High-Order Correlation Integration for Single-Cell or Bulk RNA-seq Data Analysis

Hui Tang[1], *Tao Zeng*[1]* *and Luonan Chen*[1,2,3,4]*

[1] Key Laboratory of Systems Biology, CAS Center for Excellence in Molecular Cell Science, Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, University of Chinese Academy of Sciences, Shanghai, China, [2] CAS Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming, China, [3] School of Life Science and Technology, ShanghaiTech University, Shanghai, China, [4] Shanghai Research Center for Brain Science and Brain-Inspired Intelligence, Shanghai, China

OPEN ACCESS

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

Reviewed by:
Guoxian Yu,
Southwest University, China
Lihua Zhang,
University of California, Irvine, United States

***Correspondence:**
Tao Zeng
zengtao@sibs.ac.cn
Luonan Chen
lnchen@sibs.ac.cn

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

Received: *21 December 2018*
Accepted: *09 April 2019*
Published: *26 April 2019*

Citation:
Tang H, Zeng T and Chen L (2019) High-Order Correlation Integration for Single-Cell or Bulk RNA-seq Data Analysis. Front. Genet. 10:371. doi: 10.3389/fgene.2019.00371

Quantifying or labeling the sample type with high quality is a challenging task, which is a key step for understanding complex diseases. Reducing noise pollution to data and ensuring the extracted intrinsic patterns in concordance with the primary data structure are important in sample clustering and classification. Here we propose an effective data integration framework named as HCI (High-order Correlation Integration), which takes an advantage of high-order correlation matrix incorporated with pattern fusion analysis (PFA), to realize high-dimensional data feature extraction. On the one hand, the high-order Pearson's correlation coefficient can highlight the latent patterns underlying noisy input datasets and thus improve the accuracy and robustness of the algorithms currently available for sample clustering. On the other hand, the PFA can identify intrinsic sample patterns efficiently from different input matrices by optimally adjusting the signal effects. To validate the effectiveness of our new method, we firstly applied HCI on four single-cell RNA-seq datasets to distinguish the cell types, and we found that HCI is capable of identifying the prior-known cell types of single-cell samples from scRNA-seq data with higher accuracy and robustness than other methods under different conditions. Secondly, we also integrated heterogonous omics data from TCGA datasets and GEO datasets including bulk RNA-seq data, which outperformed the other methods at identifying distinct cancer subtypes. Within an additional case study, we also constructed the mRNA-miRNA regulatory network of colorectal cancer based on the feature weight estimated from HCI, where the differentially expressed mRNAs and miRNAs were significantly enriched in well-known functional sets of colorectal cancer, such as KEGG pathways and IPA disease annotations. All these results supported that HCI has extensive flexibility and applicability on sample clustering with different types and organizations of RNA-seq data.

Keywords: high–order, integration, clustering, single-cell, bulk data analysis

INTRODUCTION

Cells, the fundamental unit in biology, can be distinguished by their size and shape using a microscope. Later, advanced technological developments have made it possible to isolate a large number of cells, and along with improvements in RNA isolation and amplification methods, next-generation sequencing technologies are used to profile the transcriptome of individual cells. Single-cell RNA sequencing (scRNA-seq) now allows for omics analysis of individual cells, which can expose exciting biological processes, novel medical insights and efficient clinical applications (Dunham et al., 2012; Kolodziejczyk et al., 2015; Wagner et al., 2016). The advances in single-cell technologies have led to more comprehensive studies for multicellular organisms than previous approaches. Recently, 10X Genomics could release a single-cell dataset of more than 1.3 million cells (2017)[1]. With the production of large amount of single-cell data, understanding the development of an organic organ requires to characterize all of its cell types, so that, it is important to quantify single-cell cell types with high quality. Conventionally, one key application of scRNA-seq is to cluster cell types based on cells' transcriptome profiles through unsupervised computational methods (Lloyd, 1982; Jaitin et al., 2014; Mahata et al., 2014; Grün et al., 2015; Kiselev et al., 2017; Jiang et al., 2018; Shi et al., 2018; Dai et al., 2019). These approaches in recently published studies show some good performances in determining different cell types (Xue et al., 2013; Patel et al., 2014; Pollen et al., 2014; Shalek et al., 2014). SAFE-clustering (Yang Y. et al., 2018) can take as input results from multiple clustering methods and scmap (Kiselev et al., 2018) can compare clusters across data sets without merging. RaceID (Grün et al., 2015) augments k-means to identify rare cell types by detecting outliers, but k-means faces the problem of global solution. Meanwhile, SC3 (Kiselev et al., 2017) adopts repeated application of *k*-means using a small subset of principal components or different initial conditions and finding the consensus clusters. SC3 is a user-friendly clustering method that works well for smaller datasets. However, it takes too long in terms of computation time because of amount of calculating correlation matrix of cells. Besides, CIDR (Lin et al., 2017) adapts hierarchical clustering (HCA) for single-cell datasets by adding an implicit imputation of zeros into the distance calculation. But, an important shortcoming of hierarchical clustering is that it is prohibitively expensive for large datasets. Therefore, the more efficient and accurate method is still urgently needed to cluster cell types.

At the same time, large amounts of bulk data have already become widely available resources along with rapid development of high throughput technologies. To take full advantage of these rich data sets, integrating multiple datasets will give more opportunities to address biological dynamics and cancer heterogeneity (Hamid et al., 2009; Wang et al., 2014). Some integration methods have been developed in recent years, such as: iClusters, SNF, NMF, and PFA (Zhang et al., 2011; Mo et al., 2013; Mahata et al., 2014; Wang et al., 2014; Shi et al., 2017). However, there are still several limitations of these approaches. For example, iClusterPlus is based on Gaussian assumption, which could not make sense when data is too heterogeneous on signal distributions. And recently developed pattern fusion analysis (PFA) can integrate multidimensional data (Shi et al., 2017) so as to provide a comprehensive way to understand biological processes and complex diseases in a multi-view manner. In theory, PFA can align local sample-patterns derived from each single data type into a global sample-pattern to characterize the sample types in a low-dimensional feature space, so that, it is expected that PFA can model the sample types (i.e., cell types) when using scRNA-seq. However, the original PFA is designed for multi-source data rather than only one source data, in addition to insufficient analysis on the sample features. Thus, it is required to extend the original PFA to sample clustering even for one source data by a unified integration framework.

To overcome above challenges, we proposed a unified computational framework for distinguishing single-cell cell types from single-cell RNA-seq data, which also keeps the ability for clustering sample types from bulk RNA-seq data. The new method named as HCI (High-order Correlation Integration), can integrate joint high-order correlation matrices, where the iterative use of Pearson's correlation coefficient in sample data are incorporated into our previously developed pattern fusion analysis method (PFA) (Shi et al., 2017). Technically, HCI integrates single-cell data sets and different distance matrices corresponding to different sample correlation feature spaces (i.e., the distance between the cells) by joint matrix factorizations.

On the one hand, HCI has been compared with other existing methods [i.e., SC3 (Kiselev et al., 2017) and SEURAT (Macosko et al., 2015)] for identifying cell types on various single-cell RNA-seq data. And the robustness of HCI was also tested in different correlation orders (e.g., one-order, second-order, different percentage of differentially expressed genes). Furthermore, a case study was conducted by HCI on a scRNA-seq dataset of Diabetes, which successfully clustered the ambiguous cells unassigned in previous study. On the other hand, HCI was also applied to analyze bulk RNA-seq data as previous PFA, e.g., bulk RNA-seq and other omics data (Schuster, 2008). By comparing HCI with the original PFA on three datasets with multiple data types (e.g., gene expression and miRNA expression), it is found that HCI can improve computational efficiency of sample clustering and can recognize gene regulatory networks in an accurate and reliable manner (Joung et al., 2007; Tran et al., 2008; Hamid et al., 2009; Peng et al., 2009).

Totally, HCI can not only cluster cell types with scRNA-seq data in an efficient way, but also capture biologically meaningful sample types as well as extracting network modules with bulk RNA-seq data or other omics data. It provides a new and general way to detect the sample-specific characteristics from the high-order correlation information in an integration manner.

MATERIALS AND METHODS

HCI pipeline schematically is shown in **Figure 1**. Input is the expression matrix **M** where columns correspond to cells or samples and rows correspond to genes or molecules, e.g., each element of **X** corresponds to the expression of a gene in a given

[1] 10X Genomics single cell gene expression datasets from https://support.10xgenomics.com/single-cell-gene-expression/datasets

cell. The analysis procedure of HCI can be summarized as several steps in follows.

Pre-processing

The gene filtering removes genes with zero expressions in all cells (or samples), which are not informative for the cell clustering. And, the normalization for each column data is carried to maintain the feature stability of each cell or sample. Then, we can get a filtered expression matrix X.

High-Order Correlation Matrix Construction

We firstly calculate $\mathbf{F}^1$, the correlation of the gene expression profiles $\mathbf{X}_{m \cdot n}$, in which the expressions of m genes are measured for n samples and x_{kj} denotes the expression level of gene k in sample j , the correlation of sample i and j can be calculated by the Pearson correlation coefficient (Rodgers and Nicewander, 1988):

$$f_{ij}^{(1)} = \frac{\sum_{k=1}^{n}(x_{ki} - x_{-i})(x_{kj} - x_{-j})}{\sqrt{\sum_{k=1}^{n}(x_{ki} - x_{-i})^2}\sqrt{\sum_{k=1}^{n}(x_{kj} - x_{-j})^2}} \quad (1)$$

where x_{ki} and x_{-i} are the expression level of gene k and the average gene expression level of sample i, respectively. Similarly, x_{kj} and x_{-j} are the expression level of gene k and the average gene expression level of sample j, respectively. Thus, we can obtain a correlation matrix $\mathbf{F}_{n \cdot n}^1$ of $\mathbf{X}$ in which $f_{i \cdot j}^1$ is its element measuring the correlation coefficient between sample i and sample j. Now, based on the matrix $\mathbf{F}_{n \cdot n}^1$, we can further calculate $\mathbf{F}_{n \cdot n}^2$ as follows:

$$f_{ij}^{(2)} = \frac{\sum_{k=1}^{n}\left(f_{ki}^{(1)} - f_{-i}^{(1)}\right)\left(f_{kj}^{(1)} - f_{-j}^{(1)}\right)}{\sqrt{\sum_{k=1}^{n}\left(f_{ki}^{(1)} - f_{-i}^{(1)}\right)^2}\sqrt{\sum_{k=1}^{n}\left(f_{kj}^{(1)} - f_{-j}^{(1)}\right)^2}} \quad (2)$$

$\mathbf{F}_{n \cdot n}^1$ is called as the first-order correlation matrix of $\mathbf{X}$, and $\mathbf{F}_{n \cdot n}^2$ is the second-order correlation matrix of $\mathbf{X}$. The advantage of this transformation with expression matrix $\mathbf{X}$ can highlight latent structures between samples with noisy (Hubert, 1985; Ren et al., 2013). In fact, we also investigated the other kind of distance matrix by using other method, such as Spearman correlation, however, $\mathbf{F}_{n \cdot n}^2$ is similar to $\mathbf{F}_{n \cdot n}^1$ due to its consideration on element rank rather than element value in matrices. Cleary, the higher-order correlation matrix can be constructed in a similar way. Therefore, in this paper, we only use the Pearson metrics to construct our high-order correlation matrices. Noted, such high-order matrix can enhance the sample clustering performance. In our prior analysis, the clustering accuracy increased quickly on the first-order correlation features, and it almost approached the highest on the second-order correlation features and tended to be saturated when the order further increased. Without loss of generality, we only used the first-order matrix and the second-order matrix to incorporate into HCI in this work.

Correlation Matrix Induced Pattern Fusion Analysis (PFA)

The input data $\mathbf{X}$ has m rows and n columns, and matrices $\mathbf{F}_{n \cdot n}^1$ and $\mathbf{F}_{n \cdot n}^2$ have n rows and n columns. We integrated these three input datasets by pattern fusion analysis. This methodology has been proved and evaluated in previous work (Shi et al., 2017), and the key steps used in our work are as follows:

The first step is to obtain the optimal local information sets of $\mathbf{U}^i, \mathbf{Y}^i$, which requires to minimize the error $\mathbf{E}^i$ as follows:

$$min \parallel E^i \parallel = min_{c^i, U^i, Y^i} \parallel W^i - (c^i 1^T + U^i Y^i) \parallel_F^2 \quad (3)$$

where $\mathbf{W}^i$ is the input data sets $\mathbf{X}$, $\mathbf{F}_{n \cdot n}^1$, $\mathbf{F}_{n \cdot n}^2$, and F is the Frobenius norm. Then, we have

$$\begin{cases} U^i = Q_{d^i}^i \\ Y^i = \left(U^i\right)^T \left(W^i - c^i 1^T\right) \\ c^i = \frac{W^i 1}{n} \end{cases} \quad (4)$$

where $\mathbf{Q}_{d^i}^i$ is an orthogonal matrix formed by the eigenvectors corresponding to the first d^i largest eigenvalues of $(\mathbf{W}^i - \mathbf{c}^i 1^T)(\mathbf{W}^i - \mathbf{c}^i 1^T)^T$. It is important noted that the sensible default values d^i of matrix X is chosen according to $\sum_{r=1}^{d^i} \delta_r / \sum_{r=1}^{p} \delta_r \geq 0.8$ and d^i is the r largest eigenvalues of $(\mathbf{W}^i - \mathbf{c}^i 1^T)(\mathbf{W}^i - \mathbf{c}^i 1^T)^T$ and the number of the non-zeros eigenvalues is p . Meanwhile, the d^i-dimension of matrix $\mathbf{F}_{n \cdot n}^1$ and $\mathbf{F}_{n \cdot n}^2$ is chosen according to $\sum_{r=1}^{d^i} \delta_r / \sum_{r=1}^{p} \delta_r \geq 0.9$ due to their different feature dimensions with $\mathbf{X}$.

And then, the adaptive optimal alignment is used to capture the global sample-pattern matrix $\mathbf{Y}$. The detailed adaption method can be seen in the original study (Shi et al., 2017), and the related parameters can be easily adjusted by the user.

Sample Clustering and Cluster Number Estimation

The global sample-spectrum $\mathbf{Y}$ obtained in the above step instead of conventional data matrix $\mathbf{X}$ can be clustered by many clustering methods, such as K-means or HCA. In this paper, K-means clustering (Ding and He, 2004) is performed on the global sample-spectrum matrix Y by using the "kmeans()" MATLAB function.

The ratio of distance between clusters (RDC) is calculated to estimate the number K of clusters. One hundred realizations of the sample clustering used K-means clustering. The number K of clusters is inferred by the average RDC number [$K = \min$ (K, the average RDC's slope is nearly 0)]. The RDC can be calculated as:

$$RDC = \frac{D_{in}}{D_{out}} \quad (5)$$

where D_{in} is the average sample distance in clusters; D_{out} is the average sample distance between clusters.

Since the reference labels of cells or samples are already known for all published datasets, the Adjusted Rand Index (ARI) (Hubert, 1985) is applied to calculate the similarity between the HCI clustering results and prior-known clusters, which can be further used to evaluate HCI and other methods [e.g., SC3 (Kiselev et al., 2017), PFA, one-order, second-order, and CV situations].

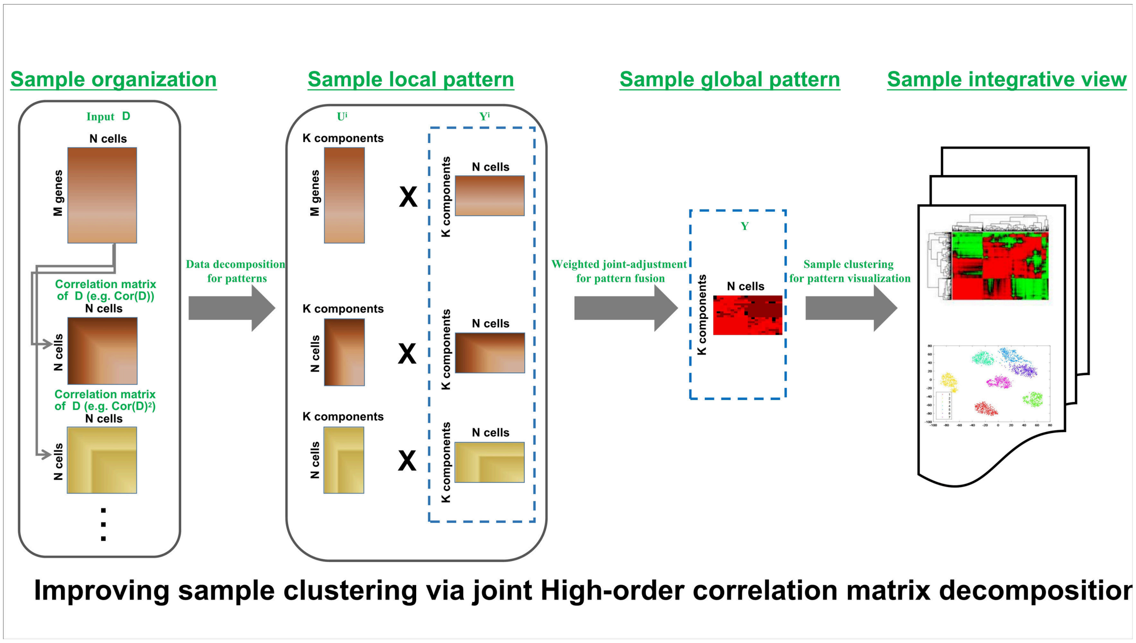

Sample organization
Sample local pattern
Sample global pattern
Sample integrative view
Input D
N cells
M genes
Correlation matrix of D (e.g. Cor(D))
N cells
N cells
Correlation matrix of D (e.g. Cor(D)²)
N cells
N cells
Data decomposition for patterns
Uⁱ
Yⁱ
K components
M genes
X
K components
N cells
K components
N cells
X
K components
N cells
K components
N cells
X
K components
N cells
Weighted joint-adjustment for pattern fusion
Y
K components
N cells
Sample clustering for pattern visualization
Improving sample clustering via joint High-order correlation matrix decompositior

FIGURE 1 | HCI framework of high-order correction matrices based on pattern fusion analysis Overview of single-cell clustering with HCI framework (see section Materials and Methods).

Molecular Network Construction for Case Study on Bulk RNA-seq Data

The multi-level network is integratively constructed by using HCI schematically shown in **Figure 4A**. In the same way, we calculated the high-order matrices $\mathbf{F}^1_{n\cdot n}$ and $\mathbf{F}^2_{n\cdot n}$ of the input datasets $\mathbf{X}_I$ (e.g., RNAseq, Methylation, MicroRNA), where n is number of samples in data. And then we integrated all input datasets $\mathbf{X}_I$ and high-order correlation matrices $\mathbf{F}^1_I$, $\mathbf{F}^2_I$ by using pattern fusion analysis method. Based on the global sample-spectrum matrix $\mathbf{Y}$, we can get the differentially expressed mRNAs (or miRNAs) from heterogeneous genomic datasets according to the coefficient matrix $\mathbf{U}^{I^*}$. In this work, we calculated a coefficient of variation for each element on the rows of $\mathbf{U}^{I}*$:

$$c_i = \frac{\delta_i}{\mu_i} \tag{6}$$

where μ_i is the average weight of mRNA i (or miRNA i) in U^{I^*}, and δ_i is the standard deviation. We can define differentially expressed mRNA (or miRNA) i if c_i is greater than a given threshold T, and they called DEGs (or DE-miRNAs).

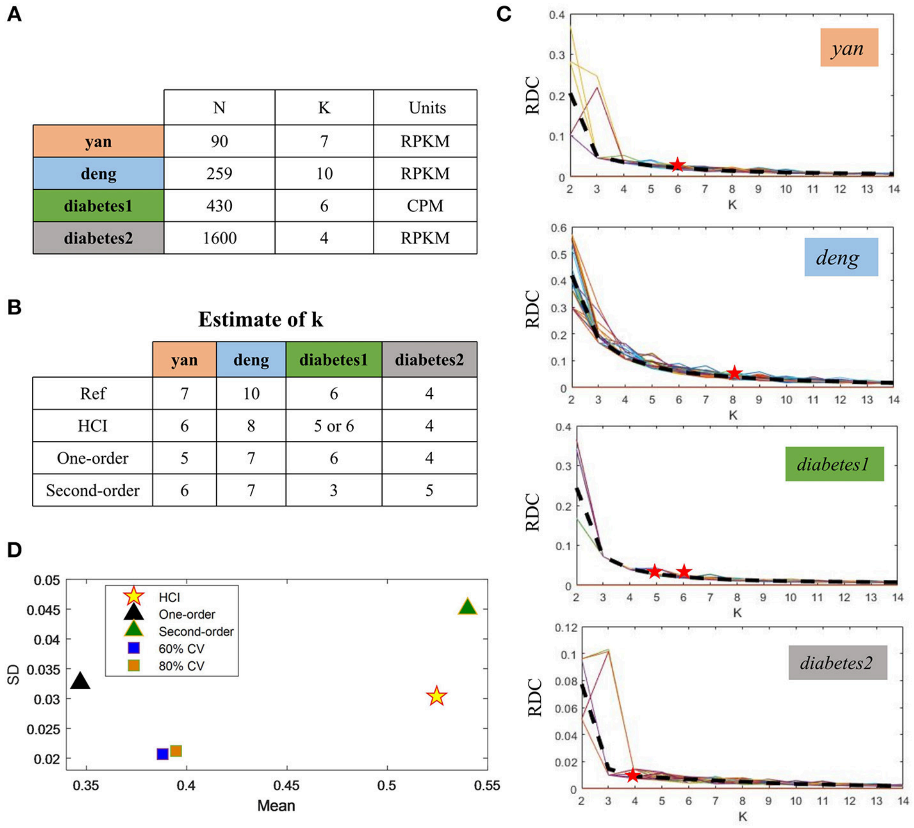
Besides, we also performed functional enrichment analysis for genes by Gene Ontology and KEGG. We also analyzed DEGs using Ingenuity Pathway Analysis (IPA), providing the association between a particular gene set and known functions, pathways, networks and associated diseases. An online database miRDB was used for miRNA target prediction and functional annotations.

We defined key genes that significantly enriched in cancer dependent on KEGG, GO and IPA analysis. We found the key genes in the DEGs, which can be linked and correlated by the combined functional couplings of protein-protein interactions of STRING. MicroRNAs which can regulate key DEGs were defined as key miRNAs (degree s > 80) (Hu et al., 2018). Cytoscape was used to reconstruct and visualize gene-gene and miRNA-gene network.

RESULTS

Performance Comparison and Robustness Evaluation

To demonstrate the performance of HCI on the single-cell datasets, we firstly downloaded four publicly available scRNA-Seq datasets (**Figure 2A**) (Yan et al., 2013; Deng et al., 2014; Wang et al., 2016; Xin et al., 2016). These datasets were selected on the basis that one can be highly confident on the cell labels as representative cells from different stages, conditions and lines. In order to quantify the similarity between the reference cell types and the clusters obtained by HCI or other comparable methods. We calculated the average ARI of the clustering results (**Figure 2D**, **Figure S1**) and estimated cluster number K according to RDC by running K-means 100 times (**Figure 2C**). Obviously, high-order correlation matrices incorporated into PFA actually improves both the accuracy and the stability of analysis solutions. We found that the accuracy was significantly improved compared with the one-order correlation matrix (only using $\mathbf{F}^1_I$) or the second-order matrix (only using $\mathbf{F}^2_I$) according to the ARI and the RDC (**Figures 2B,D**). Besides, in order to determine the robustness as a consistent performance under different conditions, the same analysis on four datasets were both



A

	N	K	Units
yan	90	7	RPKM
deng	259	10	RPKM
diabetes1	430	6	CPM
diabetes2	1600	4	RPKM

B

Estimate of k

	yan	deng	diabetes1	diabetes2
Ref	7	10	6	4
HCI	6	8	5 or 6	4
One-order	5	7	6	4
Second-order	6	7	3	5

FIGURE 2 | Accuracy and robustness evaluation **(A)** A brief introduction of four published datasets used in HCI. *N* is the number of cells in a dataset; *K* is the number of clusters originally identified; Units: FPKM is Fragments Per Kilobase of transcript per Million mapped reads, CPM is Counts of per Million mapped reads. **(B)** Number of clusters *K* predicted by HCI, One-order situation, Second-order situation, 80% CV used and 60% CV used for all datasets. Ref is the cell cluster reported in previous studies and used as reference of comparison among HCI, One-order situation, and Second-order situation. **(C)** RDC was applied 100 times in global sample-pattern matrix *Y* to each dataset. The solid lines correspond to the value of each RDC calculation. The dashed black lines correspond to the average of these solid lines. Y-coordinate in each graph represents the RDC value and the x-coordinate represents the number of cluster *K*. The star indicates *K* which we choose (see methods). **(D)** The mean and standard deviation of ARI in four datasets by running k-means 50 times separately in different situations.

repeated 50 times under different systematic conditions (e.g., 60% CV genes or 80% CV genes used) respectively, where CV genes mean ones with largest expression variances. Similarly, the performance of HCI under different correlation matrices or conditions was better (i.e., robust) than other methods according to the ARI and the RDC (**Figures 2B,D**, **Figure S1**). Overall, HCI always outperformed compared methods on distinguishing single-cell types.

Comparison of Sample-Cluster Identification With One-Level Data

We applied HCI and SC3 method to the above four datasets for evaluation and comparison on the cell clustering. We calculated the cluster number K and the running time in each individual dataset by using the R package of SC3 (Kiselev et al., 2017). On the one hand, as shown in **Table 1**, HCI performs better than SC3 across almost all datasets in estimating the number K of clusters (except for similar performance on Deng dataset). On the other hand, the running time of 2,000 cells for SC3 is more than 1 h. By contrast, the running time of HCI for 2,000 cells is <10 min as shown in **Table 2**. It is worth noted that HCI can even apply to large datasets, such as: 10k datasets from 10x genomics, with more than 10,000 cells by using MATLAB efficiently (**Table 2**, **Figure S2**). From these results, we included that, HCI has better performance than SC3 because it considers the high-order correlation information, and integrates this potential heterogeneous information by our PFA framework well.

TABLE 1 | The estimation of K compared with SC3 on real datasets.

	yan	deng	diabetes1	diabetes2
Ref	7	10	6	4
cPFA	6	7	5 or 6	4
SC3	6	9	11	13

Case Study on the scRNA-seq Data of Diabetes

We then applied HCI to a diabetes scRNA-seq data (Wang et al., 2016) with 430 annotated cells belonging to six cell types,

TABLE 2 | The running time compared with SC3 on real datasets.

	yan	deng	diabetes1	diabetes2	brain	10x
N cells	90	259	430	1,600	3,003	10,000
cPFA	5 s	18 s	31.85 s	5.8 min	16.4 min	3.1 h
SC3	7.33 min	18 min	29.67 min	101.18 min	98 min	4 h (no result)

where 205 ambiguous cells previously unassigned. For the 430 annotated cells, the RDC of HCI suggested that K is 5 or 6 (**Figures 2B**,**D**), provides the reasonable cluster number of cells. When we applied HCI to the whole cells included 430 annotated cells and 205 dropped cells, the results suggested that the K is 7. Obviously, there are potential new cell types included, and we found there were 27 annotated mesenchymal cells in the ambiguous cells. This result also showed that the other ambiguous cells can be clustered well in seven cell types separately (**Figure 3A**). Besides, the other methods (e.g., tSNE, HCA) were used to visualize the clusters of these dropped cells (**Figure 3**). As a control to this analysis, one well-known scRNA-seq analysis method SEURAT (Macosko et al., 2015) was also applied. As the results shown (**Figure S3**), HCI performed better than these traditional methods on distinguishing cell types. Noted, cluster dendrogram of global sample-pattern matrices **Y**, $\mathbf{F}^1$, and $\mathbf{F}^2$ are shown in **Figure S4** for illustrating the influence of HCI on information integration.

In addition, marker genes are particularly useful since they can usually uniquely indicate a cell cluster, e.g., α-cells with high expression on *IRX2* and *ARX*. To further interpret the biological meaning of HCI based cell clustering, we applied the 50 key marker genes of the annotated cell types to categorize the previously dropped cells which had been clustered well by HCI now. The violin plot shown the expression level of *IRX2* and *ARX* are significantly high in alpha cells previously identified and also in alpha-dropped cells newly clustered by HCI (**Figure S5**). Furthermore, it was observed a high degree of expression similarity between annotated cells and their corresponding clustered-dropped cells in these key markers (**Figures S6, S7**). Together with these results, we concluded that HCI is able to identify new cell types with high accuracy and biological significance.

Comparison of Sample-Cluster Identification With Multi-Level Data

To demonstrate the effectiveness of HCI inherited from PFA for integrating multi-level datasets, we applied HCI to three cancer omics datasets, two from the TCGA Data Portal included kidney renal clear cell carcinoma (KIRC) and Adrenocortical carcinoma (ACC), and one from the GEO (Colorectal cancer) (Sayagués et al., 2016). For the two TCGA data, the gene expression, miRNA expression and DNA methylation profiles were prepared in a similar way as those in Shi et al. (2017). As for the Colon cancer, the gene expression and miRNA expression were obtained, and we removed those mRNAs or miRNAs if they have more than 80% zero expression values across all samples. Then these datasets with 122 patients in KIRC, 79 in ACC and 51 in colon cancer were prepared, respectively (**Figure 4B**).

After carried on HCI and PFA on these datasets, respectively, we compared their results according to the RDC, which show that that HCI indeed performs better in terms of accuracy of cluster quality across datasets (**Figures 4C–E**). In this comparison, the heterogeneity factors including different complex conditions, varying data resources and dissimilar samples size would provide strong evidences to support the ability of HCI on identifying clinically relevant disease subtypes and predicting network modules involved in complex diseases (Zhang et al., 2011; Zang et al., 2016).

Case Study on the Matched mRNA and miRNA Data of Colorectal Cancer

Finally, we carried on a case study again on colorectal cancer data, especially providing the integrated mRNA-miRNA network according to the global sample-spectrum matrix **Y**. Firstly, the HCI results suggested that the normal (9 samples) and disease (42 samples) can be clustered into two discriminative groups (**Figure 5A**). Then, 6,930 differentially expressed genes and 2,976 differentially expressed miRNAs were obtained. By functional enrichment analysis on these differentially expressed genes with GO BP terms, KEGG pathways and IPA annotations, all significant physiological system development, function terms, disease and networks are listed in **Tables S1, S2**. We found that there are 2,289 genes (nearly 33% DEGs) are significantly correlated with colon cancer among all DEGs. Besides, according to the miRNA target predication from miRDB, 1,661 DEGs can be regulated by 141 DE-miRNAs (**Figure 5B**). Note that all enrichment analysis results involve 25 key genes, 14 of which can be regulated by 22 key miRNAs (**Figure 5B**). In addition, the survival risks of these genes were also evaluated as shown in **Figure 5C**.

As an illustrative instance, we constructed the gene-gene network of 25 key genes (**Figure 5D**) based on the STRING (p = 1.0e-16) (2018)[2]. The enrichment analysis results of this network are listed in **Figure 5E** (**Table S3**), and this network is significantly enriched with cytosol (P = 3.86e-05), beta-catenin destruction complex (P = 1.57e-04), colorectal cancer (P = 2.73e-46), and pathways in cancer (P = 6.82e-41). We also found that the hub genes (e.g., *MAPK8, EGF, FALGDS, CCND1, MYC*) in this network have been linked to cancer in wide literature reports. For example, the MAPK-signaling pathways have been identified as one of the most strongly associated gene markers to colorectal cancer (CRC) (Cummins et al., 2006; Barault et al., 2008; Lascorz et al., 2010; Slattery et al., 2012). *MAPK8* has been shown to interact with *MYC* which is frequently observed in numerous human cancers. Strikingly, 22 key miRNAs are correlated with 14 key genes in this network. *MiRNA-647* and *miRNA-449a* have been reported their association with colorectal cancer (Noguchi et al., 1999; Feng et al., 2018). These results revealed HCI would classify the sample types clearly and could integrate the multi-level regulatory network based on multiple heterogeneous data. All relevant DEGs and DE-miRNAs

[2] https://string-db.org/

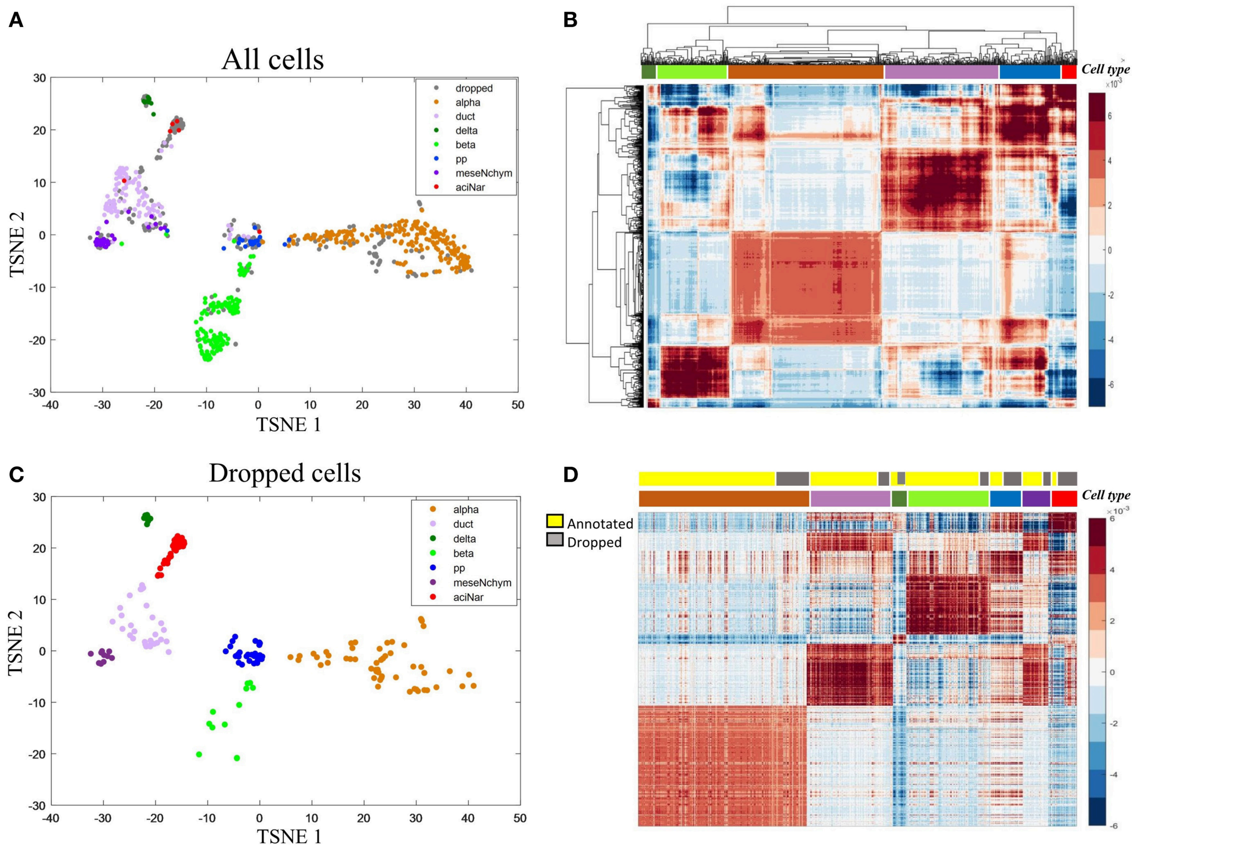
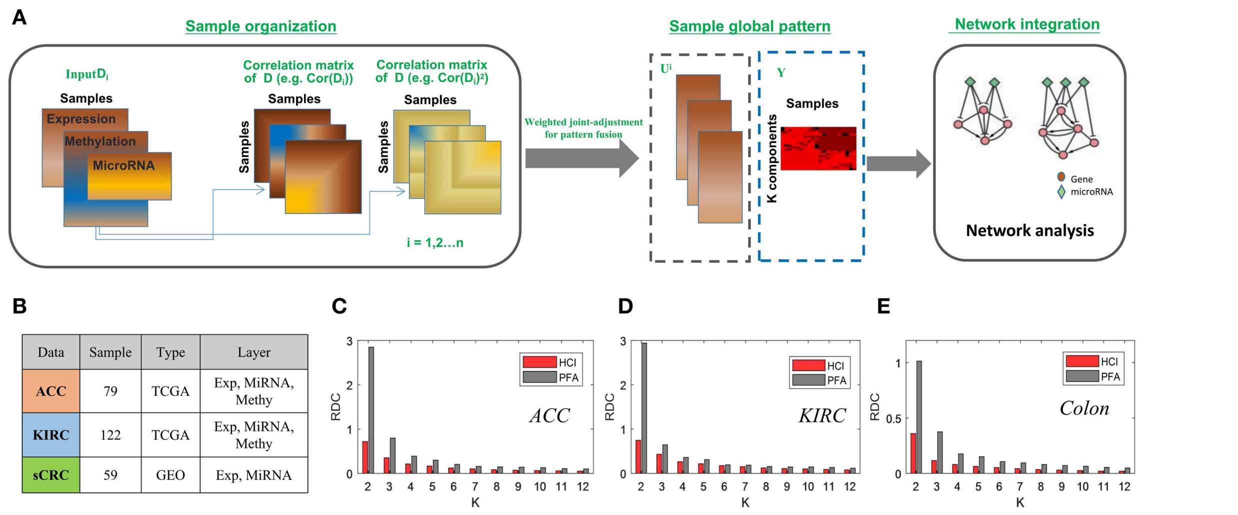


FIGURE 3 | Case study on the diabetes **(A)** The two-dimensional projection of the global-sample pattern matrix Y of diabetes1 dataset using t-SNE. Colors represent different cell types where gray dots specifically mean dropped cells. **(B)** Hierarchical clustering diagram of all cells. The colors at the top represent the references labels (see color legend in **A**). The purple represents the ductal cell and mesenchymal cell because these two cells are mixed together. **(C)** For the previously dropped cells, HCI can cluster them into different groups corresponding to known cell types well. Thus, it is more efficient on cell type identification with less non-identified cells. **(D)** Heatmap representation of Y of all cells. The first color bar represents the type of annotated cell and ambiguous cells and the second bar represents the seven cell types as the same as in **(A)** legend.

are worthy of future experimental investigation, and listed in **Tables S4, S5**.

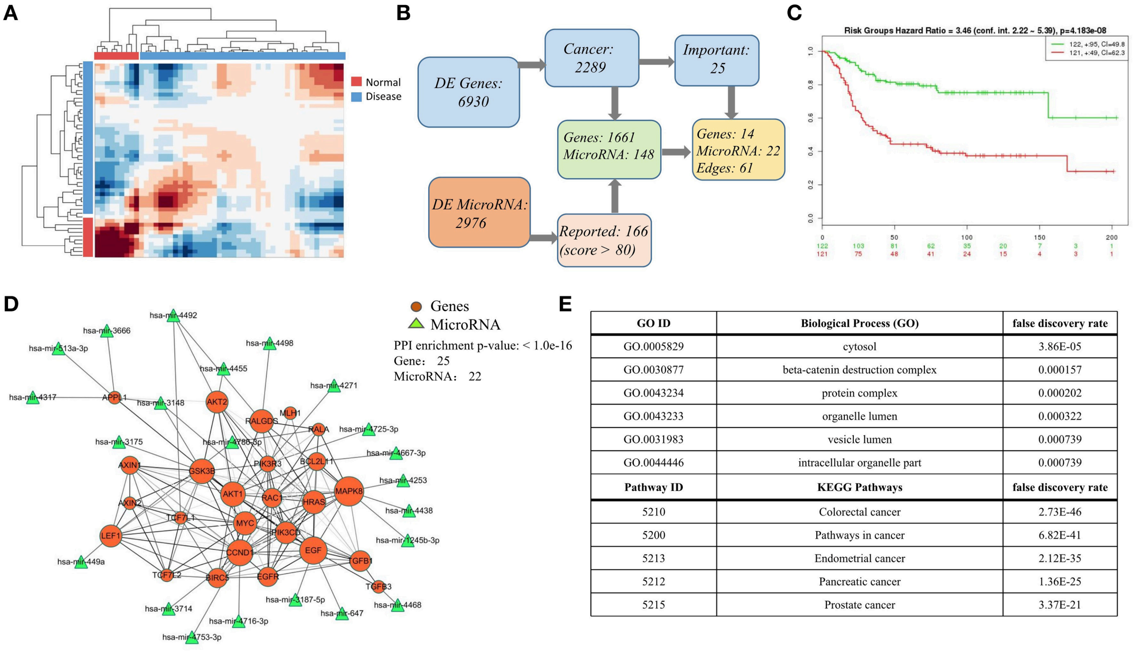
DISCUSSION AND CONCLUSION

The distinct types of biological data could provide a precise explanation for understanding the complex biological processes (Ghazalpour et al., 2006; Kutalik et al., 2008; Li et al., 2012; Zhang et al., 2012; Chen and Zhang, 2016; Zeng et al., 2016; Feng et al., 2018; Yu and Zeng, 2018). In recent decades, many approaches were proposed for analyzing single-cell data or multi-omics data to identify subtypes and construct biological networks (Gygi et al., 1999; Ding and He, 2004; Chari et al., 2010; Zhang et al., 2011; Kiselev et al., 2017; Guo et al., 2018a,b; Wang et al., 2018). However, for most methods, there are some limitations on reliably identifying the sample types by exploiting multi-datasets, such as the effect of noise on data and the computational cost. And some methods would fail to make full use of the similarity information between samples, thus making the results unreliable. Hence, in order to overcome this problem, a flexible and efficient integration method with automated information fusion and bias correction is demanded. In this work, we introduced the data-driven integrating method HCI. The key idea of this method is to incorporate the high-order similarity matrices (e.g., Pearson correlation matrix) into pattern fusion analysis, where the sample cluster or subtype structure can be actually determined benefiting from the high-order correlations. And the obtained combinatorial sample patterns from HCI could represent comprehensive characterization of inherent sample relations in data. In order to demonstrate the benefits of HCI, various evaluations have been carried on both scRNA-seq and bulk RNA-seq datasets for complex diseases. As expected, HCI effectively captured the sample (e.g., cell or patient) clusters and outperformed the existing methods under different conditions in terms of accuracy and robustness. And two deep case studies supported that HCI has satisfactory flexibility and applicability. Noted, HCI is based on PFA, which has been evaluated and compared with a few multi-view clustering methods in previous study (Shi et al., 2017). Meantime, SC3 has also been evaluated and compared with



Data	Sample	Type	Layer
ACC	79	TCGA	Exp, MiRNA, Methy
KIRC	122	TCGA	Exp, MiRNA, Methy
sCRC	59	GEO	Exp, MiRNA

FIGURE 4 | The enhanced framework flow for integrating bulk datasets and comparison of PFA **(A)** The flow chart of HCI to integrating multiple heterogeneous omics data **(B)** A brief introduction to the datasets we used in this comparison. **(C–E)** Bars correspond to the average of the RDC values by running 100 times in each dataset. Red and gray colors correspond to result of HCI and PFA respectively.



GO ID	Biological Process (GO)	false discovery rate
GO.0005829	cytosol	3.86E-05
GO.0030877	beta-catenin destruction complex	0.000157
GO.0043234	protein complex	0.000202
GO.0043233	organelle lumen	0.000322
GO.0031983	vesicle lumen	0.000739
GO.0044446	intracellular organelle part	0.000739
Pathway ID	**KEGG Pathways**	**false discovery rate**
5210	Colorectal cancer	2.73E-46
5200	Pathways in cancer	6.82E-41
5213	Endometrial cancer	2.12E-35
5212	Pancreatic cancer	1.36E-25
5215	Prostate cancer	3.37E-21

FIGURE 5 | Case study on the colorect cancer **(A)** Hierarchical clustering diagram of samples in matrix *Y*. Color bars represent the normal samples and disease samples. **(B)** The process diagram of selecting key genes. **(C)** Evaluation of the selected 25 important genes related to colorectal cancer in **(B)**. In SurvExpress, we used the average for selected genes, two risk groups and Cox fitting to generate Survival curves. The total number of each group is shown in the top right corner of graph, and the number of censoring samples is marked with +. The CI per curve is also included. *P*-value is shown in the top of figure. **(D)** The highly connected network consists mainly of 25 DEG genes and 22 miRNAs. The 22 miRNAs targets 16 genes based on miRDB database. The size of node indicates the network degree of gene. And the PPI enrichment *P*-value of genes is shown in the top right corner of this figure. **(E)** Top-ranked pathways and biological functions enriched in the 25 genes in **(D)**.

many existing approaches (Kiselev et al., 2017). Thus, in this study of scRNA data, we have directly compared HCI and SC3 on multiple datasets. It is worthy to carry on more benchmark studies in this field as a future topic (Zeng et al., 2016). Also as a future topic, we can improve HCI by further exploiting dynamics and network information, such as applying network biomarker (Zhang et al., 2015; Liu et al., 2016; Zhao et al., 2016; Liu, X. et al., 2018) or applying dynamic network biomarker

(Chen et al., 2012; Li et al., 2017; Liu et al., 2017; Liu, R. et al., 2018; Yang B. et al., 2018) for accurate and reliable clustering and classification based on omics data from the perspectives of dynamics and network.

As genomic data sources is increasing in diversity and volume, HCI can fit the data structures on both one level data or multiple level data, so that, HCI could provide new avenues for the systematic explanation of various data and complex biological phenotypes at a system-wide level. Indeed, there are still a few future topics to further extend HCI method, e.g., integrating discrete data types including somatic, SNP, and CNV information.

AUTHOR CONTRIBUTIONS

HT and TZ developed the methodology. HT executed the experiment. HT and TZ carried out the data analysis and wrote this paper. HT, TZ, and LC revised the manuscript. LC and TZ supervised the work, and LC critically reviewed the paper. All authors read and approved the final manuscript.

FUNDING

This work was supported by the National key research and development program of China (No. 2017YFA0505500), the Strategic Priority Research Program of the Chinese Academy of Sciences (CAS) (No. XDB13040700), the National Natural Science Foundation of China (NSFC) (Nos. 61403363, 11401222, 11871456, 31771476), the Shanghai Municipal Science and Technology Major Project (Grant No. 2017SHZDZX01), and the Natural Science Foundation of Shanghai (17ZR1446100).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00371/full#supplementary-material

REFERENCES

Barault, L., Veyrie, N., Jooste, V., Lecorre, D., Chapusot, C., Ferraz, J. M., et al. (2008). Mutations in the RAS-MAPK, PI(3)K (phosphatidylinositol-3-OH kinase) signaling network correlate with poor survival in a population-based series of colon cancers. *Int. J. Cancer* 122, 2255–2259. doi: 10.1002/ijc.23388

Chari, R., Thu, K. L., Wilson, I. M., Lockwood, W. W., Lonergan, K. M., Coe, B. P., et al. (2010). Integrating the multiple dimensions of genomic and epigenomic landscapes of cancer. *Cancer Metastasis Rev.* 29, 73–93. doi: 10.1007/s10555-010-9199-2

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

Chen, L., Liu, R., Liu, Z. P., Li, M., and Aihara, K. (2012). Detecting early-warning signals for sudden deterioration of complex diseases by dynamical network biomarkers. *Sci. Rep.* 2:342. doi: 10.1038/srep00342

Cummins, J. M., He, Y., Leary, R. J., Pagliarini, R., Diaz, L. A., Sjoblom, T., et al. (2006). The colorectal microRNAome. *Proc. Natl. Acad. Sci. U.S.A* 103, 3687–3692. doi: 10.1073/pnas.0511155103

Dai, H., Li, L., Zeng, T., and Chen, L. (2019). Cell-specific network constructed by single-cell RNA sequencing data. *Nucl. Acids Res.* doi: 10.1093/nar/gkz172. [Epub ahead of print].

Deng, Q., Ramsköld, D., Reinius, B., and Sandberg, R. (2014). Single-cell rna-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* 343, 193–196. doi: 10.1126/science.1245316

Ding, C., and He, X. F. (2004). Cluster structure of K-means clustering via principal component analysis. *Adv. Knowledge Discov. Data Mining Proc.* 3056, 414–418. doi: 10.1007/978-3-540-24775-3_50

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74. doi: 10.1038/nature11247

Feng, Y., Dong, Y. W., Song, Y. N., Xiao, J. H., Guo, X. Y., Jiang, W. L., et al. (2018). MicroRNA-449a is a potential predictor of colitis-associated colorectal cancer progression. *Oncol. Rep.* 40, 1684–1694. doi: 10.3892/or.2018.6566

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 Genet.* 2, 1182–1192. doi: 10.1371/journal.pgen.0020130

Grün, D., Lyubimova, A., Kester, L., Wiebrands, K., Basak, O., Sasaki, N., et al. (2015). Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* 525, 251–255. doi: 10.1038/nature14966

Guo, W. F., Zhang, S. W., Liu, L. L., Liu, F., Shi, Q. Q., Zhang, L., et al. (2018a). Discovering personalized driver mutation profiles of single samples in cancer by network control strategy. *Bioinformatics* 34, 1893–1903. doi: 10.1093/bioinformatics/bty006

Guo, W. F., Zhang, S. W., Shi, Q. Q., Zhang, C. M., Zeng, T., and Chen, L. (2018b). A novel algorithm for finding optimal driver nodes to target control complex networks and its applications for drug targets identification. *BMC Genomics* 19 (Suppl. 1):924. doi: 10.1186/s12864-017-4332-z

Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H., and Aebersold, R. (1999). Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* 17, 994–999.

Hamid, J. S., Hu, P., Roslin, N. M., Ling, V., Greenwood, C. M., and Beyene, J. (2009). Data integration in genetics and genomics: methods and challenges. *Hum Genomics Proteo.* 2009:869093. doi: 10.4061/2009/869093

Hu, J., Zeng, T., Xia, Q., Qian, Q., Yang, C., Ding, Y., et al. (2018). Unravelling miRNA regulation in yield of rice (*Oryza sativa*) based on differential network model. *Sci. Rep.* 8:8498. doi: 10.1038/s41598-018-26438-w

Hubert L, A. P. (1985). Comparing partitions. *J. Classification Classif.* 14, 193–218.

Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., et al. (2014). Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* 343, 776–779. doi: 10.1126/science.1247651

Jiang, H., Sohn, L. L., Huang, H., and Chen, L. (2018). Single cell clustering based on cell-pair differentiability correlation and variance analysis. *Bioinformatics* 34, 3684–3694. doi: 10.1093/bioinformatics/bty390

Joung, J. G., Hwang, K. B., Nam, J. W., Kim, S. J., and Zhang, B. T. (2007). Discovery of microRNA-mRNA modules via population-based probabilistic learning. *Bioinformatics* 23, 1141–1147. doi: 10.1093/bioinformatics/btm045

Kiselev, V. Y., Kirschner, K., Schaub, M. T., Andrews, T., Yiu, A., Chandra, T., et al. (2017). SC3: consensus clustering of single-cell RNA-seq data. *Nat. Methods* 14, 483–486. doi: 10.1038/nmeth.4236

Kiselev, V. Y., Yiu, A., and Hemberg, M. (2018). scmap: projection of single-cell RNA-seq data across data sets. *Nat. Methods* 15, 359–362. doi: 10.1038/nmeth.4644

Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C., and Teichmann, S. A. (2015). The technology and biology of single-cell RNA sequencing. *Mol. Cell* 58, 610–620. doi: 10.1016/j.molcel.2015.04.005

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. Biotechn.* 26, 531–539. doi: 10.1038/nbt1397

Lascorz, J., Försti, A., Chen, B., Buch, S., Steinke, V., Rahner, N., et al. (2010). Genome-wide association study for colorectal cancer identifies risk polymorphisms in German familial cases and implicates MAPK signalling pathways in disease susceptibility. *Carcinogenesis* 31, 1612–1619. doi: 10.1093/carcin/bgq146

Li, M., Li, C., Liu, W. X., Liu, C., Cui, J., Li, Q., et al. (2017). Dysfunction of PLA2G6 and CYP2C44-associated network signals imminent carcinogenesis from chronic inflammation to hepatocellular carcinoma. *J. Mol. Cell Biol.* 9, 489–503. doi: 10.1093/jmcb/mjx021

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, 2458–2466. doi: 10.1093/bioinformatics/bts476

Lin, P., Troup, M., and Ho, J. W. (2017). CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data. *Genome Biol.* 18:59. doi: 10.1186/s13059-017-1188-0

Liu, R., Wang, J., Ukai, M., Sewon, K., Chen, P., Suzuki, Y., et al. (2018). Hunt for the tipping point during endocrine resistance process in breast cancer by dynamic network biomarkers. *J. Mol. Cell. Biol.* doi: 10.1093/jmcb/mjy059. [Epub ahead of print].

Liu, X., Chang, X., Leng, S., Tang, H., Aihara, K., Chen, L. (2018). Detection for disease tipping points by landscape dynamic network biomarkers. *Nat. Sci. Rev.* doi: 10.1093/nsr/nwy162. [Epub ahead of print].

Liu, X., Chang, X., Liu, R., Yu, X., Chen, L., and Aihara, K. (2017). Quantifying critical states of complex diseases using single-sample dynamic network biomarkers. *PLoS Comput. Biol.* 13:05633. doi: 10.1371/journal.pcbi.1005633

Liu, X., Wang, Y., Ji, H., Aihara, K., and Chen, L. (2016). Personalized characterization of diseases using sample-specific networks. *Nucl. Acids Res.* 44:e16422. doi: 10.1093/nar/gkw772

Lloyd, S. P. (1982). Least-squares quantization in Pcm. *IEEE Trans. Inform. Theory* 28, 129–137.

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* 161, 1202–1214. doi: 10.1016/j.cell.2015.05.002

Mahata, B., Zhang, X., Kolodziejczyk, A. A., Proserpio, V., Haim-Vilmovsky, L., Taylor, A. E., et al. (2014). Single-cell RNA sequencing reveals T helper cells synthesizing steroids *de novo* to contribute to immune homeostasis. *Cell Rep.* 7, 1130–1142. doi: 10.1016/j.celrep.2014.04.011

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, 4245–4250. doi: 10.1073/pnas.1208949110

Noguchi, K., Kitanaka, C., Yamana, H., Kokubu, A., Mochizuki, T., and Kuchino, Y. (1999). Regulation of c-Myc through phosphorylation at Ser-62 and Ser-71 by c-Jun N-terminal kinase. *J. Biol. Chem.* 274, 32580–32587. doi: 10.1074/jbc.274.46.32580

Patel, A. P., Tirosh, I., Trombetta, J. J., Shalek, A. K., Gillespie, S. M., Wakimoto, H., et al. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* 344, 1396–1401. doi: 10.1126/science.1254257

Peng, X., Li, Y., Walters, K. A., Rosenzweig, E. R., Lederer, S. L., Aicher, L. D., et al. (2009). Computational identification of hepatitis C virus associated microRNA-mRNA regulatory modules in human livers. *BMC Genomics* 10:373. doi: 10.1186/1471-2164-10-373

Pollen, A. A., Nowakowski, T. J., Shuga, J., Wang, X., Leyrat, A. A., Lui, J. H., et al. (2014). Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.* 32, 1053–1058. doi: 10.1038/nbt.2967

Ren, X., Wang, Y., Zhang, X. S., and Jin, Q. (2013). iPcc: a novel feature extraction method for accurate disease class discovery and prediction. *Nucl. Acids Res.* 41(14):gkt343. doi: 10.1093/nar/gkt343

Rodgers, J. L., and Nicewander, W. A. (1988). 13 ways to look at the correlation-coefficient. *Am. Stat.* 42, 59–66.

Sayagués, J. M., Corchete, L. A., Gutiérrez, M. L., Sarasquete, M. E., Del Mar Abad, M., Bengoechea, O., et al. (2016). Genomic characterization of liver metastases from colorectal cancer patients. *Oncotarget* 7, 72908–72922. doi: 10.18632/oncotarget.12140

Schuster, S. C. (2008). Next-generation sequencing transforms today's biology. *Nat. Methods*, 5, 16–18. doi: 10.1038/nmeth1156

Shalek, A. K., Satija, R., Shuga, J., Trombetta, J. J., Gennert, D., Lu, D., et al. (2014). Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* 510, 363–369. doi: 10.1038/nature13437

Shi, J., Teschendorff, A. E., Chen, W., Chen, L., and Li, T. (2018). Quantifying Waddington's epigenetic landscape: a comparison of single-cell potency measures. *Brief Bioinform*. doi: 10.1093/bib/bby093. [Epub ahead of print].

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

Slattery, M. L., Lundgreen, A., and Wolff, R. K. (2012). MAP kinase genes and colon and rectal cancer. *Carcinogenesis* 33, 2398–2408. doi: 10.1093/carcin/bgs305

Tran, D. H., Satou, K., and Ho, T. B. (2008). Finding microRNA regulatory modules in human genome using rule induction. *BMC Bioinformatics* 9 (Suppl. 12): S5. doi: 10.1186/1471-2105-9-S12-S5

Wagner, A., Regev, A., and Yosef, N. (2016). Revealing the vectors of cellular identity with single-cell genomics. *Nat. Biotechnol.* 34, 1145–1160. doi: 10.1038/nbt.3711

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, 333–337. doi: 10.1038/nmeth.2810

Wang, L., Yu, X., Zhang, C., and Zeng, T. (2018) Detecting personalized determinants during drug treatment from omics big data. *Curr. Pharm. Des.* 24, 3727–3738. doi: 10.2174/1381612824666181106102111

Wang, Y. J., Schug, J., Won, K. J., Liu, C., Naji, A., Avrahami, D., et al. (2016). Single-cell transcriptomics of the human endocrine pancreas. *Diabetes* 65, 3028–3038. doi: 10.2337/db16-0405

Xin, Y., Kim, J., Okamoto, H., Ni, M., Wei, Y., Adler, C., et al. (2016). RNA sequencing of single human islet cells reveals type 2 diabetes genes. *Cell Metabo.* 24, 608–615. doi: 10.1016/j.cmet.2016.08.018

Xue, Z., Huang, K., Cai, C., Cai, L., Jiang, C. Y., Feng, Y., et al. (2013). Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature* 500, 593–597. doi: 10.1038/nature12364

Yan, L., Yang, M., Guo, H., Yang, L., Wu, J., Li, R., et al. (2013). Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nat. Struct. Mol. Biol.* 20, 1131–1139. doi: 10.1038/nsmb.2660

Yang, B., Li, M., Tang, W., Liu, W., Zhang, S., Chen, L., et al. (2018). Dynamic network biomarker indicates pulmonary metastasis at the tipping point of hepatocellular carcinoma. *Nat. Commun.* 9:678. doi: 10.1038/s41467-018-03024-2

Yang, Y., Huh, R., Culpepper, H. W., Lin, Y., Love, M. I., and Li, Y. (2018). SAFE-clustering: Single-cell Aggregated (From Ensemble) clustering for single-cell RNA-seq data. *Bioinformatics.* 8:793. doi: 10.1093/bioinformatics/bty793

Yu, X. T., and Zeng, T. (2018). Integrative analysis of omics big data. *Methods Mol. Biol.* 1754, 109–135. doi: 10.1007/978-1-4939-7717-8_7

Zang, C., Wang, T., Deng, K., Li, B., Hu, S., Qin, Q., et al. (2016). High-dimensional genomic data bias correction and data integration using MANCIE. *Nat. Commun.* 7:11350. doi: 10.1038/ncomms11305

Zeng, T., Zhang, W., Yu, X., Liu, X., Li, M., and Chen, L. (2016). Big-data-based edge biomarkers: study on dynamical drug sensitivity and resistance in individuals. *Brief Bioinform* 17, 576–592. doi: 10.1093/bib/bbv078

Zhang, S., Li, Q., Liu, J., and Zhou, X. J. (2011). A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA-gene regulatory modules. *Bioinformatics* 27, I401–I409. doi: 10.1093/bioinformatics/btr206

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. *Nucl. Acids Res.* 40, 9379–9391. doi: 10.1093/nar/gks725

Zhang, W., Zeng, T., Liu, X., and Chen, L. (2015). Diagnosing phenotypes of single-sample individuals by edge biomarkers. *J. Mol. Cell Biol.* 7, 231–241. doi: 10.1093/jmcb/mjv025

Zhao, J., Zhou, Y., Zhang, X., and Chen, L. (2016). Part mutual information for quantifying direct associations in networks. *Proc. Natl. Acad. Sci. U.S.A* 113, 5130–5135. doi: 10.1073/pnas.1522586113

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.