joint Nonnegative Matrix Factorization. The jNMF method was first proposed by Zhang et al. [5]. It can project multiple input data matrices into a common subspace, to integrate the information of each input data for analysis. Each type of genomic data as original data can be denoted as $X_I \in R^{M\times N}$ $(I = 1, 2, 3, \ldots)$. $W \in R^{M\times K}$ is the common basis matrix, and $H_I \in R^{K\times N}$ is the corresponding coefficient matrix. The objective function of jNMF can be written as

$$\begin{aligned} \min \quad & \sum_{I=1}^{P} \left\| X_I - WH_I \right\|_F^2, \\ \text{s.t.} \quad & W \geq 0, H_I \geq 0. \end{aligned} \tag{1}$$

Obviously, jNMF is the same as NMF when $P = 1$. Therefore, jNMF is the generalization model of NMF for multiple input datasets. Similar to NMF, multiplicative update rules are used to minimize the objective function. W and H_I are iteratively updated according to the following rules.

$$W_{ia} = W_{ia}\frac{\left(\sum_{I=1}^{P}\left(X_I H_I^T\right)\right)_{ia}}{\left(\sum_{I=1}^{P}\left(W\left(H_I H_I^T\right)\right)\right)_{ia}}, \tag{2}$$

$$H_{Iaj} = H_{Iaj}\frac{\left(W^T X_I\right)_{aj}}{\left(W^T W H_I\right)_{aj}}. \tag{3}$$

The jNMF method can be used to integrate and analyze multiomics data. It decomposes multiomics data matrices into multiple independent coefficient matrices and a common fusion matrix at the same time and projects high-dimensional omics data into low-dimensional spaces. Therefore, the abundant differential and complementary information of cancer multiomics data can be efficiently used, and multiomics datasets are analyzed simultaneously to obtain hidden information with biological significance.

2.2. Joint Nonnegative Matrix Factorization Based on Sparse and Graph Laplacian Regularization. Manifold learning has become a popular research topic in the domain of information science since it was first proposed in science in 2000 [25, 26]. Assuming that the data are uniformly sampled in a high-dimensional space, manifold learning can find the low-dimensional structure in the high-dimensional space and obtain the corresponding embedding mapping. Manifold learning looks for the essence of things from observed phenomena and finds the internal laws of data. The manifold assumption states that data points that are geometrically adjacent usually have similar characteristics. Therefore, an undirected weighted network/graph $G = (V; E; U)$ is constructed. $V = 1, 2, \ldots, P$ is the vertex set, E is the edge set, and U is the weight set. Edge weight $U_{j,k}(1 \le j \ne k \le q)$ is associated with edge(j, k) in E. The graph regularization with G is as follows:

$$\sum_k R_k = \sum_k \frac{1}{2}\sum_{i,j}^{N}\left(\left(f_k(x_i) - f_k(x_j)\right)^2 U_{i,j}\right) = \sum_k \frac{1}{2}\sum_{i,j}^{N}\left(\left(V_{i,j} - V_{k,j}\right)^2 U_{i,j}\right) = \sum_k V_k^T D V_k - \sum_k V_k^T U V_k = \sum_k V_k^T L V_k = T_r\left(V^T L V\right), \tag{4}$$

where $T_r(\cdot)$ is the trace of the matrix, L is the graph Laplacian matrix, and $L = D - U$. D is a diagonal matrix and $D_{i,j} = \sum_j U_{i,j}$. Intuitively, the smaller the R_k value is, the closer the two data points are. By minimizing R_k, we can obtain a sufficiently smooth mapping function on the data manifold.

To decrease the influence of noise and outliers on real data, sparse regularization is usually used to penalize the coefficient matrix. The L_0-norm, L_1-norm, and $L_{2,1}$-norm are all typical sparse regularization methods. The solution of L_0-norm is a NP-hard problem. L_1-norm is widely used because it has better optimization solution characteristics than L_0-norm. L_1-norm will tend to produce a small number of features, while the other features are all 0. Therefore, it can be used for feature selection. However, L_1-norm regularization is usually time-consuming. $L_{2,1}$-norm regularization on the coefficient matrix can generate a row sparse result, and the calculation of the $L_{2,1}$-norm is simple and convenient [23]. In this article, the $L_{2,1}$ penalty is incorporated in SG-jNMF [27]. The $L_{2,1}$-norm of a matrix Z is defined as

$$\|Z\|_{2,1} = \sum_{i=1}^{m}\sqrt{\sum_{j=1}^{n} z_{ij}^2} = \sum_{i=1}^{m}\left\|z^i\right\|_2. \tag{5}$$

2.2.1. SG-jNMF1. There are two forms of SG-jNMF methods in this article. As shown in Figure 1, the SG-jNMF1 method projects multiomics data into a common feature space. Graph regularization and a sparse penalty are applied to the fusion feature matrix. The feature matrix is constrained by graph regularization, and as much intrinsic geometric information of the original multiomics data are preserved as possible. The $L_{2,1}$-norm is used to constrain the feature matrix to reduce the influence of outliers and noise, and the objective function of integrating nonnegative matrix decomposition is constructed. The optimization problem can be expressed as

$$\begin{aligned} &\min \sum_{I=1}^{P}\left\|X_I - W H_I^T\right\|_F^2 + \sum_{I=1}^{P}\lambda_I T_r\left(W L_{I1} W^T\right) + \beta\|W\|_{2,1}, \\ &\text{s.t.} \quad W \ge 0, H_I \ge 0, \end{aligned} \tag{6}$$

where L_{I1} is the Laplacian matrix. $L_{I1} = D_{I1} - U_{I1}$, where U_{I1} is a symmetric matrix, which is the weight matrix constructed in graph regularization. D_{I1} is a diagonal matrix, and its diagonal elements are equal to the sum of the corresponding row elements or the sum of the column elements of the matrix; i.e., $D_{I1ii} = \sum_{j=1}^{n}(U_{I1ij})$.

With randomly positive initializing matrices $W$and H_I, the following update rules are executed until the algorithm converges:

$$\begin{aligned} W_{ia} &= W_{ia}\frac{\left(\sum_{I=1}^{P}\left(X_I H_I + \lambda_I U_{I1} W\right)\right)_{ia}}{\left(\sum_{I=1}^{P}\left(H_I H_I^T W + \lambda_I D_{I1} W\right) + \beta Q W\right)_{ia}}, \\ H_{Iaj} &= H_{Iaj}\frac{\left(X_I^T W^T\right)_{aj}}{\left(H_I W^T W\right)_{aj}}, \end{aligned} \tag{7}$$

where Q is a diagonal matrix, the diagonal element is $Q_{jj} = 1/\sqrt{\sum_{i=1}^{m}(W_{ij}) + \varepsilon}$, and ε is an infinitesimal positive number.

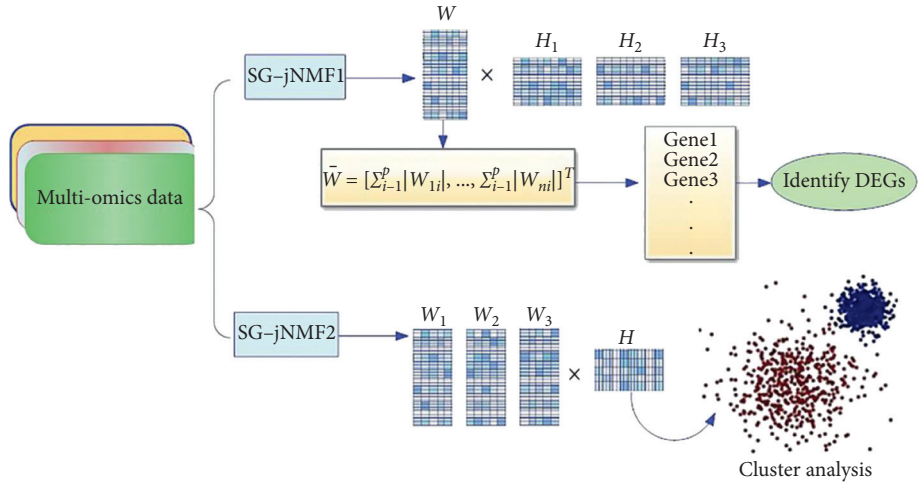

Figure 1: Framework of SG-jNMF.

2.2.2. SG-jNMF2. As seen from Figure 1, the SG-jNMF2 method projects multiomics data into a common sample space. Constraints are enforced on the common sample matrix. This method can be used to cluster multiomics data. The model can be shown by the following expression:

$$\min \sum_{I=1}^{P} \left\| X_I - W_I H^T \right\|_F^2 + \sum_{I=1}^{P} \lambda_I T_r\left(H^T L_{I2} H\right) + \beta \|H\|_{2,1},$$
$$\text{s.t.} \quad W_I \geq 0, H \geq 0. \tag{8}$$

Similarly, the algorithm iterates until it converges according to the following rules:

$$W_{Iia} = W_{Iia} \frac{(X_I H)_{ia}}{\left(W_I H^T H\right)_{ia}},$$
$$H_{aj} = H_{aj} \frac{\left(\sum_{I=1}^{P} \left(X_I^T W_I + \lambda_I U_{I2} H\right)\right)_{aj}}{\left(\sum_{I=1}^{P} \left(H W_I^T W + \lambda_I D_{I2} H\right) + \beta B H\right)_{aj}}, \tag{9}$$

where L_{I2} is the Laplacian matrix. $L_{I2} = D_{I2} - U_{I2}$, where $U_{I2}$is a symmetric matrix, which is the weight matrix constructed in graph regularization. D_{I2} is a diagonal matrix, and its diagonal elements are equal to the sum of the corresponding row elements or the sum of the column elements of the matrix; i.e., $D_{I2ii} = \sum_{j=1}^{n} (U_{I2ij})$. B is a diagonal matrix, and the diagonal element is $B_{jj} = 1/\sqrt{\sum_{i=1}^{m} (W_{ij}) + \varepsilon}$. Obviously, the objective functions of the two kinds of SG-jNMF method are both nonconvex. We can obtain the optimal solutions by minimizing the objective functions. The optimization process is shown as follows.

2.3. Optimization of SG-jNMF. Since the optimization processes of the two forms of SG-jNMF method are very similar, we only provide that of the first method. We use the multivariable alternating update rules to solve the optimization problem. Specifically, the following update steps are repeated until the algorithm converges.

2.3.1. Optimization of W. When H_I is fixed, the optimization of W is performed by minimizing the following objective function:

$$\begin{aligned} O &= \sum_{I=1}^{P} \left\| X_I - W H_I^T \right\|_F^2 + \sum_{I=1}^{P} \lambda_I T_r\left(W L_{I1} W^T\right) + \beta \|W\|_{2,1} \\ &= \sum_{I=1}^{P} \left(T_r\left(X_I^T X - 2 X_I^T W H_I^T + H_I W^T W H_I^T\right)\right) \\ &\quad + \lambda_I T_r\left(W L_{I1} W^T\right) + \beta T_r\left(W^T Q W\right). \end{aligned} \tag{10}$$

The corresponding Lagrangian function is as follows:

$$\begin{aligned} L = &\sum_{I=1}^{P} \left(T_r\left(X_I^T X - 2 X_I^T W H_I^T + H_I W^T W H_I^T\right)\right) \\ &+ \lambda_I T_r\left(W L_{I1} W^T\right) + \beta T_r\left(W^T Q W\right) \\ &+ T_r(\Phi W) + T_r\left(\Psi H_I^T\right), \end{aligned} \tag{11}$$

where $\Phi = [\phi_{il}]$ and $\Psi = [\psi_{Ia}]$ are the Lagrangian multipliers of $W$and H_I, respectively. Next, we take the first partial derivative of this Lagrangian function with respect to W:

$$\frac{\partial l}{\partial W} = \sum_{I=1}^{P} \left[-2 X H_I + 2 W H_I^T H_I + 2 \lambda_I L_{I1} W\right] + 2 \beta Q W + \Phi. \tag{12}$$

According to the KKT conditions [28], the following updating rule can be obtained:

$$W_{ia} = W_{ia} \frac{\left(\sum_{I=1}^{P} \left(X_I H_I + \lambda_I U_{I1} W\right)\right)_{ia}}{\left(\sum_{I=1}^{P} \left(H_I H_I^T W + \lambda_I D_{I1} W\right) + \beta Q W\right)_{ia}}. \tag{13}$$

2.3.2. Optimization of H_I. When W is fixed, the optimization of H_I is performed by minimizing the following objective function.

$$O = \sum_{I=1}^{P} \left\| X_I - WH_I^T \right\|_F^2, \quad \text{s.t.} \quad W \geq 0, H_I \geq 0. \tag{14}$$

The corresponding Lagrangian function is as follows:

$$l = \sum_{I=1}^{P} \left(T_r \left(X_I^T X - 2X_I^T W H_I^T + H_I W^T W H_I^T \right) \right) + T_r(\Psi) H_I^T, \tag{15}$$

and H_I runs to convergence according to the following formula:

$$H_{Iaj} = H_{Iaj} \frac{\left(X_I^T W^T \right)_{aj}}{\left(H_I W^T W \right)_{aj}}. \tag{16}$$

2.4. Convergence and Running Time. In this paper, we also demonstrate the convergence of the method through experiments. Taking the pancreatic adenocarcinoma (PAAD) dataset as an example, the convergence of the five methods is shown in Figure 2. The error function used in this article is defined as follows:

$$\text{Loss} = \sum_{I=1}^{P} \frac{\left\| X_I - WH_I^T \right\|_F^2}{\left\| X_I \right\|_F^2}. \tag{17}$$

Compared with the other four methods, SG-jNMF can converge to the smallest error value with the fastest speed.

Besides, we also tested the running time of the above methods on the PAAD dataset. The means of these five methods running 10 times on a PC are shown in Table 1. As seen in Table 1, iGMFNA has the shortest running time, followed by SG-jNMF. This is due to the introduction of sparse constraints in SG-jNMF. The running time of iNMF, iGMFNA, jNMF, and SG-jNMF methods is satisfactory.

2.5. Computational Complexity Analysis. In this part, we discuss the extra computational complexity of SG-jNMF compared to jNMF. We use big O symbol to represent the computational complexity of the algorithm. On the basis of the updating rules (3) and (4), we can easily count the arithmetic operations of each iteration in jNMF. Obviously, the cost for each iteration in jNMF is $O(\text{MNk})$. It should be noted that U_I is a sparse matrix for SG-jNMF. In addition to the multiplicative updates, constructing a K-nearest neighbor graph requires $O(N^2M)$ operations [28]. Assume that the update stops after t iterations, and the overall cost for jNMF is $O(\text{tMNk})$. The overall cost for SG-jNMF is $O(N^2M) + O(\text{tPMNk})$.

3. Results and Discussion

3.1. Data Processing. TCGA project includes a lot of gene expression profile data, DNA methylation data, copy number variation data, protein expression data, drug sensitivity data, and so on. In-depth study of these data can help us to master the mechanism of cancer occurrence and development and provide technical support for prevention, diagnosis, and treatment of cancer. In this article, four cancer datasets which are all downloaded from TCGA (https://tcgadata.nci.nih.gov/tcga/), namely, PAAD, esophageal carcinoma (ESCA), cholangiocarcinoma (CHOL), and colon adenocarcinoma (COAD), are used in these experiments. Details are listed in Table 2. To avoid the matrix dimension problem in algorithm execution, the number of genes in the four datasets is aligned to 19,876. First, RPCA is used to reduce the effects of noise and redundant information [29]. Second, the same number of samples and characteristics is retained for multiomics data of the same kind of cancer. Then, the matrices are normalized according to the standard deviation of the data such that each element of the matrix is evaluated between 0 and 1.

3.2. Clustering. When SG-jNMF2 method projects multiomics data into a common sample space, it contains all the sample information provided by the input multiomics data. To assess the clustering performance of this method, SG-jNMF2 is used to cluster the tumor samples on CHOL, PAAD, COAD, and ESCA datasets. There are four methods (iNMF, iONMF, iGMFNA, and jNMF) that perform the same experiments on the same datasets.

3.2.1. Selection of Parameters. For SG-jNMF2, clustering performance is affected by the regularization parameters. In this experiment, we empirically set the same value for λ_I with different omics data from the same cancer [30]. Therefore, there are three parameters, λ, β, and K, that need to be adjusted. λ is the graph regularization parameter, β controls the sparsity of factorization, and K is the number of nodes in the undirected graph constructed in the manifold. From Figure 3, when K is set to 3, the accuracy on the four datasets reaches a maximum. As seen from Figure 4, λ should be set to 1,000 on PAAD. When λ is equal to 0.1, the accuracy on COAD can achieve the maximum. When λ is equal to 10^{-3}, 10^{-5}, and 1, the accuracy on CHOL can achieve the maximum. When λ is equal to 10^4, the accuracy on ESCA can achieve the maximum. From Figure 5, when β is set from 10^{-5} to 10^1 for PAAD, the accuracy reaches the maximum. For ESCA and COAD, β should be set from 10^4 to 10^5. For CHOL, the value of β does not matter much.

3.2.2. Evaluation Indicators. Several indicators are used to evaluate the clustering performance of SG-jNMF2: accuracy, recall, precision, and F1-score. Accuracy is defined as

$$\text{AC} = \frac{\sum_{j=1}^{N} \delta\left(s_j, \text{map}\left(r_j \right) \right)}{N}, \tag{18}$$

where N is the total number of samples in the dataset and $\delta(x, y)$ is a singular function. When x is equal to y, the value of the function is equal to 1; otherwise, it is equal to 0. $\text{map}(r_j)$ maps the clustering label r_j to the real label s_j. The other three indicators used to evaluate clustering performance are defined as follows:

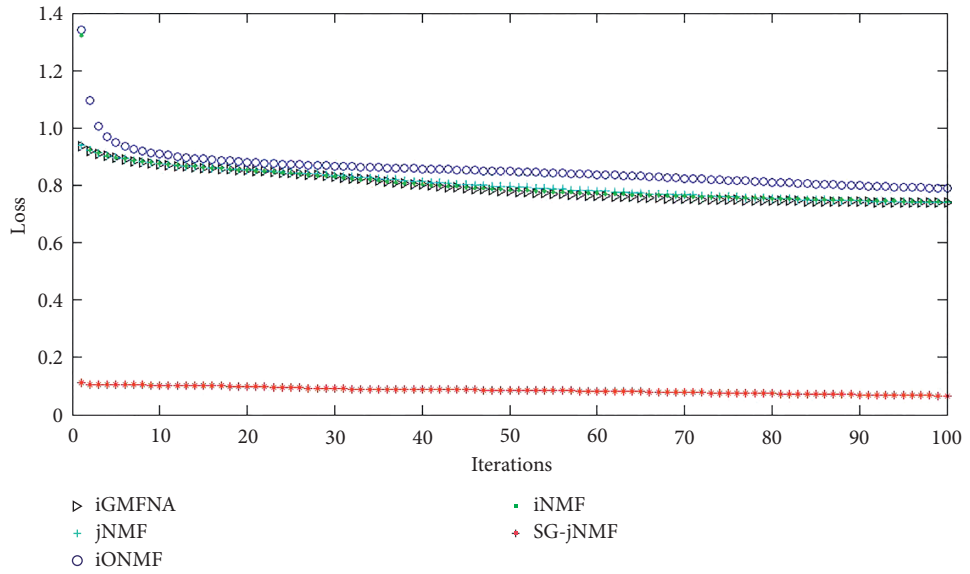

Figure 2: Comparison of convergence of five methods on PAAD dataset.

Table 1: Running time on PAAD.

Methods	Running times
iNMF	2.4261
iGMFNA	0.1312
jNMF	1.6311
iONMF	46.8843
SG-jNMF	0.5153

Table 2: Overview of multiomics datasets.

Multiomics datasets	Total number of samples	Cancer samples	Number of genes
PAAD (GE, ME, CNV)	180	176	19877
CHOL (GE, ME, CNV)	45	36	19876
ESCA (GE, ME, CNV)	192	183	19877
COAD (GE, ME, CNV)	281	262	22723

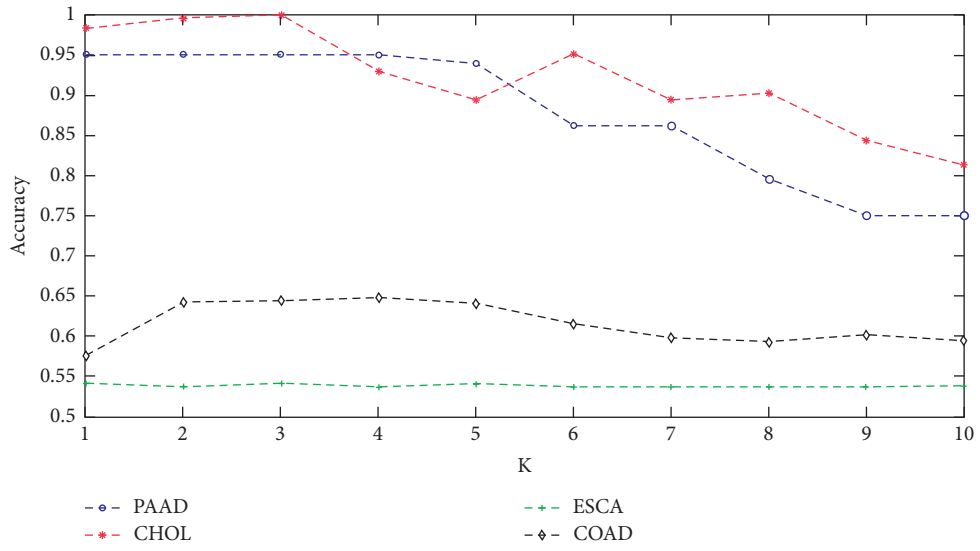

Figure 3: Accuracy of SG-jNMF varies with K.

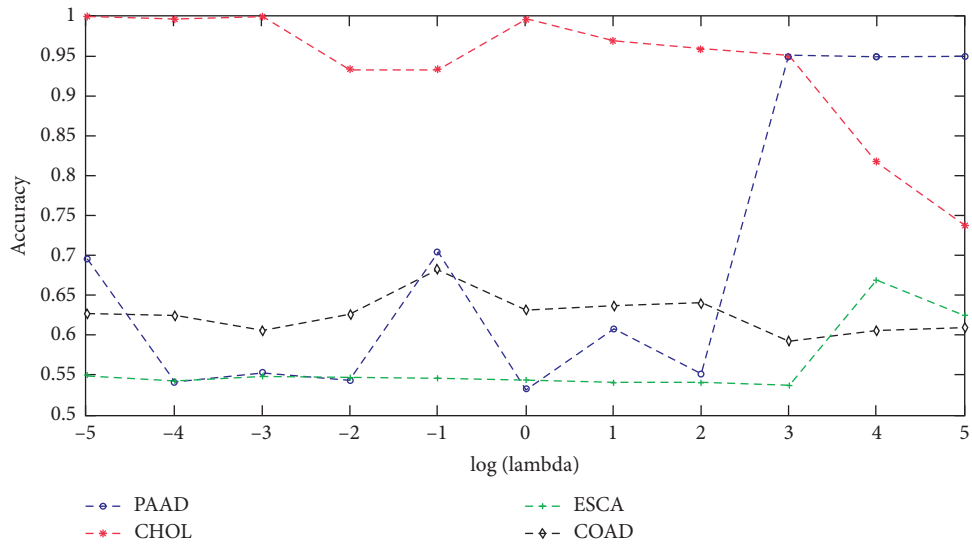

Figure 4: Accuracy of SG-jNMF varies with λ.

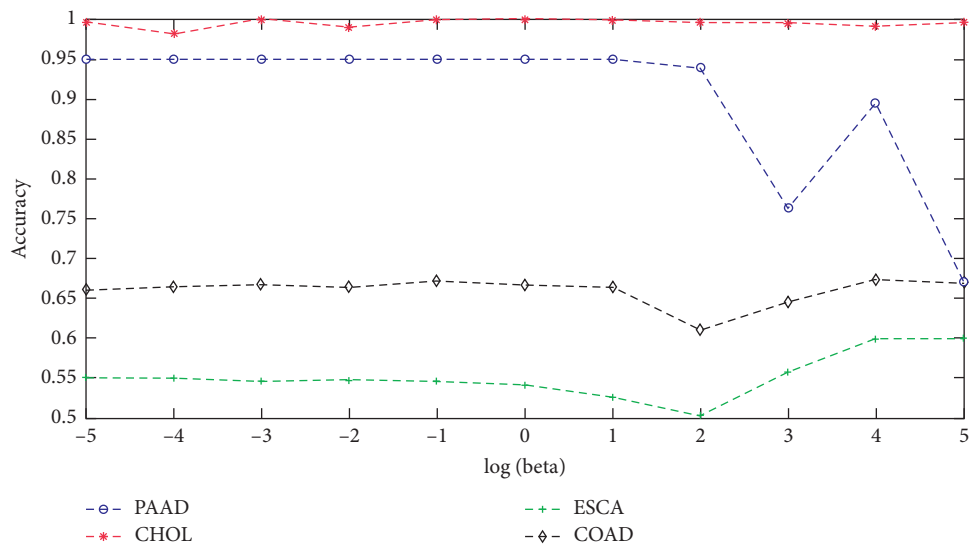

Figure 5: Accuracy of SG-jNMF varies with β.

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}},$$
$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{19}$$
$$F1 - \text{score} = \frac{2}{(1/\text{recall}) + (1/\text{precision})},$$

where TP means the number of true positives, FP is the number of false positives, and FN denotes the number of false negatives.

3.2.3. Results. In this experiment, each algorithm was run fifty times to reduce the impact of random initialization on the clustering results. We compared the accuracy, recall, precision, and F1-score of the four methods with SG-jNMF2. The mean and variance in the results are shown in Table 3. As seen in Table 3, SG-jNMF2 achieves the highest values on the four indicators mentioned above, except the recall value on the ESCA dataset. The contributions of sparse and graph regularization constraints of the algorithm are listed in Table 4. Performance improvements are measured by $\Delta_{\text{ind}} = (\text{Ind}_i - \text{Ind}_j)/(\text{Ind}_j)$, where Ind_i is the indicator of SG-jNMF and Ind_j is that of the comparison method. In particular, sparse constraints improve accuracy by 49.70%, and sparse and graph regularization constraints improve accuracy by 78.87% on the PAAD dataset. Recall and F1-score achieve more than 50% improvement on the CHOL dataset. When sparse constraints are introduced, only the recall on ESCA is reduced by 0.53%. The results on other datasets have also improved to varying degrees. In summary,

Table 3: Performance of different analysis methods.

	Datasets	iNMF	iGMFNA	jNMF	iONMF	SG-jNMF
Accuracy	PAAD	53.56 (0.00)	63.44 (3.05)	53.11 (0.01)	56.23 (0.72)	**95.00 (0.00)**
	CHOL	90.22 (1.71)	97.33 (0.04)	93.78 (0.49)	90.04 (0.96)	**99.11 (0.00)**
	ESCA	54.17 (0.00)	54.58 (0.00)	53.96 (0.00)	58.70 (1.46)	**66.87 (0.01)**
	COAD	61.87 (0.01)	63.35 (0.02)	59.15 (0.09)	62.13 (0.22)	**68.84 (0.01)**
Recall	PAAD	51.47 (4.68)	58.07 (4.84)	53.34 (4.50)	48.44 (1.20)	**67.33 (2.26)**
	CHOL	50.78 (2.01)	56.17 (1.84)	55.39 (2.04)	46.41 (1.49)	**88.06 (0.47)**
	ESCA	50.98 (0.29)	**51.31 (0.29)**	48.86 (0.27)	51.13 (0.30)	51.04 (0.12)
	COAD	49.34 (1.25)	49.15 (1.30)	49.65 (1.28)	46.73 (1.08)	**56.18 (1.51)**
Precision	PAAD	97.79 (0.05)	98.60 (0.04)	97.52 (0.05)	97.71 (0.05)	**99.09 (0.01)**
	CHOL	58.55 (0.20)	62.21 (0.20)	63.36 (0.21)	60.54 (1.72)	**91.00 (0.10)**
	ESCA	94.40 (0.05)	95.20 (0.07)	95.90 (0.08)	95.67 (0.06)	**98.39 (0.02)**
	COAD	49.34 (1.25)	49.15 (1.30)	49.65 (1.28)	46.73 (1.08)	**56.18 (0.51)**
F1-score	PAAD	63.54 (3.70)	66.70 (1.29)	67.84 (1.71)	64.42 (1.86)	**77.15 (2.12)**
	CHOL	53.47 (2.01)	58.19 (1.91)	58.03 (2.04)	51.61 (1.63)	**89.18 (1.04)**
	ESCA	66.00 (0.18)	66.47 (0.19)	64.52 (0.18)	66.41 (0.16)	**66.96 (0.16)**
	COAD	63.71 (1.34)	63.56 (1.27)	64.25 (1.30)	61.23 (1.14)	**71.11 (0.55)**

Table 4: The contribution of graph regularization and sparse constraints to clustering performance.

	Datasets	Beta = 0 (%)	Lambda = 0, beta = 0 (%)
Accuracy	PAAD	49.70	78.87
	CHOL	1.82	5.68
	ESCA	22.52	23.93
	COAD	8.67	16.38
Recall	PAAD	15.90	26.22
	CHOL	56.77	58.98
	ESCA	−0.53	4.46
	COAD	14.30	13.15
Precision	PAAD	0.49	1.60
	CHOL	46.28	43.62
	ESCA	3.35	2.59
	COAD	6.18	4.92
F1-score	PAAD	15.67	13.72
	CHOL	53.26	53.68
	ESCA	0.70	3.78
	COAD	11.88	10.68

the performance of the integrated NMF in analyzing multiomics data greatly improves by introducing sparse constraints and graph regularization constraints.

3.3. Identifying Co-DEGs. First, three matrices (DM, GE, and CNV of PAAD) are input into the SG-jNMF1 model and are projected into a common feature space. Second, we sum the common feature matrix in rows. Finally, we sort the elements in the sum vector in descending order. The top 100 genes are selected as Co-DEGs. These 100 genes are compared with pancreatic cancer genes exported from GeneCards (URL:http://www.genecards.org). Co-DEGs with relevance scores above 4 are listed in Table 5. CDKN2A is frequently mutated or deleted in many tumors. It plays an important role as a tumor suppressor gene. Studies have shown that the mutation of CDKN2A is closely related to the development of pancreatic cancer in families [31]. It is frequently seen in many tumors that mutation and overexpression of CCDN1 can alter the process of the cell cycle. Wang et al. identified pancreatitis-associated genes and found that CCND1 was involved in the pathway of pancreatic cancer [32]. Research on transcriptome sequencing shows that PTF1A maintains the expression of genes in all cellular processes. Deletion of PTF1A leads to an imbalance, cell damage, and acinar metaplasia, which is directly related to the development of pancreatic cancer [33]. Scientists have explored the effects of GRP on human intestinal and pancreatic peptides. Therefore, SG-jNMF1 can effectively integrate the information of multiomics data to identify Co-DEGs closely related to the disease.

We also use SG-jNMF1 to integrate three gene expression datasets from ESCA, CHOL, and COAD to identify Co-DEGs associated with all three diseases. Partially Co-DEGs and their relevance scores with ESCA, CHOL, and COAD are shown in Table 6. The relevance score of CHEK2 with ESCA is up to 77.66. Allelic variation in CHEK2 has a strong relationship with the risk of esophageal cancer [34]. Relevance score of CHEK2 with COAD is 29.65. The germline variation in CHEK2 is also closely related to the

Table 5: Co-DEGs identified by SG-jNMF on PAAD.

Name	Relevance score	Associated diseases	Related pathways
CDKN2A	91.19	Melanoma, cutaneous malignant 2, and melanoma-pancreatic cancer syndrome	Modulation and signaling and cell cycle role of SCF complex in cell cycle regulation
CCDN1	53.49	Multiple myeloma and Von Hippel–Lindau syndrome	ATF-2 transcription factor network and DNA damage response
PTF1A	33.66	Pancreatic and cerebellar agenesis and pancreatic agenesis 2	Developmental biology and regulation of beta-cell development
GRP	21.84	Duodenal ulcer and lung disease	Peptide ligand-binding receptors and signaling by GPCR

Table 6: Co-DEGs identified by SG-jNMF of ESCA, CHOL, and COAD.

Genes	Related diseases	Relevance score with ESCA	Relevance score with CHOL	Relevance score with COAD
CHEK2	Prostate cancer and Li–Fraumeni syndrome 2	77.66	0.35	29.65
BRAF	Cardiofaciocutaneous syndrome 1 and lung cancer	55.71	13.44	67.81
RARB	Microphthalmia, syndromic 12 and chromosome 3P deletion	28.39	2.5	23.41
NFE2L2	Immunodeficiency, developmental delay, and hypohomocysteinemia and lung squamous cell carcinoma	25.81	2.54	28.18

risk of colorectal cancer [35]. Frequent mutations in BRPA have been widely reported in human malignancies, including esophageal cancer, cholangiocarcinoma, and colon cancer [36–38]. This provides a computational method for the study of Co-DEGs in multiple diseases.

4. Conclusions

In this paper, we propose an integrative matrix factorization method (SG-jNMF) used to analyze heterogeneous multiomics data. The novel method jointly projects multiomics data matrices into a common low-dimensional space. Two forms of SG-jNMF enable multiomics data to be analyzed from both the sample and feature perspectives. This integrative analysis method can consider the local association of data and decrease the interference of noise and redundant information in the heterogeneous multiomics data. Experimental results show that the new method is superior to existing methods in analyzing heterogeneous multiomics data. Another significant advantage of SG-jNMF is that it can flexibly handle multiple input data of various types. This flexibility means that the input data can be different types of data (GE, ME, CNV, etc.) for the same disease or the same type of data for different diseases. We can use this method to identify Co-DEGs associated with a particular disease and detect common Co-DEGs associated with several diseases. This provides an efficient calculation method for biological and medical research. Next, we will use the correlation between Co-DEGs to build a gene coexpression correlation network, and further study the function of gene modules and related pathways.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported in part by the grants from the National Natural Science Foundation of China, nos. 61902215 and 61702299.

References

[1] J. N. Weinstein, E. A. Collisson, G. B. Mills et al., "The cancer genome atlas pan-cancer analysis project," *Nature Genetics*, vol. 45, no. 10, pp. 1113–1120, 2013.

[2] C. Curtis, S. P. Shah, S.-F. Chin et al., "The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups," *Nature*, vol. 486, no. 7403, pp. 346–352, 2012.

[3] Y. L. Wang and X. Wang, "Fault diagnosis of wind turbine's converter based on memristive neural network," *Applied Mechanics and Materials*, vol. 705, no. 3, pp. 333–337, 2014.

[4] J.-X. Liu, Y.-L. Gao, C.-H. Zheng, Y. Xu, and J. Yu, "Block-constraint robust principal component analysis and its application to integrated analysis of tcga data," *IEEE Transactions on Nanobioscience*, vol. 15, no. 6, pp. 510–516, 2016.

[5] S. Zhang, C.-C. Liu, W. Li, H. Shen, P. W. Laird, and X. J. Zhou, "Discovery of multi-dimensional modules by integrative analysis of cancer genomic data," *Nucleic Acids Research*, vol. 40, no. 19, pp. 9379–9391, 2012.

[6] Z. Yang and G. Michailidis, "A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data," *Bioinformatics*, vol. 30, no. 1, pp. 1–8, 2015.

[7] M. Strazar, M. Zitnik, B. Zupan, J. Ule, and T. Curk, "Orthogonal matrix factorization enables integrative analysis of multiple RNA binding proteins," *Bioinformatics*, vol. 32, no. 10, pp. 1527–1535, 2016.

[8] H.-Q. Wang, C.-H. Zheng, and X.-M. Zhao, "jnmfma: a joint non-negative matrix factorization meta-analysis of

transcriptomics data," *Bioinformatics*, vol. 31, no. 4, pp. 572–580, 2014.

[9] R. Argelaguet, B. Velten, D. Arnol et al., "Multi-omics factor analysisâĂŤa framework for unsupervised integration of multi-omics data sets," *Molecular Systems Biology*, vol. 14, no. 6, pp. 1–13, 2018.

[10] P. Ray, L. Zheng, J. Lucas, and L. Carin, "Bayesian joint analysis of heterogeneous genomics data," *Bioinformatics*, vol. 30, no. 10, pp. 1370–1376, 2014.

[11] Y.-A. Huang, Z.-H. You, X. Chen, K. Chan, and X. Luo, "Sequence-based prediction of protein-protein interactions using weighted sparse representation model combined with global encoding," *BMC Bioinformatics*, vol. 171, no. 1, p. 184, 2016.

[12] W. Min, J. Liu, and S. Zhang, "Network-regularized sparse logistic regression models for clinical risk prediction and biomarker discovery," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 15, no. 3, pp. 944–953, 2018.

[13] J. Xi, M. Wang, and A. Li, "Discovering mutated driver genes through a robust and sparse co-regularized matrix factorization framework with prior information from mirna expression patterns and interaction network," *BMC Bioinformatics*, vol. 19, no. 1, p. 214, 2018.

[14] Y. Pei, N. Chakraborty, and K. Sycara, "Nonnegative matrix tri-factorization with graph regularization for community detection in social networks," *ACM Transactions on Intelligent Systems and Technology, (TIST)*, vol. 8, no. 1, p. 1, 2016.

[15] L.-Y. Dai, C.-H. Zheng, J.-X. Liu et al., "Integrative graph regularized matrix factorization for drug-pathway associations analysis," *Computational Biology and Chemistry*, vol. 78, pp. 474–480, 2019.

[16] K. Zhan, J. Shi, J. Wang, and F. Tian, "Graph-regularized concept factorization for multi-view document clustering," *Journal of Visual Communication and Image Representation*, vol. 48, pp. 411–418, 2017.

[17] J. Wu, S. Pan, X. Zhu et al., "Boosting for multi-graph classification," *IEEE Transactions on Cybernetics*, vol. 45, no. 3, pp. 416–429, 2015.

[18] E. Zhang and X. Ma, "Regularized multi-view subspace clustering for common modules across cancer stages," *Molecules*, vol. 23, no. 5, p. 1016, 2018.

[19] J. Wang, J.-X. Liu, C.-H. Zheng, Y.-X. Wang, X.-Z. Kong, and C.-G. Wen, "A mixed-norm Laplacian regularized low-rank representation method for tumor samples clustering," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 16, no. 1, pp. 172–182, 2019.

[20] Z. Cui, "L 2, 1-grmf: an improved graph regularized matrix factorization method to predict drug-target interactions," *BMC Bioinformatics*, vol. 20, no. 8, p. 287, 2019.

[21] F. Liu, S. Xue, J. Wu et al., "Deep learning for community detection: progress, challenges and opportunities," 2020, https://arxiv.org/abs/2005.08225.

[22] J. Wu, X. Zhu, C. Zhang, and P. S. Yu, "Bag constrained structure pattern mining for multi-graph classification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 10, pp. 2382–2396, 2014.

[23] Z. Zhang, "Jointly learning structured analysis discriminative dictionary and analysis multiclass classifier," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 8, pp. 3798–3814, 2017.

[24] Z. Zhang, Y. Zhang, S. Li et al., "Flexible auto-weighted local-coordinate concept factorization: a robust framework for unsupervised clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 1, p. 1, 2019.

[25] L.-Y. Dai, C.-M. Feng, J.-X. Liu, C.-H. Zheng, J. Yu, and M.-X. Hou, "Robust nonnegative matrix factorization via joint graph laplacian and discriminative information for identifying differentially expressed genes," *Complexity*, vol. 2017, Article ID 4216797, 11 pages, 2017.

[26] H. S. Seung and D. D. Lee, "The manifold ways of perception," *Science*, vol. 290, no. 5550, pp. 2268-2269, 2000.

[27] F. Nie, H. Huang, X. Cai, and C. H. Ding, "Efficient and robust feature selection via joint l2, 1-norms minimization," in *Proceedings of Advances in Neural Information Processing Systems*, pp. 1813–1821, British Columbia, Canada, December 2010.

[28] D. Cai, "Graph regularized nonnegative matrix factorization for data representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1548–1560, 2010.

[29] M.-X. Hou, Y.-L. Gao, J.-X. Liu, L.-Y. Dai, X.-Z. Kong, and J. Shang, "Network analysis based on low-rank method for mining information on integrated data of multi-cancers," *Computational Biology and Chemistry*, vol. 78, pp. 468–473, 2019.

[30] N. Yu, Y.-L. Gao, J.-X. Liu, J. Shang, R. Zhu, and L.-Y. Dai, "Co-differential gene selection and clustering based on graph regularized multi-view NMF in cancer genomic data," *Genes*, vol. 9, no. 12, p. 586, 2018.

[31] D. B. Zhen, K. G. Rabe, S. Gallinger et al., "Brca1, brca2, palb2, and cdkn2a mutations in familial pancreatic cancer: a pacgene study," *Genetics in Medicine*, vol. 17, no. 7, p. 569, 2015.

[32] D. Wang, Z.-M. Zhu, Y.-L. Tu et al., "Identfication of key miRNAs in pancreatitis using bioinformatics analysis of microarray data," *Molecular Medicine Reports*, vol. 14, no. 6, pp. 5451–5460, 2016.

[33] C. Q. Hoang, M. A. Hale, Ana C. Azevedo-Pouly et al., "Transcriptional maintenance of pancreatic acinar identity, dierentiation, and homeostasis by ptf1a," *Molecular and Cellular Biology*, vol. 36, no. 34, pp. 3033–3047, 2016.

[34] H. Gu, W. Qiu, Y. Wan et al., "Variant allele of CHEK2 is associated with a decreased risk of esophageal cancer lymph node metastasis in a Chinese population," *Molecular Biology Reports*, vol. 39, no. 5, pp. 5977–5984, 2012.

[35] L. H. Williams, D. Choong, S. A. Johnson, and I. G. Campbell, "Genetic and epigenetic analysis of CHEK2 in sporadic breast, colon, and ovarian cancers," *Clinical Cancer Research*, vol. 12, no. 23, pp. 6967–6972, 2006.

[36] C. H. Maeng, J. Lee, P. van Hummelen et al., "High-throughput genotyping in metastatic esophageal squamous cell carcinoma identifies phosphoinositide-3-kinase and BRAF mutations," *PloS One*, vol. 7, no. 8, Article ID e41655, 2012.

[37] S. Robertson, O. Hyder, R. Dodson et al., "The frequency of KRAS and BRAF mutations in intrahepatic cholangiocarcinomas and their correlation with clinical outcome," *Human Pathology*, vol. 44, no. 12, pp. 2768–2773, 2013.

[38] S. Ogino, K. Nosho, G. J. Kirkner et al., "CpG island methylator phenotype, microsatellite instability, BRAF mutation and clinical outcome in colon cancer," *Gut*, vol. 58, no. 1, pp. 90–96, 2009.