Bioinformatics, 2017, 1–9
doi: 10.1093/bioinformatics/btx176
Advance Access Publication Date: 12 April 2017
Original Paper

OXFORD

Systems biology

Pattern fusion analysis by adaptive alignment of multiple heterogeneous omics data

Qianqian Shi[1], Chuanchao Zhang[1,2], Minrui Peng[1], Xiangtian Yu[1], Tao Zeng[1,*], Juan Liu[2,*] and Luonan Chen[1,*]

[1]Key Laboratory of Systems Biology, CAS Center for Excellence in Molecular Cell Science, Innovation Center for Cell Signaling Network, Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences; University of Chinese Academy of Sciences, Shanghai 200031, China and [2]State Key Laboratory of Software Engineering, School of Computer, Wuhan University, Wuhan 430072, China

*To whom correspondence should be addressed.
Associate Editor: Jonathan Wren

Received on October 15, 2016; revised on March 18, 2017; editorial decision on March 23, 2017; accepted on March 27, 2017

Abstract

Motivation: Integrating different omics profiles is a challenging task, which provides a comprehensive way to understand complex diseases in a multi-view manner. One key for such an integration is to extract intrinsic patterns in concordance with data structures, so as to discover consistent information across various data types even with noise pollution. Thus, we proposed a novel framework called 'pattern fusion analysis' (PFA), which performs automated information alignment and bias correction, to fuse local sample-patterns (e.g. from each data type) into a global sample-pattern corresponding to phenotypes (e.g. across most data types). In particular, PFA can identify significant sample-patterns from different omics profiles by optimally adjusting the effects of each data type to the patterns, thereby alleviating the problems to process different platforms and different reliability levels of heterogeneous data.

Results: To validate the effectiveness of our method, we first tested PFA on various synthetic datasets, and found that PFA can not only capture the intrinsic sample clustering structures from the multi-omics data in contrast to the state-of-the-art methods, such as iClusterPlus, SNF and moCluster, but also provide an automatic weight-scheme to measure the corresponding contributions by data types or even samples. In addition, the computational results show that PFA can reveal shared and complementary sample-patterns across data types with distinct signal-to-noise ratios in Cancer Cell Line Encyclopedia (CCLE) datasets, and outperforms over other works at identifying clinically distinct cancer subtypes in The Cancer Genome Atlas (TCGA) datasets.

Availability and implementation: PFA has been implemented as a Matlab package, which is available at http://www.sysbio.ac.cn/cb/chenlab/images/PFApackage_0.1.rar.

Contact: lnchen@sibs.ac.cn, liujuan@whu.edu.cn or zengtao@sibs.ac.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Large amounts of multiple heterogeneous omics data have become widely available with rapid advance of high throughput technologies (Schuster, 2008). Conventional studies based only on single data types (e.g. gene expression) could account for very limited number of the biological systems or disease complexity (Chari *et al.*, 2010; Gygi *et al.*, 1999; Hamid *et al.*, 2009). Nowadays, it has been recognized that the distinct types of biological data provide partly dependent views for the whole landscape of complex diseases, and thus integrating multiple data types gives more opportunities to

address many biological or medical issues, e.g. coherent genomic alterations (Chen and Zhang, 2016; Kutalik *et al.*, 2008; Zhang *et al.*, 2011, 2012), and cancer heterogeneity (Hamid *et al.*, 2009; Wang *et al.*, 2014). Therefore, the integrative analysis of various omics data is increasingly becoming an important component of the genomics and bioinformatics (Gevaert, 2008).

Plenty of integration approaches aim to infer the correlated patterns in multiple-dimensional genomic profiles, named as 'co-modules' or 'molecule-patterns' (Chen and Zhang, 2016; Ghazalpour *et al.*, 2006; Li *et al.*, 2012; Zhang *et al.*, 2012), shedding light on coherent understanding of biological systems. However, for many cases, very limited molecular features share similar phenotypes across data types, thus resulting in unreliable identification of prognostic distinct subtypes. Besides, diverse data sources could not provide identical information to uncover 'sample-patterns' (e.g. cancer heterogeneity) due to systematic noises or scale differences. Thus, the confident signals supported by a handful of data types (i.e. complementary sample-patterns) were often missing, and the shared noises could be strengthened. To address these issues, several integration approaches were developed in recent years. For example, iClusterPlus (Mo *et al.*, 2013), an enhanced method derived from iCluster (Shen *et al.*, 2010), considers different variable types using a couple of generalized linear models. However, iClusterPlus is based on Gaussian assumption, which could not discover the patterns when data is too heterogeneous on signal distributions. Another popular method called as similarity network fusion (SNF), was proposed by Wang et al. (Wang *et al.*, 2014), to construct sample networks from each data type, and then fuse them into one weighted sample-similarity network. In their work, SNF treated all available data types equally, which may lead to diluting signals and make a false 'fusion' due to different platforms or reliability levels of data types. Later, Meng et al. presented moCluster (Meng *et al.*, 2015), which is based on PCA. But, simple PCA could not deal with low signal-to-noise data, resulting in the unstable performance of moCluster. Hence, a data-driven method, which not only can exploit the data structures but also has the ability of the automated information fusion and bias correction, was still lacking.

To overcome the challenges above, we proposed Pattern Fusion Analysis (PFA), a novel computational framework that can identify the integrated 'sample-patterns' across heterogeneous genomic profiles, based on an adaptive optimization strategy. In particular, PFA can align local sample-patterns derived from each data type into a global sample-pattern to characterize phenotypes. In such a way, PFA is capable to capture biologically meaningful sample-pattern in a low-dimensional feature space as well as to quantitatively evaluate how informative each data type or even a sample is to support this global sample-pattern. To validate the effectiveness of PFA, we have tested PFA on several synthetic datasets, and found that PFA with great robustness, could accurately obtain the intrinsic sample-wide patterns compared to the established integration methods (such as iClusterPlus, SNF and moCluster). Furthermore, we have applied PFA to CCLE (The Cancer Cell Line Encyclopedia) (Barretina *et al.*, 2012) datasets and three multi-level cancer datasets in TCGA (The Cancer Genome Atlas) (Cancer Genome Atlas Research, 2013). PFA successfully identified distinct subgroups of cells or patients with biological or clinical importance, providing novel insights into cancer heterogeneity.

2 Materials and methods

2.1 Method overview

PFA can be summarized with the following two steps, as schematically shown in Figure 1: (i) Extracting local information including local 'sample-patterns' from each data type by dimension reduction method (Fig. 1B and C); (ii) Capturing the global sample-spectrum or sample-pattern by an adaptive optimization strategy across multiple data types (Fig. 1D). Here, a sample-pattern is represented by sample-to-sample associations or similarities, and Figure 1B and D was redrawn by modifying the similar figure from that in (Zhang *et al.*, 2012).

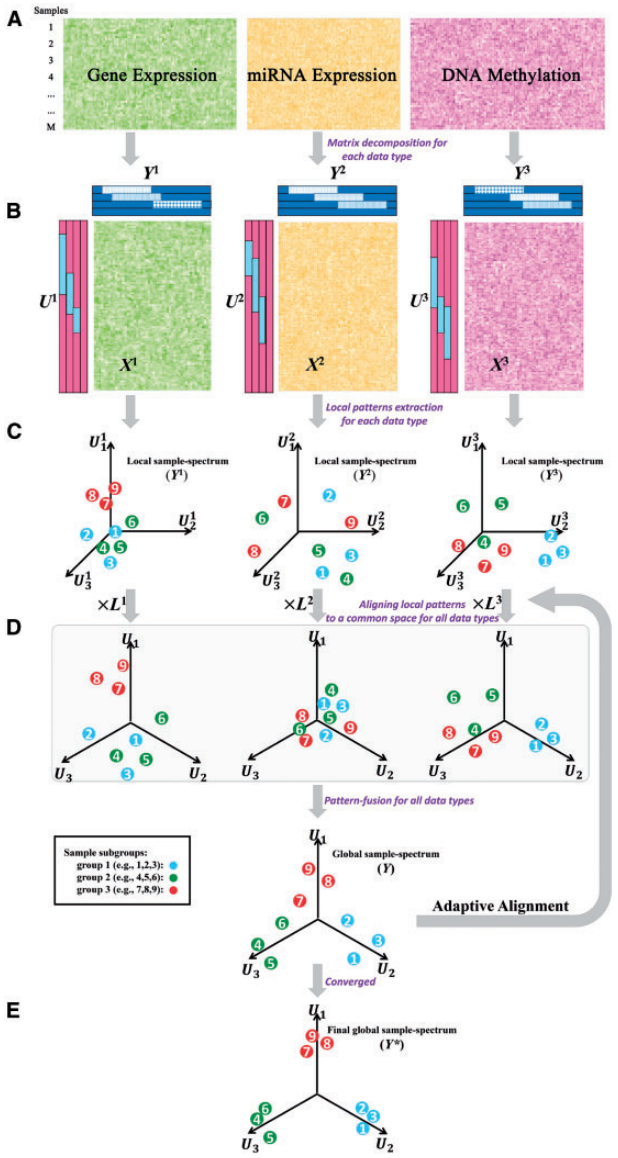

Fig. 1. Overview of Pattern Fusion Analysis (PFA). (**A**) Multiple biological data types for the same set of samples as the input of PFA. The rows of data matrices correspond to samples and the columns correspond to biological elements (e.g. genes, miRNAs and DNA methylation sites). (**B**) Local information extraction by dimension reduction approach for each data type. The local sample-spectrum (Y^i, $1 \le i \le 3$; in dark blue color) of each data type in low-dimensional feature space (U^i, $1 \le i \le 3$; in carmine color) is derived from the corresponding biological data (X^i, $1 \le i \le 3$) independently by principal component analysis. (**C**) Representation of local sample-patterns for each data type in relative low-dimensional feature spaces. (**D**) Illustration of iterative pattern-fusion process. The local sample-spectrum (Y^i, $1 \le i \le 3$) is projected onto a global space by 'alignment matrix' (L^i, $1 \le i \le 3$) and fused in an adaptive way by 'correction matrix' (W^i, $1 \le i \le 3$), i.e. iteratively update the global sample-spectrum (Y). (**E**) The final integrated global sample-spectrum (Y^*) represents the inherent sample-pattern. Each node indicates a subject and node color marks the subsets of samples

Firstly, given k genomic profiles for the same cohort of subjects (e.g. patients or samples), PFA derives genomic features and sample-patterns for every type of data $X^i (i \in \{1,\ldots,k,\ k \geq 2)$ independently through principal component analysis (PCA: $X^i \sim U^i Y^i$) (Fig. 1A and B). This step is to make dimensionality reduction, i.e. reduce redundancy, by an orthonormal matrix U^i, and gain those subsets of features which can represent the local information of each data type (Fig. 1C).

In the next step, the key essence of PFA is to recognize optimal global sample-spectrum Y by fusing these obtained sample-patterns (e.g. common or complementary patterns) across multiple local sample-spectrums $Y^1,\ldots,Y^k$ (Fig. 1C and D). This pattern-fusion step iteratively updates two matrices, i.e. the 'alignment matrix' $L^i (i \in \{1,\ldots,k\})$ and the 'correction matrix' $W^i (i \in \{1,\ldots,k\})$ for all data types, mainly based on the Local Tangent Space Alignment (LTSA) theory (Zhang and Zha, 2003). 'Correction matrices' measure the contributions by each data type (or even one sample) to the global sample-pattern, and their values are initialized equally, thereafter the values would become larger if the i_{th} data type (or sample) could provide more supportive information for sample-pattern identification and vice versa. In each iteration, an appropriate alignment matrix L^i projects different local Y^i onto a common low-dimensional space for sample alignment and pattern fusion, thus delineating a global spectrum Y to characterize the population heterogeneity. Along with the iterations, the updated matrix W^i can be used to re-adjust various biases (or noises) of data and re-evaluate the effects by any Y^i to Y, until the convergence is reached with an optimal global sample-spectrum Y^* (Fig. 1E).

2.2 Extracting local information by dimension reduction for each data type

The input data matrix X^i has h^i rows (i.e. biological elements) and n columns (i.e. samples). Each column vector x^i_j measures sample j in the h^i-dimensional feature space. To reduce redundancy, we firstly extract the local information for every sample j in a space of smaller dimensionality d^i corresponding to a given data type i by the following formula:

$$x^i_j = c^i + U^i y^i_j + \varepsilon^i_j \ (1 \leq j \leq n, 1 \leq i \leq k) \tag{1}$$

$$\text{or } X^i = c^i \mathbf{1}^T + U^i Y^i + E^i \tag{2}$$

where U^i is an orthonormal $h^i \times d^i$ matrix representing a d^i-dimensional feature subspace, which mostly describes the variability of samples, and Y^i is a $d^i \times n$ matrix indicating the local sample-spectrum in such d^i-dimensional space. Besides, c^i is a constant vector (column vector) with h^i elements, and $\mathbf{1}$ is a vector with n elements, i.e. $\mathbf{1}=(1,\ldots,1)^T$, where T is the transpose operator. E^i is the error matrix.

To obtain the optimal local information sets of U^i and Y^i, it is required to minimize the reconstruction error E^i, i.e.

$$\min\|E^i\| = \min_{c^i,U^i,T^i} \|X^i - \left(c^i \mathbf{1}^T + U^i Y^i\right)\|^2_F \tag{3}$$

where F is the Frobenius norm.

This problem can be easily solved based on Theorem 1 in Supplementary Information and the optimal solution is given as follows:

$$\begin{cases} U^i = Q^i_{d^i} \\ Y^i = \left(U^i\right)^T \left(X^i - c^i \mathbf{1}^T\right) \\ c^i = \dfrac{X^i \mathbf{1}}{n} \end{cases} \tag{4}$$

where $Q^i_{d^i}$ is an orthogonal matrix formed by the eigenvectors corresponding to the first d^i largest eigenvalues of $(X^i - c^i 1^T)(X^i - c^i 1^T)^T$.

From this theoretical results, the following Algorithm 1 can extract the local sample-spectrum Y^i of each data type, where the d^i-dimension is chosen according to $\sum_{r=1}^{d^i} \delta_r / \sum_{r=1}^{p} \delta_r \geq 0.8$ and δ_r is the r_{th} largest eigenvalues of $\left(X^i - c^i 1^T\right)\left(X^i - c^i 1^T\right)^T$ and the number of the non-zero eigenvalues is p. Note that, these local patterns are in different feature spaces, depending on their data types, and thus should not be directly compared with each other at this time.

Algorithm 1 Algorithm to extract the local sample-spectrum of each biological data type

Input: the profile of i_{th} data type, i.e. $X^i = [x^i_1, x^i_2, \ldots, x^i_n]$ and principal component number d^i

Output: the local sample-spectrum based on i_{th} data type: $Y^i = [y^i_1, y^i_2, \ldots, y^i_n]$

1. Data profile centralization: $\bar{X}^i = X^i(I - 1/\mathbf{1}\mathbf{1}^T)$
2. Doing the eigenvalue decomposition of the matrix $\bar{X}^i(\bar{X}^i)^T$, *i.e.* $\bar{X}^i(\bar{X}^i)^T = U^i \Lambda^i (U^i)^T$, U^i is an orthogonal matrix; and Λ^i is a diagonal matrix
3. Computing the local sample-spectrum Y^i, *i.e.* $Y^i = (U^i_{d^i})^T \bar{X}^i$, $U^i_{d^i}$ is an orthogonal matrix, deriving from the d^i largest eigenvalues of U^i

2.3 Capturing global sample-spectrum across data types by adaptive optimal alignment

Next, we align all local patterns above (in different feature spaces from different data types) to a common feature space so as to combine the local patterns across most data types. As motivated by the fact that each layer of biological data would interpret sample-wide pattern with different viewpoints, we further propose an adaptive integration approach to capture signals and remove bias in the local sample-spectrum fusion process. Briefly, the adaption method can be derived by optimizing the fitness function in the following:

$$\min \sum_{i=1}^{k} \frac{1}{V^i} \sum_{j=1}^{n} w^i_j \left\| y_j - b - L^i y^i_j \right\|^2 = \min \sum_{i=1}^{k} \frac{1}{V^i} \left\| Y - b\mathbf{1}^T - \left(L^i Y^i\right) W^i \right\|^2_F$$
$$\geq \min \sum_{i=1}^{k} \frac{1}{V^i} \left\| Y \left(I - \frac{\mathbf{1}\mathbf{1}^T}{n} \right) \left(I - \left(Y^i W^i\right)^{\dagger} \left(Y^i W^i\right) \right) \right\|^2_F = \min tr\left(Y \Phi Y^T\right) \tag{5}$$

$$\Phi = \sum_{i=1}^{k} \frac{1}{V^i} \Phi_i = \sum_{i=1}^{k} \frac{1}{V^i} \left(I - \frac{\mathbf{1}\mathbf{1}^T}{n} \right) \left(I - \left(Y^i W^i\right)^{\dagger} \left(Y^i W^i\right) \right) \left(\left(I - \frac{\mathbf{1}\mathbf{1}^T}{n} \right) \left(I - \left(Y^i W^i\right)^{\dagger} \left(Y^i W^i\right) \right) \right)^T \tag{6}$$

$$V^i = \left\| \left(I - \frac{\mathbf{1}\mathbf{1}^T}{n} \right) \left(I - \left(Y^i W^i\right)^{\dagger} \left(Y^i W^i\right) \right) \right\|_F \tag{7}$$

where y^i_j presents the coordinates of sample j in the local sample space identified on data type i; and y_j indicates the global coordinates of sample j, which is a d-dimensional vector; and the 'alignment' matrix L^i is used to construct the unified affine space in order to align the i_{th} local information to the global information; and b is used for matrix Y centralization as a d-dimensional vector. w^i_j is a

scalar measuring the weight of sample j in data type i, forming a diagonal matrix, called the 'correction' matrix W^i which could balance the effects of different data types, i.e. $W^i = diag\left(sqrt(w_1^i),\ sqrt(w_2^i), \ldots, sqrt(w_n^i)\right)$; and the weight value would become larger when sample j could introduce fewer error or more effective information to discern global sample-pattern based on the data type i. And V^i is a scalar and used for scale normalization on the i_{th}-layer data. I denotes as the identity matrix and the symbol † is the pseudo-inverse operator of a matrix.

The equal condition of Eq. (5) would be established, when b and L^i satisfy the following conditions as:

$$\begin{cases} b = \dfrac{Y\mathbf{1}}{n} \\ L^i = Y\left(I - \dfrac{\mathbf{1}\mathbf{1}^T}{n}\right)\left(Y^i W^i\right)^\dagger \end{cases} \tag{8}$$

Actually, the function (5) has no unique optimal solution on the matrix Y and cannot ensure the sparsity of the optimal weight $W^i (i = 1, \ldots, k)$. Thus, it is necessary to add the orthogonality constraints of the matrix Y and L_2-norm of the 'correction matrix' W. Finally, by adding all those soft constraints, the extended objective function of PFA with variables Y and W is defined as follows:

$$\begin{aligned} &\min tr\left(Y\Phi Y^T\right) + \lambda \|W\|_2^2 \\ &s.t. \begin{cases} YY^T = I \\ W^T\mathbf{1} = 1 \\ W >= 0 \end{cases} \end{aligned} \tag{9}$$

where $W^T = \left(w_1^1,\ w_2^1 \ldots, w_n^1, \ldots w_1^i, w_2^i \ldots,\ w_n^i, \ldots w_1^k, w_2^k \ldots, w_n^k\right) = (w_1, \ldots, w_{i*j} \ldots, w_M),\ (M = k * n)$, and λ is a tuning parameter, used to balance the solution accuracy of matrix Y and the sparsity of estimated weight W. To solve the optimization problem (9), we alternatively minimize the objective function with respect to Y and W.

When fixing W, we can update Y by computing the eigenvectors of the matrix Φ with eqn. (6). And the optimal Y is shown by formula:

$$Y = \Psi_d \tag{10}$$

where the matrix Ψ_d is formed by the d eigenvectors of Φ, corresponding to the 2_{nd} to the $(d+1)_{th}$ smallest eigenvalues, and $d = min(d^1, d^2, \ldots, d^k)$

Then fixing Y, the optimal W is the solution of the following problem:

$$\begin{aligned} &\min \Delta^T W + \lambda \|W\|_2^2 \\ &s.t. \begin{cases} W^T\mathbf{1} = 1 \\ W \geq 0 \end{cases} \end{aligned} \tag{11}$$

where $\Delta = \left(\varphi_1, \varphi_2, \ldots, \varphi_{i*j}, \ldots, \varphi_M\right),\ M = k*n$ with $\varphi_{i*j} = \frac{\|y_j - b - L^i y_j^i\|^2}{V^i}$. Assume that Δ are sorted in increasing order, i.e. $\varphi_1 \leq \ldots \leq \varphi_M$.

Then, the optimal solution of the problem (11) is given as follows based on Theorem 2 in Supplementary Information:

$$\begin{cases} w_m = \dfrac{\theta - \varphi_m}{2\lambda}, m = 1, \ldots, N \\ w_m = 0, m = N+1, \ldots, M \\ \theta = \dfrac{2\lambda + \sum_{m=1}^{N} \varphi_m}{N} \\ N = \arg\max_m \left(\theta - \varphi_m > 0\right) \end{cases} \tag{12}$$

From function (12), we can see that the optimal W has only N nonzero entries and can be calculated analytically once we know the optimal N. In practice, to find the optimal N, a simple approach is to check the case from $N = M$ to $N = 1$ decreasingly. When $\theta - \varphi_m > 0$, we determine the value of N, as shown in Algorithm 2.

Algorithm 2 Algorithm to calculate optimal W

Input: Δ (in ascending order) and λ
Output: optimal weight matrix W

1. for $N \leftarrow M$ to 1 do
2. if $N = \arg\ \max_m(\theta - \varphi_m > 0)$ then
3. break
4. end if
5. end for
6. $w_m = \frac{\theta - \varphi_m}{2\lambda}$, for $m = 1, \ldots, N$; $w_m = 0$, for $m = N+1, N+2, \ldots, M$

Based on function (9) and Algorithm 2, the global sample-spectrum can be captured using another algorithm named Algorithm 3.

Algorithm 3 Algorithm to capture the global sample-spectrum Y

Input: the local sample-spectrum of each data type, Y^i $(i = 1,2,\ldots,k)$, and λ
Output: optimal weight matrix W and the global sample-spectrum Y

1. Initilize W
2. Optimizing Y according to formula (10) in main text
3. Optimizing W according to Algorithm 2
4. Repeating step 2 and 3 until Y reaches convergence

2.4 Iterative updating process and clustering method

The procedure from local sample-spectrums extracted by Algorithm 1 based on each data type, to the global sample-spectrum

aggregating across genomic data types identified by Algorithm 3, is presented in a summarized Algorithm 4. Such process has the advantage to reduce bias and obtain an optimal global sample-pattern.

Algorithm 4 The iterative updating process for PFA

Input: the profile of i_{th} data types, i.e. $X^i = [x_1^i, x_2^i, \ldots, x_n^i]$, principal component number d^i and tuning parameter λ
Output: the global sample-spectrum Y

1. Computing the local sample-spectrum of each data type according to Algorithm 1
2. Optimizing Y according to Algorithm 3

Furthermore, the global sample-spectrum, i.e. the sample coordinates in the certain space, can be clustered by any traditional clustering algorithms. In this paper, we adopted k-means clustering (Ding and He, 2004) as a downstream analysis of the main PFA.

3 Results

3.1 Evaluation of PFA on simulated examples

To demonstrate the ability of PFA on subgroup identification by fusing the sample-patterns from various data types, numerical experiments were conducted based on the simulated datasets (Meng *et al.*, 2016), with comparison to several existing methods as iClusterPlus, SNF and moCluster. In addition, the efficiency of our adaptive optimization strategy of PFA was also evaluated in the simulated examples.

3.1.1 Synthetic data

Two categories of numeric datasets were generated for a systematic evaluation against the ranged bias. Each contains 3 data layers (or types) and 400 samples, created from real genomic profiles with biological variation levels under the pre-determined sample clusters by Singular Value Decomposition (SVD), as described in Supplementary Information. The data of relevant layers come from the corresponding GSE51557 (Conway *et al.*, 2015), GSE73002 (Shimomura *et al.*, 2016) and GSE10645 (Nakagawa *et al.*, 2008) (i.e. DNA methylation data, miRNA expression and RNA expression, respectively). We called the 'good-condition' numeric example as simData1 (Supplementary Fig. S1) where all the layers of data could provide partial but effective information to describe the global sample-spectrum (e.g. common or complementary sample-patterns), and named the 'bad-condition' numeric example as simData2 where signals in one of the three layers are substantially insufficient (e.g. Layer II in Fig. 2A), providing little information on the integrated sample-pattern (see Supplementary Information). Briefly, the 400 samples in simData1 and simData2 were randomly selected from the above real genomics data, while four sample clusters (namely, 1–100, 101–200, 201–300, 301–400) are distinctly distributed, where 1–300 are designed with complementary sample-patterns and 301–400 are shared across the given numeric data layers (see Supplementary Information). Note that the global sample-pattern cannot be recovered by any single data layer in the two datasets (Fig. 2A and Supplementary Fig. S1A).

In addition, another simulated dataset (i.e. simData3) was generated to test the power of the adaptive optimization strategy, an important scheme in our method. It contains 6 data layers and 300 samples, randomly generated from real biological dataset GSE10645. Across the six data layers, 3 sample clusters (i.e. 1–100, 101–200, 201–300) were established in Supplementary Fig. S2. In particular, the layer I presents a clear separation, while from layer I to layer VI, the noise levels become progressively strong, making it increasingly difficult to discover such pattern structure (see Supplementary Information).

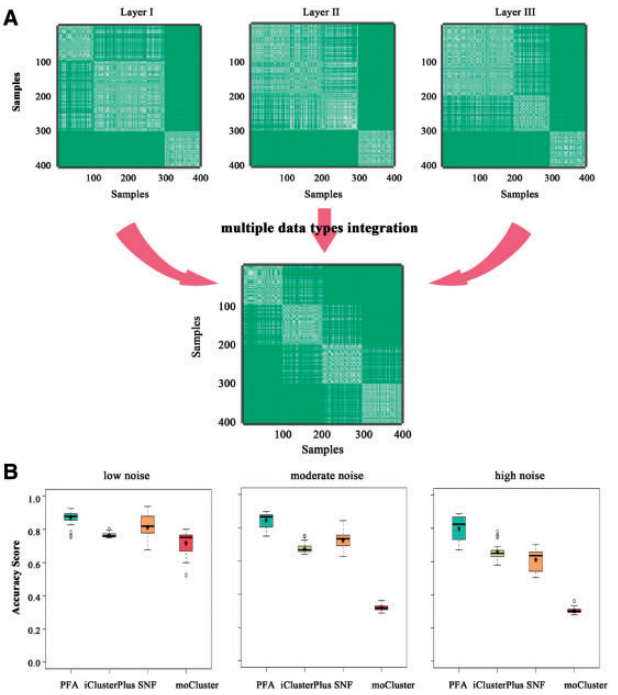

Fig. 2. Sample-pattern diagram of simData2 and comparison of different integration methods. (**A**) Three local sample-patterns based on single data layers, are integrated into a global four-cluster sample-pattern by PFA. Each layer of data could distinguish incomplete clusters, i.e. layer I data classifies samples of 1–100, 101–300 and 301–400 clearly; layer II data classifies samples of 301–400, blurring 1–100/201–300 and 101–200; layer III data classifies samples of 1–200, 201–300 and 301–400 clearly. (**B**) The clustering accuracy comparison among PFA, iClusterPlus, SNF and moCluster under different noise conditions, measures their effectiveness on detecting integrated sample-patterns

3.1.2 Evaluation and comparison on sample-cluster identification

We applied PFA and the other methods to the first two simulated datasets (i.e. simData1 and simData2) for evaluation and comparison on sample-pattern discovery. In order to determine the robustness of these methods in a consistent manner, the datasets were both 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 a criterion, defined as Accuracy Score=(number of correct cluster assignments/total number of cluster assignments). Overall, according to all the results, PFA always succeeded to piece local information together, distinguishing the pre-designed 4 clusters. In addition, the PFA model outperformed over other approaches in terms of both accuracy and consistency across distinct noise strength (Fig. 2B). iClusterPlus also performed stably in the 'good-condition' and 'bad-condition' examples, probably because of its Monte Carlo sampling procedure (Mo *et al.*, 2013), but when the added noise was moderate or high, the precision of iClusterPlus was always low (around 60%). In the 'good-condition', SNF could better

Table 1. The performance of PFA and the other three approaches on simData1 and simData2

	simData1			simData2		
	Low noise	Moderate noise	High noise	Low noise	Moderate noise	High noise
PFA	0.95±0.05	0.86±0.04	0.80±0.06	0.87±0.02	0.84±0.06	0.79±0.12
iClusterPlus	0.89±0.02	0.74±0.03	0.67±0.02	0.76±0.01	0.68±0.03	0.65±0.03
SNF	0.96±0.06	0.83±0.04	0.73±0.09	0.81±0.06	0.72±0.05	0.63±0.08
moCluster	0.70±0.08	0.66±0.13	0.57±0.06	0.71±0.05	0.32±0.02	0.30±0.02

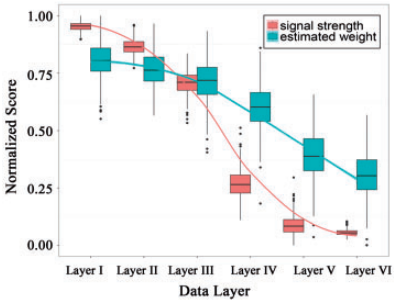

Fig. 3. Evaluation of the adaptive strategy in PFA. Min-max normalization transforms the signal strength (i.e. silhouette scores) and estimated weight (i.e. 'correction' measurements) between 0 and 1. Fit lines were created under 'loess' regression

capture the sample-pattern due to its nonlinearity (Wang *et al.*, 2014), but nevertheless SNF fused more noises in simData2 subgroup identification when the signal of a local pattern became weaker in 'bad-condition', thus pulling down its performance (Fig. 2B, Supplementary Fig. S1B). It is worth to note that moCluster performed the worst against noise and in identifying complicated data structures among these methods (Fig. 2B). As indicated in Table 1 for the summarized simulation results, the performance of our PFA and the previous SNF was similar when all the data layers could provide effective information, while our approach achieved obvious superior data adaptability to SNF and others when signal-to-noise ratio of any used data type was disturbed (i.e. Layer II in simData2). Under all conditions, our method indeed has stable and better ability to identify global sample spectrum than the traditional approaches.

3.1.3 Evaluation of adaptive optimization strategy of PFA

We further demonstrate the effectiveness of our PFA process, i.e. the adaptive optimization strategy, which helps to reduce the side effect of 'dirty' data types based on inherent characteristics of data. In the integration results, weight measurements (i.e. 'correction' matrix) will quantitatively evaluate the contribution by each data type (or sample) to the global sample-pattern identification. Here, we used simData3 to test the utility of our novel strategy and took silhouette score (Rousseeuw, 1987), which measures cluster coherence, as a signal strength indicator in each data layer. PFA was applied to the numeric examples for 1000 times, and then the association between our estimated weight and signal strength was shown in Figure 3. The falling trend across varying data layers is consistent with the evaluation criterion as expected (Spearman correlation test, P-value $< 1.0 \times 10^{-15}$). In other words, enhanced to traditional PCA, PFA is capable to recognize data of high signal-to-noise ratios adaptively and automatically eliminate the effects of noisy data types (or noisy samples) in the integration process. Therefore, such adaptive optimization enables our method to perform robust and powerful in diverse situations as shown in Figure 2.

3.2 Study on CCLE data

We next applied PFA to CCLE datasets (Barretina *et al.*, 2012), where two data types, i.e. matched gene expression data of Affymetrix-HT-HG-U133V2 array and copy number variation (CNV) profiles of Affymetrix-GW-SNP6 array for 415 cell lines, were obtained, representing 11 different tumor types (intra-group cell line number > 25; Supplementary Table S1). When multiple transcript probes point to the same gene, their average value is kept for this gene. And the tissue-specific genes were downloaded from The Human Protein Atlas (Uhlen *et al.*, 2015) and PaGenBase (Pan *et al.*, 2013), while tumor associated genes were collected from GeneCards (Safran *et al.*, 2010). Since upper aerodigestive tract (UADT) contains a number of organs, e.g. tongue, trachea and esophagus, their gene union was regarded as UADT tissue-specific gene set, in the following analysis. One-sided t-test was used to identify the over-expressed tissue-specific genes between certain tumor cell lines and the rest, and two-sided t-test was adopted for identifying the differential expressed/CNV tumor-specific genes. FDR correction was implemented in the process. A high proportion of over- or differential-expressed genes indicates a high score for preserving tissue- or tumor-specific characteristics.

At first, we carried out PFA algorithm to figure out histological tumor patterns with CCLE data. Although the clusters identified by PFA could fit 11 original groups (Supplementary Fig. S3A), the pattern structures between the cell lines seem to have other explanations (clustering overlap ratio: 71%). From Figure 4C and Supplementary Fig. S4, we noted that tumor types in terms of expression data or CNV profiles show different capacities for preserving tissue- and tumor-specific characteristics, which would challenge the histological separations and suggest molecular-based sample-patterns. Thus, derived from data itself, a 9-cluster sample-pattern (Supplementary Fig. S3B) reassigns cell lines as an optimal solution by the current model, and a similar result was obtained in Mo's work (Mo *et al.*, 2013). Different from their analysis, we carefully illustrated sample-patterns based on single and combined data types, and also assessed the clusters using tissue/tumor similarity measurement (Ravasi *et al.*, 2010; Sandberg and Ernberg, 2005).

As expected, the molecular-based sample-spectrum identified by PFA is a spatial picture fusing different side views from single data types (Fig. 4A). For example, among the 'clear' integrative clusters, the gene expression and CNV data can both support the separation of c2 (short for cluster2), c3, c5 and c9 (i.e. shared patterns across multiple data types). In addition, the pattern structures of c1 and c4 are strongly supported by the gene expression profiles, which make the final fused patterns similar to expression-based subgroups (i.e. weakly complementary patterns across multiple data types). Moreover, our model even did a relative fuzzy assignment of c6, c7 and c8 due to the weak expression signals and almost blank CNV information (i.e. strongly complementary patterns across multiple data types), also implying PFA as an effective data-driven method.

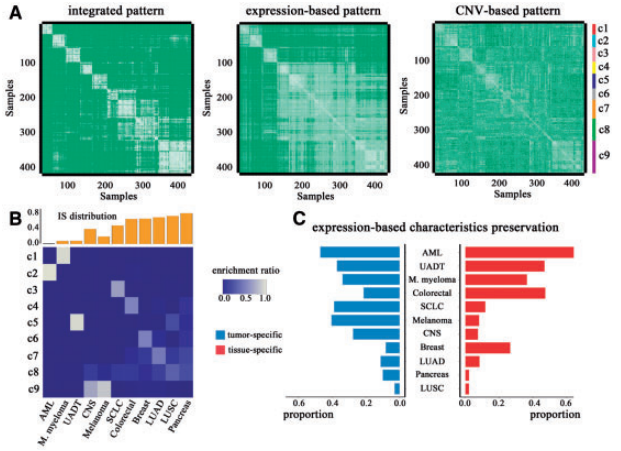

Fig. 4. CCLE: analyses on 11 histological tumor types. (**A**) Local sample-patterns based on single data types (i.e. gene expression and CNV) for 415 cell lines, are integrated into the global 9-cluster sample-pattern by PFA. (**B**) Illustration of associations between histological tumors and integrated cell clusters. The tumor-cluster enrichment ratio R_j^i, is calculated as $(N^i \cap N_j)/N_j$, where N^i, N_j denote number of cell lines assigned in cluster i (i=1,2,...,9) and tumor j (j=1,2,...,11), respectively. Tumor Identity Score is computed as $IS_j = \sum_i^9 R_j^i \log\left(R_j^i/Q_j\right)$, where Q_j is the fractional separation of tumor type j under a null model assuming uniform separation across clusters. (**C**) Tumor types in terms of gene expression profiles show different capacities for preserving tumor- and tissue-specific characteristics. The proportion of up-regulated genes to tissue-specific genes measures the tissue-specific characteristics preservation, and the proportion of differential-expressed genes to tumor associated genes measures the tumor-specific characteristics preservation

Given the overall partitions, we investigated the associations between tumor types and the identified cell clusters, calculating tumor-cluster enrichment ratios, e.g. all the cell lines of acute myelocytic leukemia (AML) are assigned to c2, thus AML-c2 enrichment ratio is 1.0 and AML-ci (i = 1,3,4,...,9) ratios all equal to 0.0. Hence, we could define an Identity Score (IS), which uses entropy to quantify the extent to heterogeneity between cell lines originating from the same tumor type (Fig. 4B). According to the definition, a minimal IS = 0 would be reported for cell lines within the tumor concentrated only in a single cluster, while a maximal IS ≅1 would be reported for cell lines uniformly separated into 9 clusters. And we found that the IS distribution was negatively correlated with maintenance of tissue-specific characteristics at transcription level (Spearman correlation -0.65), as well as more in concordance with tumor-specific gene proportions in terms of expressions and CNV profiles (Spearman correlation -0.85 and -0.85). It means that when cell lines hold more tissue- or tumor-specific characteristics, they prefer to be grouped together and the IS of this tumor type achieves low; on the contrary, when IS is high, it indicates the cell lines would lose their tissue or tumor identity so that the molecular aberrations of great variation underlie different clusters or subgroups. For example, AML and multiple myeloma (M. myeloma) cell lines (IS: 0.0 and 0.07, respectively) keep relatively high proportions of tissue- and tumor-specific gene expression characteristics (Fig. 4C), and are correctly separated, while pancreatic, LUSC (lung squamous cell carcinoma) and LUAD (lung adenocarcinoma) cell lines (IS: 0.70, 0.64 and 0.61, correspondingly) show great tumor heterogeneity, which agreed with previous findings using CCLE datasets (Barretina *et al.*, 2012; Mo *et al.*, 2013).

Next, we investigated the associated characteristics across the composited tumor types within the same fuzzy integrated cluster. For instance, the two main tumor types in c9 are melanoma (IS: 0.17) and central nervous system (CNS) (IS: 0.34), for which 90% melanoma and 74% CNS cell lines are clustered together in c9. Though melanoma cell lines lose many tissue-specific characteristics, they maintain tumor-specific expressions and CNV profiles well (Fig. 4C, Supplementary Fig. S4), to keep tumor homogeneity. Additionally, these melanoma tumor characteristics prefer to be shared with intra-group CNS cell lines (Fisher's exact test, P-value $< 2 \times 10^{-16}$ and = 0.0087 for expression and CNV, respectively), other than inter-group CNS cell lines (P-value = 0.00032 and 0.028), indicating cell lines within c9 more similar at molecular levels. While, another cluster of c5 groups 96% UADT squamous carcinoma (IS: 0.07) and 30% LUSC cell lines together in an alternative way. UADT cell lines have dys-regulation of 37% of tumor-specific genes, but few overlap with c5-LUSC cell lines (Fisher's exact test, P-value = 0.16). Actually, these c5-LUSC cell lines seem to 'gain' those UADT tissue-specific characteristics. Of 262 UADT tissue-specific genes (genes shared with lung tissue removed), 24, 138, 194, 138 were differentially expressed (FDR < 0.05) between c5-UADT cell lines and c5-LUSC, LUAD, SCLC (small cell lung carcinoma), other LUSC ones, respectively. Besides, these differential expressed genes are highly overlapped among the three control groups (Supplementary Fig. S5A). All these results indicate that c5-LUSC may overcome the tissue 'barrier' and become UADT-like (P-value = 0.0045), which was previously implicated in clinical studies (Delahaye-Sourdeix *et al.*, 2015; Hsieh *et al.*, 1997). In addition, through the top 5 over-expressed UADT tissue-specific gene panel (Supplementary Fig. S5B), the c5-LUSC subgroup exactly corresponds to a basal-S100A8 subtype identified on expression data (Wilkerson *et al.*, 2010) with important diagnostic implications. Not coincidentally, such shared molecular patterns (i.e. 9 clusters) between tumors are also observed and verified in Mo's work (Mo *et al.*, 2013). Hence, we can see the pan-cancer integrated patterns would summarize molecular aberrations across tissue origins or tumor types and provide new insights on clinical strategies (Weinstein *et al.*, 2013). Besides, these results also support that PFA indeed has superior ability to reveal sample-patterns (e.g. shared and complementary patterns) across data types of distinct signal-to-noise ratios.

3.3 Study on TCGA data

To demonstrate the effectiveness of PFA for addressing clinical issues (i.e. prognostic prediction), we also applied PFA to the omics data of three cancer sites from the TCGA Data Portal (accessed March, 2016) with most samples in high-risk on survival: kidney renal clear cell carcinoma (KIRC), lung squamous cell carcinoma (LUSC) and glioblastoma multiforme (GBM). For each individual of these tumor types, its corresponding gene expression, miRNA expression and DNA methylation profiles were arranged. To get the largest sample size for analysis, we used different platforms for these cancer types: gene expression data of Illumina-Hiseq-RNASeq platform in KIRC, of Illumina-HiSeq-RNASeqV2 platform in LUSC and of Affymetrix-HT-HG-U133A in GBM; miRNA expression data of Illumina-GA-miRNAseq platform in KIRC and LUSC, of Agilent-miRNA-8X15K platform in GBM; DNA methylation data of Illumina-HumanMethylation-27 platform in KIRC, LUSC and GBM. And publicly available clinical information for each patient was also downloaded, especially including the overall survival data. Then we identified and removed those samples if these samples have more than 20% missing data in any data type. In addition, those features with 20% or more missing values across patients were

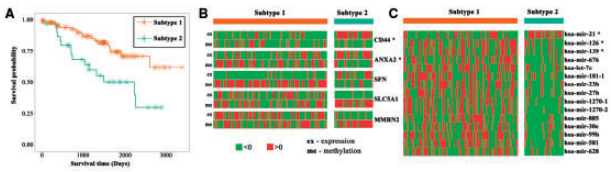

Fig. 5. TCGA: kidney renal cell carcinoma application. (**A**) Illustration of prognostic difference between two subtypes identified by PFA. The figure in (**B**) shows the identified signatures of genes, whose RNA expression (**ex**) and DNA methylation (**me**) profiles in (**B**) are concordantly differential between patients. Differential-regulated miRNAs in the corresponding subtypes are shown in (**C**). * marks the previously validated biomarkers

Table 2. Prognostic prediction of different methods on three tumor types

Cancer types	KIRC (2 clusters)	LUSC (3 clusters)	GBM (3 clusters)
PFA	2.2×10^{-3}	7.3×10^{-4}	1.6×10^{-6}
iClusterPlus	8.6×10^{-2}	4.6×10^{-1}	8.3×10^{-2}
SNF	1.5×10^{-2}	6.0×10^{-2}	3.8×10^{-4}
moCluster	4.7×10^{-2}	1.4×10^{-2}	1.2×10^{-2}

*P-value in each table cell represents statistical significance of log-rank test.

abandoned. And the average values among other samples were imputed for remaining missing data. After such a pre-process, the three datasets with 122 patients in KIRC, 107 in LUSC and 215 in GBM were prepared for our PFA method. The complex contexts, varying in data platforms and sample sizes as heterogeneity, would provide strong evidences to determine whether PFA could identify clinically relevant subtypes (Zang *et al.*, 2016).

After implementing PFA, 2 clusters (i.e. subtypes) for KIRC, 3 clusters for LUSC, and 3 clusters for GBM can be assigned according to the silhouette implications (Supplementary Fig. S6). Together with the available clinical index of each patient, a subgroup-based Cox log-rank model was built and it evaluated the survival risks between the identified subtypes from KIRC (P-value $= 2.2 \times 10^{-3}$), LUSC (P-value $= 7.3 \times 10^{-4}$) and GBM (P-value $= 1.6 \times 10^{-6}$) respectively, as the survival curves shown in Figure 5A and Supplementary Fig. S7. Given that the mentioned methods achieved the same number of clusters, the different performance across cancers (Table 2) indicated that PFA could capture the inherent characteristics of samples and has the best clinical prognosis efficiency comparing with iClusterPlus, SNF and moCluster.

We next analyzed the identified KIRC subtypes and highlighted their clinical differences as further validations. Firstly, we found that the obvious difference of survival time between the two groups (Fig. 5A) highly correlated with tumor metastatic status (Fisher's exact test, P-value $= 3.4 \times 10^{-4}$), implying the separation created by PFA is biologically significant. In addition to KIRC associated features (Supplementary Figs. S8–S10), we performed two-sample one-sided t-test to identify differential mRNA expressions, DNA methylations and miRNA expressions (FDR < 0.05). A set of 5 genes differentially expressed and their anti-methylations were found to be (negatively) associated with such expression pattern (Zeng and Li, 2010), revealing methylation-induced gene expressions (Fig. 5B). In this case, the patients in subtype 2 corresponding to CD44, ANXA2 overexpression, showed a shorter survival time, consistent with previous findings (Li *et al.*, 2015b; Yang *et al.*, 2015). Besides, has-mir-21, ever demonstrated important roles in KIRC regulation and metabolism (Cancer Genome Atlas Research, 2013), also predicts the poor prognosis when up-regulated (Fig. 5C). In addition, our analysis revealed under-expression of has-mir-126 and has-mir-139 in metastatic patients with poor clinical outcomes (Li *et al.*, 2015a), supporting the identified KIRC classification scheme as well. These validated or potential concordant events shared between integrated subtypes would clarify tumor biology and reveal functional associations based on multiple data types simultaneously.

4 Discussion

The distinct types of biological data could provide different viewpoints for understanding the complex biological phenomena (Chen and Zhang, 2016; Ghazalpour *et al.*, 2006; Kutalik *et al.*, 2008; Li *et al.*, 2012; Zhang *et al.*, 2012). In the recent decade, many integration approaches for multi-omics data were proposed to identify 'co-modules' or 'molecule-patterns', including the 'ping-pong' algorithm (Kutalik *et al.*, 2008), non-negative matrix factorization (NMF) methods (Zhang *et al.*, 2012) and generalized regression methods (Mo *et al.*, 2013). However, for many cases, only very limited molecular features or molecule-patterns, underlying the overall sample-spectrum, share similar phenotypes across data types, and thus making the identified subtypes of the clinic samples unreliable. Hence, to overcome such problems, an integration method with automated information fusion and bias correction is demanded.

In this paper, we proposed a data-driven integration approach named as PFA. The key concept of our method is that the sample cluster-/subtype-structure is actually determined, but we look into it from different perspectives (e.g. gene expression, DNA methylation, etc.). Thus, the process to develop a comprehensive perspective could be summarized as adaptive data alignment, including automated information fusion and bias correction in accordance with data inherent structures. Through this scheme (Supplementary Fig. S12A), PFA performs to identify global sample-spectrum with biologically functional interpretations, and its brief summary is as follows. Firstly, PFA obtains the local sample-patterns from all data types by PCA. Then, it aligns those local sample-patterns to a common feature space and synthesizes the global sample-pattern across most data types. During this process, the contributions by each data type (or individual sample) on the global sample-spectrum would be quantitatively measured and the effects of bias or systematic noises would be iteratively decreased to better fit the data. The repeated correction will end when it reaches convergence. After the adaptive optimal alignment, the combinatorial sample-pattern could represent comprehensive characterization, which would be more close to inherent relations in data. To demonstrate the benefits of our adaptive optimization strategy, we carried out both synthetic examples and real cancer datasets, as well as made a comparison to the state-of-the-art integration methods (i.e. iClusterPlus, SNF and moCluster) (Supplementary Fig. S12B). As expected, we found that our estimated weight was identical to signal strength measured by an independent criterion (i.e. silhouette score). In the simulation studies (Supplementary Fig. S11), PFA effectively recovered the designed subgroups and outperformed over the relevant methods, in terms of clustering accuracy and robustness. Besides, the sample-patterns identified by PFA in CCLE datasets and TCGA datasets were proven to be highly correlated with the clinical data, indicating the effectiveness of PFA in deriving biologically meaningful information. Opposed to analyzing individual datasets separately, PFA provides a way to objectively fuse multi-omics data, and it also reveals molecule-patterns across different layers of data, i.e. methylation-induced genes for KIRC, which support the identified sample-spectrum sharing phenotypes. Moreover, the quantitative

assessment would provide a highlighted biological insight into precision medicine, or the approach could be generally applied to other subjects requiring various data sources' integration.

Although our adaptive alignment can best fit the data structures, it still may make a 'weak' or 'wrong' fusion when the biases consistently exist in the majority of data types. For example, in the LUSC subtype identification, the survival curves between two clusters are very close (Supplementary Fig. S7A), and all the other methods mentioned in the contexts cannot provide a clear prognostic separation under the same condition. This is probably because those data types (i.e. expression, miRNA and methylation) are not sufficient to distinguish some sub-cohorts and more data types relevant to LUSC heterogeneity (e.g. somatic mutation or CNV) or clinical covariates (e.g. tumor size, metastatic status, etc.) should be considered. And our method currently has not particularly dealt with discrete data types, e.g. somatic mutation, which should be improved as well in the future work. We also believe the adaptive optimal PFA could be extended to uncover more sophisticated biological features by integrating multi-layer heterogeneous data in time course (Chen *et al.*, 2012).

Acknowledgements

The authors would like to thank Professor Michael K. Ng and Dr. Chuan Chen (Hong Kong Baptist University) for helpful discussions and suggestions.

Funding

This paper was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (CAS) (No. XDB13050000), National Key R&D Program - Special Project on Precision Medicine (2016YFC0903400), National Program on Key Basic Research Project (2014CB910504), NSFC (91439103, 91529303, 31200987, 81471047) and the Natural Science Foundation of Shanghai (17ZR1446100).

Conflict of Interest: none declared.

References

Barretina,J. *et al*. (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, **483**, 603–607.

Cancer Genome Atlas Research,N. (2013) Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature*, **499**, 43–49.

Chari,R. *et al*. (2010) Integrating the multiple dimensions of genomic and epigenomic landscapes of cancer. *Cancer Metast. Rev.*, **29**, 73–93.

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

Chen,L. *et al*. (2012) Detecting early-warning signals for sudden deterioration of complex diseases by dynamical network biomarkers. *Sci. Rep.*, **2**, 342.

Conway,K. *et al*. (2015) Racial variation in breast tumor promoter methylation in the Carolina Breast Cancer Study. *Cancer Epidemiol. Biomarkers Prev.*, **24**, 921–930.

Delahaye-Sourdeix,M. *et al*. (2015) A rare truncating BRCA2 variant and genetic susceptibility to upper aerodigestive tract cancer. *J. Natl. Cancer Inst.*, **107**, djv037–djv037.

Ding,C. and He,X.F. (2004) Cluster structure of K-means clustering via principal component analysis. *Lect. Notes Artif. Int.*, **3056**, 414–418.

Gevaert,O. (2008) Integrating microarray and proteomics data to predict the response on cetuximab in patients with rectal cancer. *Pac. Symp. Biocomput. Pac. Symp. Biocomput.*, **13**, 166–177.

Ghazalpour,A. *et al*. (2006) Integrating genetic and network analysis to characterize genes related to mouse weight. *PLoS Genet.*, **2**, e130.

Gygi,S.P. *et al*. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.*, **17**, 994–999.

Hamid,J.S. *et al*. (2009) Data integration in genetics and genomics: methods and challenges. *Hum. Genomics Proteomics*, **2009**, 869093.

Hsieh,W.C. *et al*. (1997) Temporal relationship between cancers of the lung and upper aerodigestive tract. *Jpn. J. Clin. Oncol.*, **27**, 63–66.

Kutalik,Z. *et al*. (2008) A modular approach for integrative analysis of large-scale gene-expression and drug-response data. *Nat. Biotechnol.*, **26**, 531–539.

Li,W. *et al*. (2012) Identifying multi-layer gene regulatory modules from multi-dimensional genomic data. *Bioinformatics*, **28**, 2458–2466.

Li,M. *et al*. (2015a) MicroRNAs in renal cell carcinoma: a systematic review of clinical implications (Review). *Oncol. Rep.*, **33**, 1571–1578.

Li,X. *et al*. (2015b) Prognostic value of CD44 expression in renal cell carcinoma: a systematic review and meta-analysis. *Sci. Rep.*, **5**, 13157.

Meng,C. *et al*. (2016) moCluster: identifying joint patterns across multiple omics data sets. *J. Proteome Res.*, **15**, 755–765.

Mo,Q. *et al*. (2013) Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proc. Natl. Acad. Sci. U. S. A.*, **110**, 4245–4250.

Nakagawa,T. *et al*. (2008) A tissue biomarker panel predicting systemic progression after PSA recurrence post-definitive prostate cancer therapy. *PloS One*, **3**, e2318.

Pan,J.B. *et al*. (2013) PaGenBase: a pattern gene database for the global and dynamic understanding of gene function. *PloS One*, **8**, e80747.

Ravasi,T. *et al*. (2010) An atlas of combinatorial transcriptional regulation in mouse and man. *Cell*, **140**, 744–752.

Rousseeuw,P.J. (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, **20**, 53–65.

Safran,M. *et al*. (2010) GeneCards Version 3: the human gene integrator. *Datab. J. Biol. Datab. Curat.*, **2010**, baq020.

Sandberg,R. and Ernberg,I. (2005) Assessment of tumor characteristic gene expression in cell lines using a tissue similarity index (TSI). *Proc. Natl. Acad. Sci. U. S. A.*, **102**, 2052–2057.

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

Shen,R. *et al*. (2010) Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, **26**, 292–293.

Shimomura,A. *et al*. (2016) Novel combination of serum microRNA for detecting breast cancer in the early stage. *Cancer Sci.*, **107**, 326–334.

Uhlen,M. *et al*. (2015) Proteomics. Tissue-based map of the human proteome. *Science*, **347**, 1260419.

Wang,B. *et al*. (2014) Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods*, **11**, 333–337.

Weinstein,J.N. *et al*. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.*, **45**, 1113–1120.

Wilkerson,M.D. *et al*. (2010) Lung squamous cell carcinoma mRNA expression subtypes are reproducible, clinically important, and correspond to normal cell types. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.*, **16**, 4864–4875.

Yang,S.F. *et al*. (2015) Annexin A2 in renal cell carcinoma: expression, function, and prognostic significance. *Urol. Oncol.*, **33**, 22 e11–21.

Zang,C.Z. *et al*. (2016) High-dimensional genomic data bias correction and data integration using MANCIE. *Nat. Commun.*, **7**, 11305.

Zeng,T. and Li,J.Y. (2010) Maximization of negative correlations in time-course gene expression data for enhancing understanding of molecular pathways. *Nucleic Acids Res.*, **38**, e1.

Zhang,S. *et al*. (2011) A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA-gene regulatory modules. *Bioinformatics*, **27**, i401–i409.

Zhang,S. *et al*. (2012) Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Res.*, **40**, 9379–9391.

Zhang,Z. and Zha,H. *et al* (2003) Principle manifolds and nonlinear dimensionality reduction via tangent space alignment. *SIAM J. Sci. Comput.*, **26**, 313–338.