frontiers in Genetics

ORIGINAL RESEARCH
published: 20 August 2019
doi: 10.3389/fgene.2019.00744

Multi-view Subspace Clustering Analysis for Aggregating Multiple Heterogeneous Omics Data

Qianqian Shi[1], Bing Hu[2], Tao Zeng[3,4] and Chuanchao Zhang[5]**

[1] Hubei Key Laboratory of Agricultural Bioinformatics, College of Informatics, Huazhong Agricultural University, Wuhan, China, [2] Department of Applied Mathematics, College of Science, Zhejiang University of Technology, Hangzhou, China, [3] Key Laboratory of Systems Biology, Institute of Biochemistry and Cell Biology, Shanghai Institute of Biological Sciences, Chinese Academy of Sciences, Shanghai, China, [4] Shanghai Research Center for Brain Science and Brain-Inspired Intelligence, Shanghai, China, [5] Wuhan Institute of Huawei Technologies, Wuhan, China

Integration of distinct biological data types could provide a comprehensive view of biological processes or complex diseases. The combinations of molecules responsible for different phenotypes form multiple embedded (expression) subspaces, thus identifying the intrinsic data structure is challenging by regular integration methods. In this paper, we propose a novel framework of "Multi-view Subspace Clustering Analysis (MSCA)," which could measure the local similarities of samples in the same subspace and obtain the global consensus sample patterns (structures) for multiple data types, thereby comprehensively capturing the underlying heterogeneity of samples. Applied to various synthetic datasets, MSCA performs effectively to recognize the predefined sample patterns, and is robust to data noises. Given a real biological dataset, i.e., Cancer Cell Line Encyclopedia (CCLE) data, MSCA successfully identifies cell clusters of common aberrations across cancer types. A remarkable superiority over the state-of-the-art methods, such as iClusterPlus, SNF, and ANF, has also been demonstrated in our simulation and case studies.

Keywords: multi-view subspace clustering analysis, data integration, heterogeneity, low-rank representation, graph diffusion

OPEN ACCESS

Edited by:
Shihua Zhang,
Academy of Mathematics and Systems Science (CAS), China

Reviewed by:
Wenyuan Li,
University of California, Los Angeles, United States
Jinyu Chen,
Academy of Mathematics and Systems Science (CAS), China

****Correspondence:***
Qianqian Shi
qqshi@mail.hzau.edu.cn
Chuanchao Zhang
chaozhangchuan@163.com

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

Received: *21 December 2018*
Accepted: *16 July 2019*
Published: *20 August 2019*

Citation:
Shi Q, Hu B, Zeng T and Zhang C (2019) Multi-view Subspace Clustering Analysis for Aggregating Multiple Heterogeneous Omics Data. Front. Genet. 10:744. doi: 10.3389/fgene.2019.00744

INTRODUCTION

The rapid advance of high throughput technologies makes large amounts of various omics data available to study biological problems (Schuster, 2008). While, different types of data could provide complementary or common information to each other since a biological system consists of a series of highly ordered molecular and cellular events (Wang et al., 2014; Ma and Zhang, 2017; Shi et al., 2017a). Thus, compared to single data types (e.g., gene expression), the integration of multiple omics data is more likely to completely understand the molecular mechanisms underlying particular biological processes or complex diseases, and therefore offers more opportunities to better address biological or medical issues, e.g., to identify cancer subtypes with different biological or clinical outcomes (Xiong et al., 2012; Chen and Zhang, 2016; Shi et al., 2017a).

So far, quite a lot of data-integration methods have been proposed and they can be briefly summarized into two main categories: firstly, to extract signals from each data type; secondly, to acquire comprehensive information by a sample-centric integration (Arneson et al., 2017; Zhang et al., 2017a). In addition, these data integration methods mainly depend on two strategies, one is space projection method (Fan et al.,

2016), and the other one is metric (similarity measures) fusion technique (Wang et al., 2014). These ideas match the nonlinear characteristics of biological systems and should really work when capturing the whole phenotype landscape.

However, their solutions to obtain the sample or gene patterns from multiple data domains are really distinct from each other. The earliest proposed methods identify multi-dimensional genomic modules (e.g., mRNA-miRNA functional pairs) (Ghazalpour et al., 2006; Kutalik et al., 2008; Li et al., 2012; Zhang et al., 2012; Chen and Zhang, 2016), which present high correlations over the samples in data sets. Such "co-modules" can only uncover common sample structures across data types and likely lead to biased clustering because much phenotype-associated differential information is missing. Later, Mo et al. developed a method, iClusterPlus (Mo et al., 2013), which considers different properties of omics data (e.g., continuous, count or binary valued variables) through corresponding linear regression models. However, some assumptions held by this method are too strong for heterogeneous tumor samples, and may also lose biologically meaningful information. As a nearly assumption-free and fast approach, SNF (Wang et al., 2014) (similarity network fusion) can overcome such issues and it uses local structure preservation method (i.e., *K*-nearest neighbors) to adjust sample similarity networks for each data type. But, SNF can only characterize pair-wise Euclidean (or other) distances in the sample neighborhoods, and is sensitive to local data noises or outliers. Recently, Ma and Zhang proposed ANF, an "update" of SNF, which incorporates weights of views for each data type (Ma and Zhang, 2017). ANF presents more general and interpretable power than SNF, but it still reserves the unstable nature of pair-wise clustering. Notably, increasing biological evidence suggests that distinct regulatory mechanisms preside over physiological phenotypes (e.g., Waddington's canalization) or even the tumor cell states (Mark and Aviv, 2002). Cell types or patients present extremely strong heterogeneity due to the different master gene sets, implying that these individuals are scattered in multiple biological states (feature subspaces) even at a single data level (Shi et al., 2017b; Haghverdi et al., 2018). That means the pair-wise similarity measurement (e.g., in SNF) can't capture the true heterogeneity spanning in different subspaces, further leading to inaccurate integrative clustering. Thus, the more effective integration approach is still lacking.

Motivated by above requirements from methodology and biology study, we propose a novel framework called "Multi-view Subspace Clustering Analysis (MSCA)" by using representation-based methods (e.g., low-rank representation, namely LRR) (Lin et al., 2011; Liu et al., 2013). LRR or relevant subspace clustering algorithms are originally developed and applied in image recognition (Zhang et al.; Cao et al., 2015; Gao et al., 2016; Brbić and Kopriva, 2017; Zhang et al., 2017b). These methods enable to recover the signal spaces of the images, providing a better description of the visual patterns. Furthermore, they generate a block-diagonal representation graph of samples, which measures sample similarities by linear combinations of the remaining samples, presenting more robust than pair-wise clustering. However, when applied to highly heterogenous data, such as biological omics profiles, these methods are often fragile since they assume linear embedded structures underlie the original data and can't exploit the local geometric relationships of objects (Zhuang et al., 2015). Hence, we should improve the utility of subspace clustering to be more appropriated for biological cases. In our proposed MSCA model, we incorporate the advantage of local structure preservation to force the representations to be locally linear at each data type, and capture the integrative clustering pattern by fusing the multiple informative graphs from local sample representations. In particular, MSCA implements two steps of nonlinear pattern identification for different omics data during pattern fusion, where the multi-view is able to recover more details of systems' complexity and heterogeneity. To validate the effectiveness of our method, we firstly applied MSCA to various synthetic datasets, and found that MSCA not only successfully recognizes the predefined subgroups with a better performance than several state-of-the-art methods, but also shows great robustness on different parameters' variation. In addition, MSCA has demonstrated a good ability to yield biologically relevant subgroups of tumor cells of multiple origins in CCLE (Barretina et al., 2012) data set.

METHODS

Method Overview

MSCA takes two steps as schematically shown in **Figure 1**: i) Construction of sample representation matrix from each type of genomic profiles by a subspace clustering algorithm (**Figures 1A**, **B**); ii) Graph diffusion process of sample similarity matrices, which are derived from the representation matrices corresponding to all data types (**Figure 1C**). MSCA was implemented as a Matlab package and is freely available at https://github.com/ZCCQQWork/MSCA.

The representation graph Z of step (i) presents each single sample as a linear combination of the remaining ones in the same subspace/cluster, and therefore it can be shown as a block-diagonal and sparse matrix. Such low-rank characteristic of Z makes it more robust to data outliers and capable to retain more structural information of data, thus paving a good way for the next integrative. After that, MSCA implements the graph diffusion step (ii). It makes information propagate across multiple graphs in an iteration way. And this could fuse biological signals from the involved genomic data. After a few iterations, MSCA converges to the optimal graph (**Figure 1D**), as a multi-view similarity measurement, revealing the underlying relationship of samples. Note that both the steps follow nonlinear criteria, to maximize the chance of characterizing the true complexity and heterogeneity of data, and especially the common information will strengthen the supported sample patterns whereas discordant local structures will weaken their similarities.

Extracting the Sample Representation Graph From Each Data Type

Suppose we describe a genomic profile (e.g., mRNA expression) with h biological measurements and n samples as a data matrix $X = [x_1, x_2, \ldots, x_n]$, x_i and x_j correspond to two samples; then the representation relationships of all samples can be calculated as follows:

$$\min_{Z,E} \|Z\|_* + \lambda \|E\|_{2,1} \quad \text{s.t.} \begin{cases} X = XZ + E \\ Z^T \mathbf{1} = \mathbf{1} \\ Z_{ij} = 0, (i,j) \in \bar{\Omega} \end{cases} \tag{1}$$

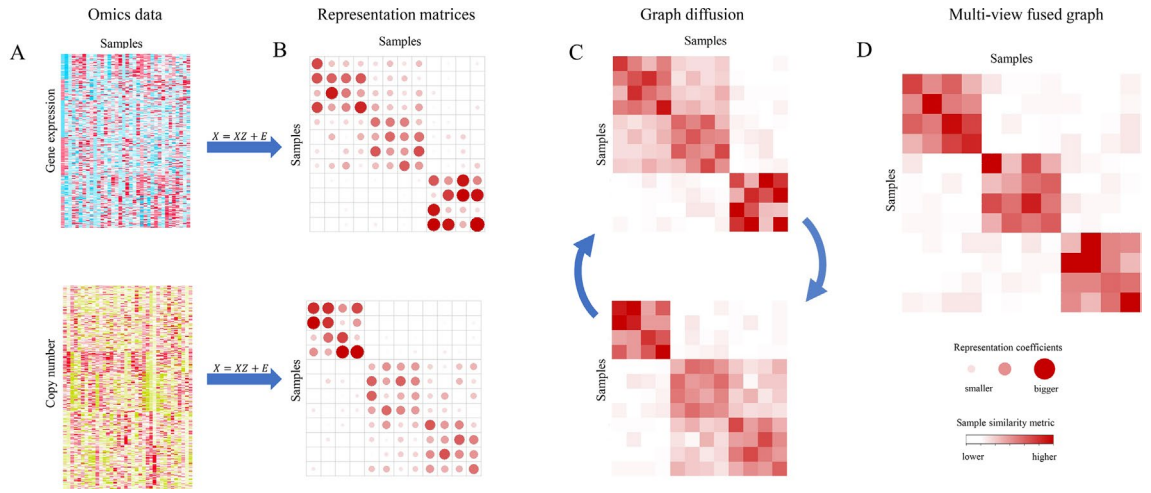

FIGURE 1 | Overview of Multi-view Subspace Clustering Analysis (MSCA). **(A)** Different biological data types for the same set of samples. **(B)** Sample representation matrices for each data type. Coefficients are represented by dots, and bigger redder ones mean larger values. **(C)** Cross-graph diffusion process to integrate multiple similarity matrices, updating aggregated information iteratively. **(D)** Final integrative result when step in **(C)** reaches convergence. Color of square in graphs indicates sample-to-sample similarities. *X* denotes as each omic data matrix, and *Z* presents the representation matrix. *E* is the error matrix (see *Methods*).

where $Z = [z_1, z_2, \ldots, z_n]$ is a $n \times n$ matrix containing all the coefficient measurements between pairs of samples $x_i (1 \le i \le n)$, and z_i is a coefficient vector of sample i. $\|\mathbf{Z}\|_*$ represents the nuclear norm of $\mathbf{Z}$, i.e., the sum of all singular values of Z; $\|E\|_{2,1} = \sum_{j=1}^{n} \sqrt{\sum_{i}^{h} \left(e_{ij}\right)^2}$ and is $l_{2,1}$-norm of the error matrix E, where e_{ij} is the (i,j)-$_{\text{th}}$ entry of matrix E.

Note that, in the first constraint condition, the linear representation of samples can capture the global structure in data, thus a large similarity coefficient means the two samples are spatially close. Next in the second constraint condition, $\mathbf{1}$, as an all-one vector, is used to normalize Z that $\sum_i Z_{ij} = 1$. And in the third constraint condition, $\bar{\Omega}$ denotes as the complement of Ω, where Ω is a set of edges between the samples in a predefined adjacency graph. For example, if x_i and x_j are not graph neighbors, we have $(i,j) \in \bar{\Omega}$. In this work, we use K-nearest neighbors to predetermine the sample local structure in terms of pair-wise Euclidean distances. Then, the tuning parameter λ is used to balance the two optimization terms, which could be selected according to their respective properties, or tuned empirically. For the selection of parameters K and λ, the section *Evaluation of MSCA on Synthetic Examples* has more detailed discussions. Given solving problem (1), we obtain the optimal solution Z^*, which is block-diagonal indicating that samples in the same subspace are clustered together due to the comprehensive considerations/constraints of global and local data structures. The corresponding sample affinity matrix W is obtained by $W = \left(\left|Z^*\right| + \left|Z^{*T}\right|\right)/2$, which can be passed on to the next step integration.

In fact, the optimization problem (1) can be solved *via* ADMM (alternating direction method of multipliers) algorithm (Lin et al., 2010) as below. Firstly, this problem can be converted to an equivalent problem:

$$\min_{L_{\bar{\Omega}}(Z)=0, E} \|J\|_* + \lambda \|E\|_{2,1}$$
$$\text{s.t.} \begin{cases} X = XZ + E \\ Z^T \mathbf{1} = \mathbf{1} \\ J = Z \end{cases} \tag{2}$$

And its augmented Lagrangian function is:

$$L_\mu(Z,E,J) = \|J\|_* + \lambda\|E\|_{2,1} + \langle Y_1, X - XZ - E\rangle + \langle Y_2, \mathbf{1}^T - \mathbf{1}^T Z\rangle + \langle Y_3, Z - J\rangle + \frac{\mu}{2}\left(\|X - XZ - E\|_F^2 + \|\mathbf{1}^T - \mathbf{1}^T Z\|_F^2 + \|Z - J\|_F^2\right) \tag{3}$$

where μ is a penalty parameter larger than 0. $||*||_{\text{F}}$ denotes the Frobenious norm, and Y_1, Y_2 and Y_3 are Lagrangian multipliers corresponding to three constraints in equation (2) respectively; $L_{\bar{\Omega}}(Z) = 0$ corresponds to the third constraint condition in original optimization equation (1). As known, the above problem can be minimized orderly to update the variables Z, J, E by fixing the other variables, respectively, according to ADMM.

Suppose at k times of updates, we acquire $Z^k, J^k, E^k, Y_1^k, Y_2^k$ and Y_3^k, and the alternate process with update functions can be summarized in below:

Firstly, assuming all the other five matrices are fixed, we can compute J^{k+1}:

$$\begin{aligned} J^{k+1} &= \arg\min_J \|J\|_* + \langle Y_3^k, Z^k - J\rangle + \frac{\mu^k}{2}\|Z^k - J\|_F^2 \\ &= \arg\min_J \|J\|_* + \frac{\mu^k}{2}\left\|Z^k + \frac{Y_3^k}{\mu^k} - J\right\|_F^2 \end{aligned} \tag{4}$$

Secondly, assuming J^{k+1}, Z^k, Y_1^k are fixed, we can compute E^{k+1}:

$$E^{k+1} = \arg\min_E \lambda\|E\|_{2,1} + \langle Y_1^k, X - XZ^k - E\rangle + \frac{\mu^k}{2}\|X - XZ^k - E\|_F^2$$
$$= \arg\min_E \lambda\|E\|_{2,1} + \frac{\mu^k}{2}\left\|X - XZ^k + \frac{Y_1^k}{\mu^k} - E\right\|_F^2 \tag{5}$$

Thirdly, assuming J^{k+1}, E^{k+1}, Y_1^k, Y_2^k and Y_3^k are fixed, we can compute the updated Z from following optimization problem:

$$\min_{L_{\bar{\Omega}}(Z)=0} \langle Y_1^k, X - XZ - E^{k+1}\rangle + \langle Y_2^k, \mathbf{1}^T - \mathbf{1}^T Z\rangle + \langle Y_3^k, Z - J^{k+1}\rangle$$
$$+ \frac{\mu^k}{2}\left(\|X - XZ - E^{k+1}\|_F^2 + \|\mathbf{1}^T - \mathbf{1}^T Z\|_F^2 + \|Z - J^{k+1}\|_F^2\right) \tag{6}$$

In fact, this problem is equivalent to

$$\min_{L_{\bar{\Omega}}(Z)=0} \left\|X - XZ - E^{k+1} + \frac{Y_1^k}{\mu^k}\right\|_F^2 + \left\|\mathbf{1}^T - \mathbf{1}^T Z + \frac{Y_2^k}{\mu^k}\right\|_F^2$$
$$+ \left\|Z - J^{k+1} + \frac{Y_3^k}{\mu^k}\right\|_F^2 \tag{7}$$

Then, it can be further linearized with respect to Z *at* Z^k based on LADMAP (linearized alternating direction method with adaptive penalty) algorithm (Lin et al., 2011):

$$\min_{L_{\bar{\Omega}}(Z)=0} \left\langle -X^T\left(X - XZ^k - E^{k+1} + \frac{Y_1^k}{\mu^k}\right) - \mathbf{1}\left(\mathbf{1}^T - \mathbf{1}^T Z^k + \frac{Y_2^k}{\mu^k}\right)\right.$$
$$\left. + \left(Z^k - J^{k+1} + \frac{Y_3^k}{\mu^k}\right), Z - Z^k\right\rangle + \frac{\eta}{2}\|Z - Z^k\|_F^2 \tag{8}$$

where $\eta = \|X\|_2^2 + \|\mathbf{1}^T\|_2^2 + 1$.

In the end, we obtain Z^{k+1} according to the following updating rule:

$$Z^{k+1} = \arg\min_{L_{\bar{\Omega}}(Z)=0} < H^k, Z - Z^k > + \frac{\eta}{2}\|Z - Z^k\|_F^2$$
$$= \arg\min_{L_{\bar{\Omega}}(Z)=0} \frac{\eta}{2}\left\|Z - Z^k + \frac{H^k}{\eta}\right\|_F^2$$
$$= \begin{cases} \left(Z^k - \frac{H^k}{\eta}\right)_{ij}, (i,j) \in \Omega \\ 0, \quad (i,j) \in \bar{\Omega} \end{cases} \tag{9}$$

where

$$H^k = -X^T\left(X - XZ^k - E^{k+1} + \frac{Y_1^k}{\mu^k}\right) - \mathbf{1}\left(\mathbf{1}^T - \mathbf{1}^T Z^k + \frac{Y_2^k}{\mu^k}\right)$$
$$+ \left(Z^k - J^{k+1} + \frac{Y_3^k}{\mu^k}\right)$$

Fourthly, assuming that E^{k+1}, Z^{k+1} and J^{k+1} are fixed, we can calculate simultaneously:

$$Y_1^{k+1} = Y_1^k + \mu^k(X - XZ^{k+1} - E^{k+1}) \tag{10}$$

$$Y_2^{k+1} = Y_2^k + \mu^k(\mathbf{1}^T - \mathbf{1}^T Z^{k+1}) \tag{11}$$

$$Y_3^{k+1} = Y_3^k + \mu^k(Z^{k+1} - J^{k+1}) \tag{12}$$

All the above subproblems can form a closed loop until convergence, and the whole step to derive the graph weight matrix W can be briefly summarized in Algorithm 1.

ALGORITHM 1 Algorithm to extract the sample representation matrix for each data type.

Input: the profile of i_{th} data type, i.e. $X^i = \left[x_1^i, x_2^i, ..., x_n^i\right]$, tuning parameter λ, and nearset neighbors parameter K.

Output: the sample representation matrix W^i of i_{th} data type.

1. Obtain neighbors in data X^i using K-nearset neighbour method, and assign the parameter Ω

2. Solve the equation (1) by updating (4), (5), (9)-(12) until the iteration converges and obtain the optimal Z^*

3: Construct the sample similarity matrix W^i by $W^i = \left(|Z^*| + |Z^{*T}|\right)/2$

Capturing Multi-View Graph From Various Omics Data

Given m different genomics data types, we could obtain respective affinity matrices W^i, $i = 1, 2, \ldots, m$ as nonlinear similarity measurements of all samples by above Algorithm 1. This step would fuse individual affinity graphs to a systematic one. The graph diffusion process is implemented like SNF ever does (Wang et al., 2014). In this step, we continue to take advantage of locality-preserving strategy and define a kernel matrix, S, to ensure samples in the same neighborhood still stay close across data sources. Simultaneously, we normalized the raw affinity matrix W to a new status matrix P, which keeps the original information and reduces the scale bias. Note that matrix P still carries the full information about the similarity of each sample to all others whereas matrix S only encodes the similarity to the local neighborhoods for each sample.

For the m different biological data types, matrices P^i and S^i of the i-th data type are obtained by equations (13) and (14) based on (W^i, $i = 1, 2, \ldots, m$).

$$P^i(i,j) = \begin{cases} \frac{W^i(i,j)}{2\sum_{k \neq i} W^i(i,k)}, j \neq i \\ 1/2, j = i \end{cases} \tag{13}$$

$$S^i(i,j)=\begin{cases}\dfrac{W^i(i,j)}{\sum_{k\in N_i}W^i(i,k)}, j\in N_i\\ 0, \quad otherwise\end{cases} \tag{14}$$

where N_i is the K nearest neighbors of the sample x_i based on W^i.

The key step of MSCA is to iteratively update status matrix in graph diffusion across data types as follows:

$$\begin{aligned}P^1_{t+1}&=S^1\times\left(\frac{\sum_{k\neq1}P^k_t}{m-1}\right)\times(S^1)^T\\ &\cdots\\ P^i_{t+1}&=S^i\times\left(\frac{\sum_{k\neq i}P^k_t}{m-1}\right)\times(S^i)^T\\ &\cdots\\ P^m_{t+1}&=S^m\times\left(\frac{\sum_{k\neq m}P^k_t}{m-1}\right)\times(S^m)^T\end{aligned} \tag{15}$$

where P^i_{t+1} is the status matrix of i-th data type after $t+1$ iterations and $P^i_1=P^i$ represent the initial status matrix at $t=1$.

The equation (15) updates the status matrices each time generating m parallel interchanging diffusion processes. After t steps, the overall status matrix or multi-view matrix $W^\#$ is computed as:

$$W^\#=\frac{\sum_{i=1}^{m}P^i_t}{m} \tag{16}$$

Iterative Updating Process and Clustering Method

Given a series of sample representation matrices generated by Algorithm 1, the iterative integration process is summarized as Algorithm 2.

ALGORITHM 2 The Iterative Updating Process for MSCA.

Input: The profile of the m data types, i.e., $X=[X^1,X^2,...,X^m]$, tuning parameter λ, and nearset neighbors parameter K.
Output: The multi-view similarity matrix $W^\#$ across m data types
1. Computing the representation matrix W^i (i = 1,2,..m) of each data type according to Algorithm 1
2. Updating the status matrix P^i (i = 1,2...m) of each data type by the equation (15) until the process reaches convergence
3. Capturing the multi-view similarity matrix W$^\#$ by the equation (16)

Therefore, the final undirected graph $W^\#$, involving multi-layer signals, i.e., local and global information, is capable to present the intrinsic complexity of data. The multi-view fused matrix can be applied into spectral clustering algorithm [e.g., Ratio Cuts (Ding et al., 2013)] to identify the meaningful groups of samples, e.g., prognostic different subtypes, or other potential applications.

RESULTS

Evaluation of MSCA on Synthetic Examples

To demonstrate the ability of MSCA on multi-view subgroups identification, simulation experiments are conducted, with comparison to the above mentioned methods (Mo et al., 2013; Wang et al., 2014; Cao et al., 2015; Brbić and Kopriva, 2017; Ma and Zhang, 2017; Zhang et al., 2017b). In addition, the selection of parameters in MSCA has also been discussed in these synthetic examples.

Synthetic Data

Two categories of numeric data sets have been considered for a complete evaluation. Each contains two types of data and 90 samples underlying predefined sample structures by singular value decomposition (Meng et al., 2015). To preserve feature characteristics (e.g., amount, diversity and variance) of biological data types (e.g., gene expression and methylation profiles), the two data types in synthetic examples are directly generated from real data sets (i.e., GSE49278 and GSE49277) (Assié et al., 2014) (**Supplementary Information**). And each data type could provide partial but effective information to describe the whole sample patterns (e.g., type 1 and type 2 in **Figure 2A** and **Supplementary Figure S1A**). We called the "weak heterogeneity" numeric example as simData1 where samples are distributed in a single subspace and the "strong heterogeneity" one as simData2 where different manifold subspaces exist. Briefly, the 90 samples with three established clusters (namely, 1-30, 31-60, 61-90) in simData1 and simData2 are randomly selected from real data, where samples 31-90 present similar distributions from data type 1; and 1-60 appear close from data type 2. But the samples in 31-90 and 1-60 would have different embedded structures or manifold subspaces. Note that the true clusters cannot be recovered by any single data type in both synthetic examples (**Figure 2A** and **Supplementary Figure S1A**).

Evaluation and Comparison Based on Cluster Identification

We first applied MSCA and the other methods to the generated data sets (i.e., simData1 and simData2) with predetermined clustering structures. To avoid accidental events, both the data sets were randomly repeated 500 times under different systematic conditions (i.e., low: 0% extra noises; moderate: 10% extra noises; high: 30% extra noises), respectively. And the performance of each algorithm was measured by adjusted Rand index (ARI) (Santos and Embrechts, 2009), and a high value indicates an identical clustering. According to all the results, MSCA always succeeded to piece the information of each data type together, brilliantly distinguishing the pre-designed three clusters (**Figure 2B** and **Supplementary Figure S1B**). Given simData1 of less heterogeneity, all of the compared methods almost perform excellent (**Supplementary Figure S1B**). However, when complexity increases, a great performance difference among different methods comes out. Our MSCA model still performed accurately and robustly to identify sample patterns, even across varying noise strengths (**Figure 2B**). But the pair-wise clustering-based methods,

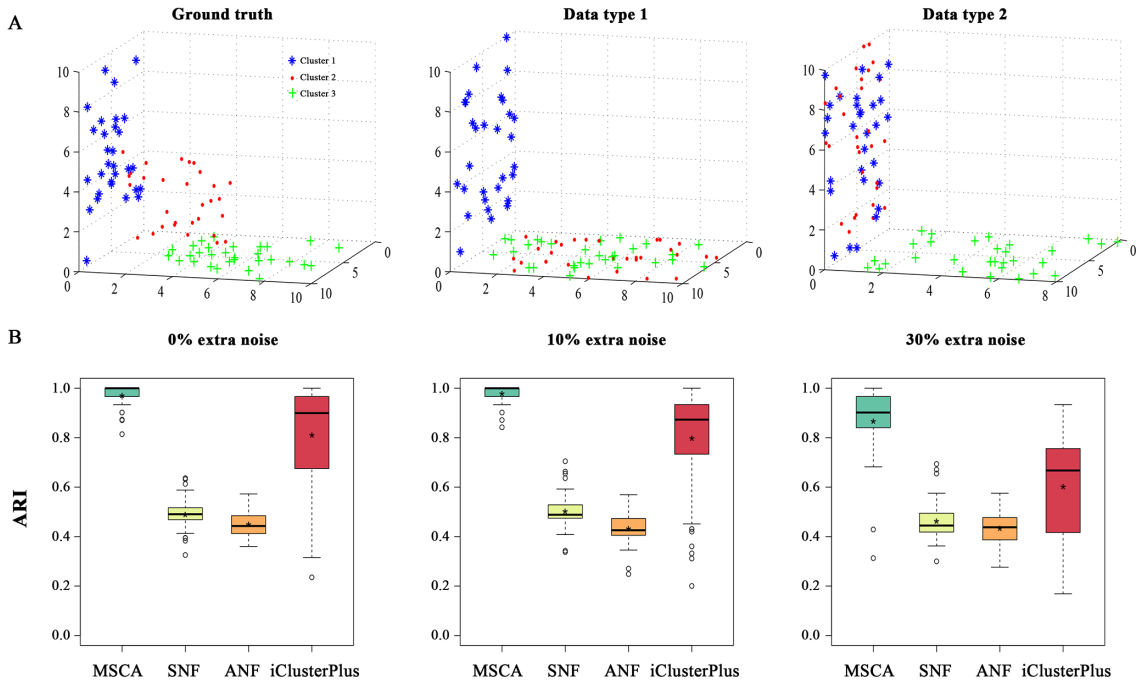

FIGURE 2 | A simulation study on simData2. **(A)** 3D Illustration of sample patterns in different feature spaces. Data points, i.e., samples, are colored and shaped by their true cluster labels. Clean cluster boundaries only can be seen in an integrative affine space. Points in two clusters may be mislabeled in a single coordinated space, i.e., Cluster 2 and Cluster 3 for data type 1, Cluster 1 and Cluster 2 for type 2. **(B)** The clustering accuracy comparison among MSCA, SNF, ANF and iClusterPlus under different noise conditions, measures their effectiveness on detecting integrated sample-patterns.

i.e., SNF and ANF, obviously can't recognize the multiple manifolds embedded in high-dimensional space. Even for those subspace clustering algorithms, they didn't perform that well when integrating data sets with biological characteristics (**Supplementary Figure S2**), thus highlighting the feasibility of MSCA for biological cases. While, iClusterPlus performed the second best on accuracy, but the accuracy ranges manifested "long-tail" to expose the unstable nature of iClusterPlus. It's probably because iClusterPlus uses random sampling procedure to solve equations (Mo et al., 2013), and is sensitive to data noises. In all, the novel nonlinear similarity measurement in MSCA is demonstrated to be robust to data noises and heterogeneity, which helps provide a more accurate multi-view for sample patterns in multi-level dataset.

Robustness Analysis of MSCA Under Different Parameters

There are two parameters, i.e., λ and K (see *Methods*), in MSCA model, thus it is crucially important to examine their effects on the MSCA performance. In particular, the parameter K determines the predefined neighborhoods, which constrains the solutions of sample representation matrices. Under different selections of K or λ, we use simData2 to test the robustness of MSCA. To avoid results by chance, we repeated 1,000 times and take the average ARI values as evaluation measurement. According to all the results (**Figure 3**), MSCA performs stable and accurate in a wide range of K and λ. Once again, the advantage of combining low-rank presentation and local preservation makes MSCA more parameter-independent, and brings a novel light on developing new bioinformatic tools for integrating heterogeneous biological data.

Study on CCLE Data

To demonstrate the effectiveness of MSCA to address practical issues, we have applied MSCA to CCLE datasets (Barretina et al., 2012) with the matched mRNA expression profiles by Affymetrix Human Genome U133 Plus 2.0 array and copy number data by Affymetrix SNP Array 6.0. Though it contains thousands of cell lines, we only kept 415 cell lines, whereby more than 25 cells have the same tissues of origin (**Supplementary Table S1**). For each tissue, we obtained its specific expressed genes from two databases: The Human Protein Atlas (Uhlen et al., 2015) and PaGenBase (Pan et al., 2013). Several organs belong to upper aerodigestive tract cancer (UADT), including tongue, trachea and esophagus etc., thus, all their gene sets were treated as UADT specific genes. While tumor associated genes were collected from GeneCards (Safran et al., 2010) and top 100 by the provided relevance scores were selected to illustrate corresponding aberration patterns among different subgroups. We adopted one-sided Wilcoxon signed-rank test to identify the tissue-specific genes between one of the clusters and all the remaining ones. More highly expressed genes with $P <$ 0.05 (adjusted by FDR) indicate the cluster strongly correlated with a certain tissue of origin. Similarly, differential expression or copy number was calculated using two-sided Wilcoxon signed-rank test for each single gene. A significant P-value shows gene expression or copy number in one group dominates the other cell lines and we

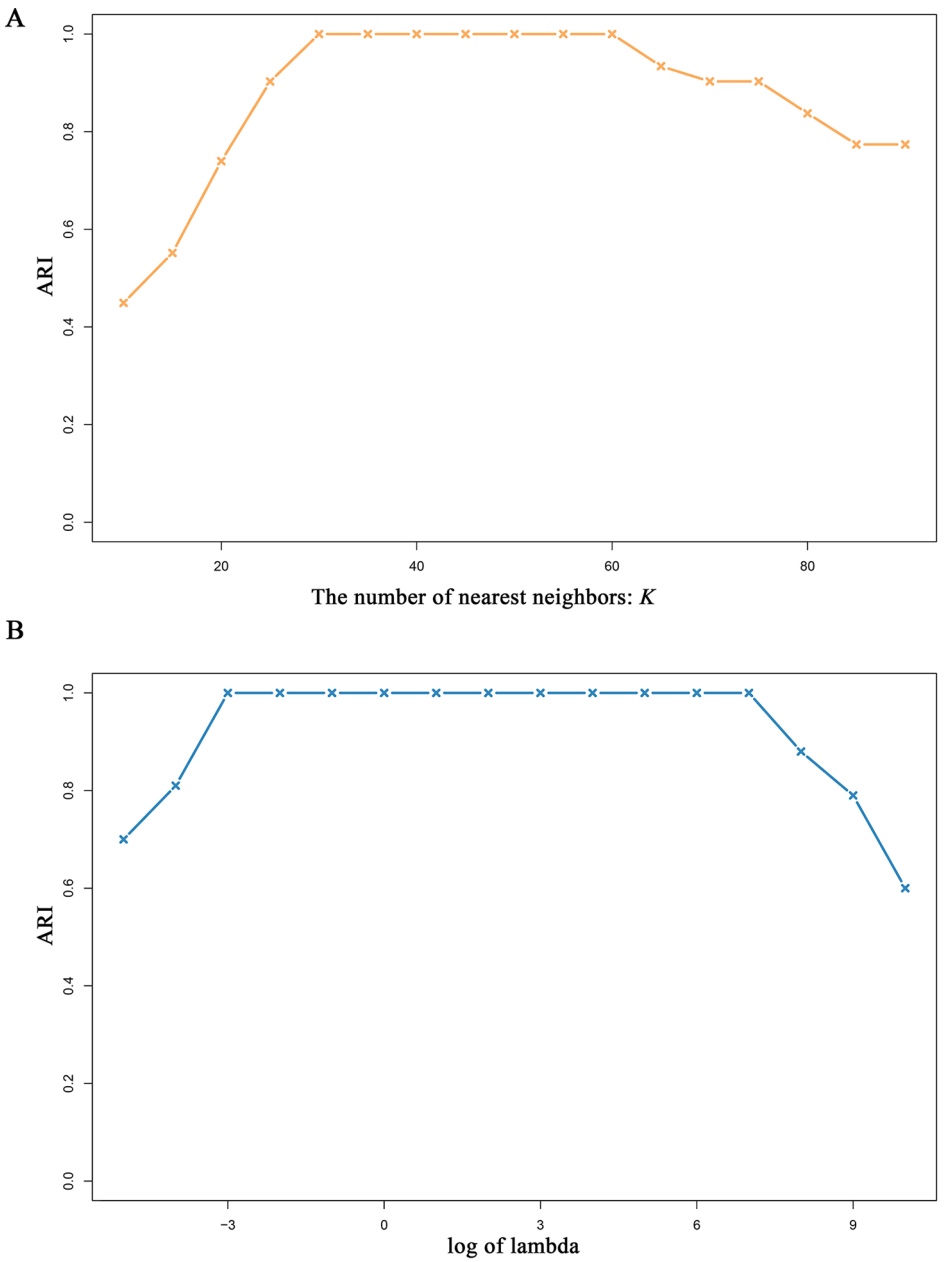

FIGURE 3 | Performance of MSCA under different parameters. Varying selected nearest neighbor number *K* or the tuning parameter λ, MSCA identifies the predefined clusters in simData2.

regard those differential genes with $P < 0.05$ (after FDR correction) as cluster markable features. Though clusters may share markable features, we count the number of shared clusters to measure the inter-cluster heterogeneity.

Firstly, we used the silhouette score (Rousseeuw, 1999) to evaluate how coherent the identified clusters are, and then we assigned the cell lines into nine clusters (**Supplementary Figure S3**). Among the compared methods, we observed MSCA had a better silhouette score, indicating superior subgroup identification for CCLE samples (**Figure 4A**). Then, we compared the integrative clusters with the original tissue groups (**Figure 4B**), and found some cell lines still manifest high lineage dependency (Pearson correlation 0.42). For example, all the AML or M. myeloma cell lines are assigned to single clusters (i.e., cluster1 and cluster5, respectively), separating from other solid tumor ones. Accordingly, the cluster1 preserves about 77% blood genes and cluster5 holds 85% lymph associated genes (**Supplementary Figure S4**). Besides, the characteristic preservation of tissue specificity for some clusters can explain their homogeneity in turn. But beyond all that, we can see different histological cancer cell lines are grouped into the same integrative clusters because they share gene alterations (**Supplementary Figures S5, S6**). Notably,

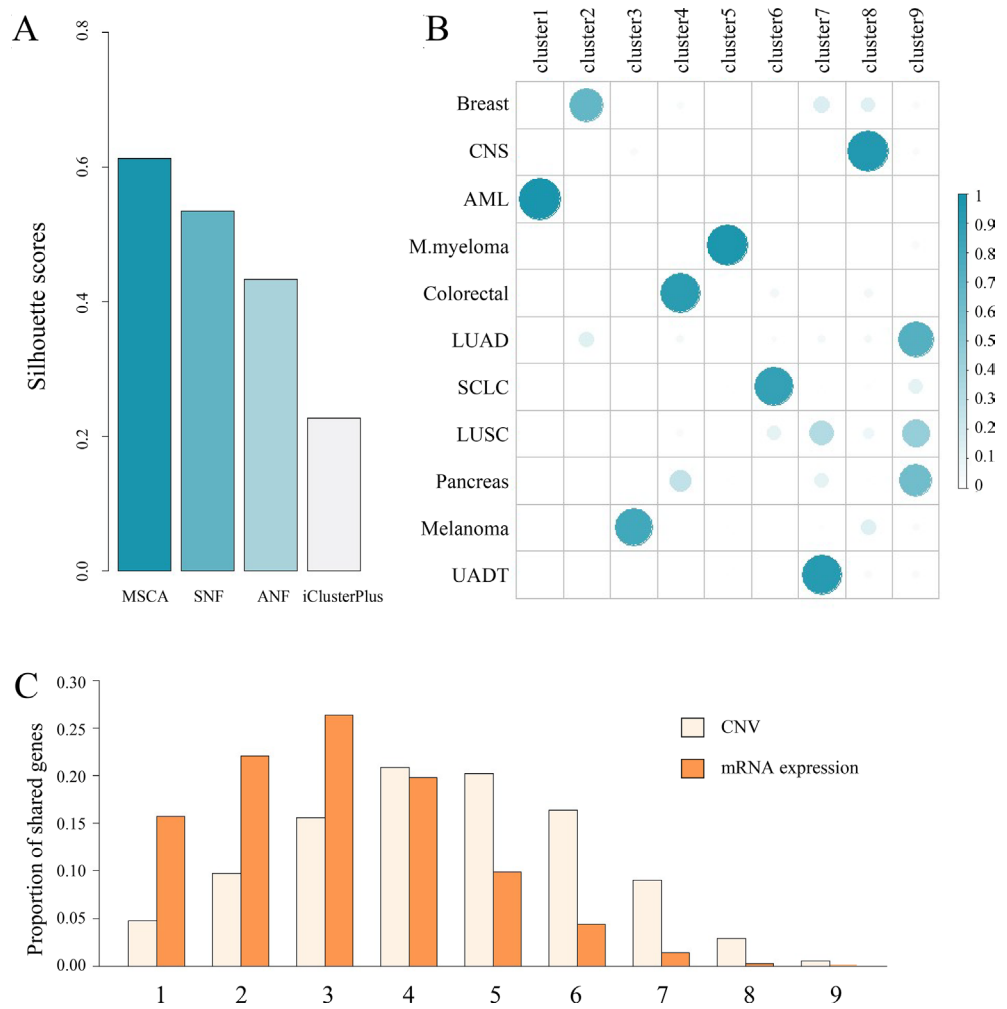

FIGURE 4 | A case study on CCLE dataset. **(A)** Comparison of the integrative clustering results obtained by MSCA, SNF, ANF and iClusterPlus. Note that the number of clusters is the same, i.e., 9 for the four methods. **(B)** Illustration of associations between histological origins (i.e., rows) and integrated cell clusters (i.e., columns). Sum of each row equals 1. **(C)** A brief summary of remarkable features across clusters. *x*-axis indicates the number of shared groups. Breast, breast cancer; CNS, central nervous systems; AML, acute myelocytic leukemia; M.myeloma, multiple myeloma; Colorectal, colorectal cancer; LUAD, lung adenocarcinoma; SCLC, small cell lung cancer; LUSC, lung squamous cell carcinoma; Pancreas, pancreas cancer; UADT, upper aerodigestive tract cancer.

the markable features between clusters, especially those copy number variants (**Figure 4C**), tend to be held by only few clusters, revealing strong heterogeneity between MSCA identified clusters (P-value < 10^{-12} and < 10^{-23} for expression and copy number data respectively, identified by sample shifting test for 5,000 times). Thus, the integrated pan-cancer analysis by MSCA may challenge the tissue original separation and indicate the common molecular aberrations across tumor types.

DISCUSSION

It's widely acceptable that integration of distinct types of biological data could provide more complete information to understand system complexity and disease heterogeneity (Ghazalpour et al., 2006; Kutalik et al., 2008; Li et al., 2012; Zhang et al., 2012; Zhang et al., 2017c). Over the past decades, the integration methods have progressed to get closer to biological details, from focusing on common information to specific signals, from critical hypothesis to assumption-free, and from linear models to nonlinear methods, etc. However, it is still a challenging task for bioinformatics to more accurately capture the underlying sample/gene structures from multiple omics data.

Here, we propose the MSCA model with the capacity to identify precise manifolds of samples in data space. In fact, our MSCA method is very similar to a previously published method, SNF (and ANF), which attempts to recognize sample patterns based on cross-view diffusion. However, the biggest difference is that SNF regards all the samples in the same feature space, nevertheless MSCA considers therein embedded multiple subspaces, i.e., different functional molecule sets. We carried out both synthetic examples and a real cancer dataset to demonstrate the capacities of MSCA. In the *in silico* studies, MSCA effectively fused the concordant

information associated in certain sample subgroups and outperformed several state-of-the-art integrative methods, in terms of clustering accuracy and robustness. In real case study, the sample patterns derived by MSCA correspond to biological differences using independent knowledge and analytic methods. Beyond that, we believe it can also help other studies which need integration of various data sources, in addition to complex diseases.

Though MSCA implements two nonlinear steps, proven to be effective in theory and practice, the problem of over-learning might still exist because we use the local similarities twice (see *Methods*). Such design may lead to bias when data types contain a lot of shared noises, which is worth careful consideration and improvement. Furthermore, MSCA has currently dealt with continuous data types (e.g., mRNA expression, copy number variant), the effectiveness on other forms of data, e.g., binary data (somatic mutation), category data (clinical covariates), still needs to be continuously improved.

AUTHOR CONTRIBUTIONS

CZ and QS completed the majority of the project and wrote the article. TZ and BH revised the article.

FUNDING

This paper was supported by the National Natural Science Foundation of China (No. 61802141), Natural Science Foundation of Hubei Province (No. 2018CFB098) and Huazhong Agricultural University Scientific & Technological Self-innovation Foundation (No. 2662017QD043).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00744/full#supplementary-material

REFERENCES

Arneson, D., Shu, L., Tsai, B., Barrere-Cain, R., Sun, C., and Yang, X. (2017). Multidimensional integrative genomics approaches to dissecting cardiovascular disease. *Front. Cardiovasc. Med.* 4 (Suppl 2), 8. doi: 10.3389/fcvm.2017.00008

Assié, G., Letouzé, E., Fassnacht, M., Jouinot, A., Luscap, W., Barreau, O., et al. (2014). Integrated genomic characterization of adrenocortical carcinoma. *Nat. Genet.* 46 (6), 607–612. doi: 10.1038/ng.2953

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., et al. (2012). The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 483 (7391), 603. doi: 10.1038/nature11003

Brbić, M., and Kopriva, I. (2017). Multi-view low-rank sparse subspace clustering. *Pattern Recognit.* (73), 247–258. doi: 10.1109/TCYB.2018.2883566

Cao, X., Zhang, C., Fu, H., Liu, S., and Zhang, H. (2015). "Diversity-induced Multi-view Subspace Clustering," in *Computer Vision & Pattern Recognition*. (Boston, USA: IEEE), 586–594. doi: 10.1109/CVPR.2015.7298657

Chen, J., and Zhang, S. (2016). Integrative analysis for identifying joint modular patterns of gene-expression and drug-response data. *Bioinformatics* 32 (11), 1724–1732. doi: 10.1093/bioinformatics/btw059

Ding, L., Gonzalez-Longatt, F. M., Wall, P., and Terzija, V. (2013). Two-step spectral clustering controlled islanding algorithm. *IEEE T. Power Syst.* 28 (1), 75–84. doi: 10.1109/TPWRS.2012.2197640

Fan, Y., He, R., and Hu, B. G. (2016). "Global and local consistent multi-view subspace clustering," in *Pattern Recognition.* (Kuala Lumpur, Malaysia: IEEE), 564–568. doi: 10.1109/ACPR.2015.7486566

Gao, H., Nie, F., Li, X., and Huang, H. (2016). "Multi-view Subspace Clustering," in *IEEE International Conference on Computer Vision*. (New York, USA: IEEE), 4238–4246. doi: 10.1109/ICCV.2015.482

Ghazalpour, A., Doss, S., Zhang, B., Wang, S., Plaisier, C., Castellanos, R., et al. (2006). Integrating genetic and network analysis to characterize genes related to mouse weight. *PLoS Genetics* 2 (8), e130. doi: 10.1371/journal.pgen.0020130

Haghverdi, L., Lun, A. T., Morgan, M. D., and Marioni, J. C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* 36 (5), 421. doi: 10.1038/nbt.4091

Kutalik, Z., Beckmann, J. S., and Bergmann, S. (2008). A modular approach for integrative analysis of large-scale gene-expression and drug-response data. *Nat. Biotechnol.* 26 (5), 531–539. doi: 10.1038/nbt1397

Li, W., Zhang, S., Liu, C. C., and Zhou, X. J. (2012). Identifying multi-layer gene regulatory modules from multi-dimensional genomic data. *Bioinformatics* 28 (19), 2458–2466. doi: 10.1093/bioinformatics/bts476

Lin, Z., Chen, M., and Ma, Y. (2010). The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *Eprint Arxiv* 9. doi: 10.1016/j.jsb.2012.10.010

Lin, Z., Liu, R., and Su, Z. (2011). Linearized alternating direction method with adaptive penalty for low-rank representation. *Adv. Neural Inf. Process. Syst.* 2011, 612–620.

Liu, G., Lin, Z., Yan, S., Sun, J., Yu, Y., and Ma, Y. (2013). Robust recovery of subspace structures by low-rank representation. *IEEE Trans. Pattern Anal. Mach. Intell.* 35 (1), 171–184. doi: 10.1109/TPAMI.2012.88

Ma, T., and Zhang, A. (2017). "Integrate multi-omic data using Affinity Network Fusion (ANF) for cancer patient clustering," in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM).* (Kansan City, MO, USA: IEEE), 398–403. doi: 10.1109/BIBM.2017.8217682

Mark, L. S., and Aviv, B. (2002). Waddington's canalization revisited: developmental stability and evolution. *Proc. Natl. Acad. Sci. U.S.A.* 99 (16), 10528–10532. doi: 10.1073/pnas.102303999

Meng, C., Helm, D., Frejno, M., and Kuster, B. (2015). moCluster: identifying joint patterns across multiple omics data sets. *J. Proteome Res.* 15(3), 755–765. doi: 10.1021/acs.jproteome.5b00824

Mo, Q., Wang, S., Seshan, V. E., Olshen, A. B., Schultz, N., Sander, C., et al. (2013). Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proc. Natl. Acad. Sci. U.S.A.* 110 (11), 4245–4250. doi: 10.1073/pnas.1208949110

Pan, J. B., Hu, S. C., Shi, D., Cai, M. C., Li, Y. B., Zou, Q., et al. (2013). PaGenBase: a pattern gene database for the global and dynamic understanding of gene function. *PLoS One* 8 (12), e80747. doi: 10.1371/journal.pone.0080747

Rousseeuw, P. J. (1999). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20 (20), 53–65. doi: 10.1016/0377-0427(87)90125-7

Safran., M., Dalah, I., Alexander, J., Rosen, N., Iny Stein, T., Shmoish, M., et al. (2010). GeneCards Version 3: the human gene integrator. *Database (Oxford),* 2010, baq020. doi: 10.1093/database/baq020

Santos, J. M., and Embrechts, M. (2009). "On the use of the adjusted rand index as a metric for evaluating supervised classification," in *Artificial Neural Networks-icann, International Conference*, 2009. (Limassol, Cyprus: Springer, Berlin, Heidelberg), 175–184. doi: 10.1007/978-3-642-04277-5_18

Schuster, S. (2008). Next-generation sequencing transforms today's biology. *Nat. Methods* 5 (1), 16. doi: 10.1038/nmeth1156

Shi, Q., Zhang, C., Peng, M., Yu, X., Zeng, T., Liu, J., et al. (2017a). Pattern fusion analysis by adaptive alignment of multiple heterogeneous omics data. *Bioinformatics* 33 (17), 2706–2714. doi: 10.1093/bioinformatics/btx176

Shi, Q., Zhang, C., Guo, W., Zeng, T., Lu, L., Jiang, Z., et al. (2017b). Local network component analysis for quantifying transcription factor activities. *Methods* 124, 25–35. doi: 10.1016/j.ymeth.2017.06.018

Uhlen, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., et al. (2015). Proteomics. Tissue-based map of the human proteome. *Science* 347 (6220), 1260419. doi: 10.1126/science.1260419

Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., et al. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* 11 (3), 333–337. doi: 10.1038/nmeth.2810

Xiong, Q., Ancona, N., Hauser, E. R., Mukherjee, S., and Furey, T. S. (2012). Integrating genetic and gene expression evidence into genome-wide association analysis of gene sets. *Genome Res.* 22 (2), 386–97. doi: 10.1101/gr.124370.111

Zhang, C., Fu, H., Hu, Q., Cao, X., Xie, Y., Tao, D., et al. (2018) Generalized latent multi-view subspace clustering. *IEEE Trans. Pattern Anal. Mach. Intell.* PP (99), 1–1. doi: 10.1109/TPAMI.2018.2877660

Zhang, C., Liu, J., Shi, Q., Yu, X., Zeng, T., and Chen, L. (2017a). "Integration of multiple heterogeneous omics data," in *IEEE International Conference on Bioinformatics and Biomedicine*. (Kansan City, MO, USA: IEEE), 564–569. doi: 10.1109/bibm.2016.7822582

Zhang, C., Hu, Q., Fu, H., Zhu, P., and Cao, X. (2017b). "Latent Multi-view Subspace Clustering," in *Computer Vision & Pattern Recognition*. (Honolulu, Hawaii: IEEE), 4279–4287. doi: 10.1109/CVPR.2017.461

Zhang, C., Liu, J., Shi, Q., Zeng, T., and Chen, L. (2017c). Differential function analysis: identifying structure and activation variations in dysregulated pathways. *Sci. China Inf. Sci.* 60 (1), 012108. doi: 10.1007/s11432-016-0030-6

Zhang, S. H., 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 (19), 9379–91. doi: 10.1093/nar/gks725

Zhuang, L., Wang, J., Lin, Z., Yang, A. Y., Ma, Y., and Yu, N. (2015). Locality-preserving low-rank representation for graph construction from nonlinear manifolds. *Neurocomputing* 175 (PA), 715–722. doi: 10.1016/j.neucom.2015.10.119

Conflict of Interest Statement: 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.